Estimating hidden concentrations using a simple RESCUE model with the Ensemble Kalman Filter: proof-of-concept using simulated laboratory measurements

Overview

It is easier to measure the intracellular concentrations of reporter proteins (mCherry, GFP-mCherry) than to measure the concentrations of substrate and product mRNA. A good kinetic model that produces similar reporter protein concentrations might be able to infer these hard-to-measure mRNA concentrations. However, if the parameters used in the kinetic model (e.g., mRNA doubling time) are wrong, the outcomes of the model cannot be trusted. This limits the ability of using kinetic models to infer the concentrations of substrate and product mRNA.

To measure the uncertainties in the model outcomes due to uncertainties in the parameters, it is common to run an ensemble of models, with different sets of parameters. These parameter sets are usually drawn at random from a normal distribution with a spread that represents the uncertainty in the parameters. The ensemble of model outcomes can then be used to generate concentration estimates (mean concentrations) and the uncertainties (standard deviation).

What if we could somehow use laboratory-measured reporter protein concentrations to improve the accuracy and precision of the estimate for all of the concentrations? One such method is the Ensemble Kalman Filter (EnKF). The EnKF broadcasts information from the laboratory measurements to all model concentrations in the ensemble through the statistical connections between all concentrations and the concentrations of the reporter proteins in the ensemble. In other words, there is a potential to use the EnKF to infer hard-to-measure mRNA concentrations!

In this project, we explored this potential of the EnKF using simulated laboratory measurements of mCherry and GFP-mCherry concentrations. Here, we will describe the setup of a simple model of the RESCUE system and outline the EnKF algorithm we used. Following that, the results of applying the EnKF will be described and discussed. Some directions for future work will also be suggested.

If the reader is interested in a more detailed explanation of the EnKF, see Appendix A.

Simple model of RESCUE

The RESCUE system is made up of 6 components:

1. substrate mRNA (S),
2. product mRNA (P),
3. mRNA editing enzyme (E),
4. enzyme-substrate mRNA (ES),
5. GFP-mCherry protein that is generated from S (GC), and,
6. GFP protein that is generated from P (G).

In this project, we set up a simple model that simulates the intracellular evolution of the concentrations of the RESCUE system components. These processes are:

1. equilibrium binding of E to S,
2. production of P from ES,
3. production and decay of S,
4. production and decay of P,
5. production of GC from S and decay of GC, and,
6. production of C from P and the decay of C.

In the model, the equations that govern these processes and their parameters can be found in Appendix B.

After some tuning of the parameters, we managed to achieve equilibrium concentrations of G and GC on the order of 1000s of protein molecules per cells (Figure 1). These values are consistent with a result from the wet laboratory: they experienced sensor oversaturation when they tried to measure fluorescence in the cells.

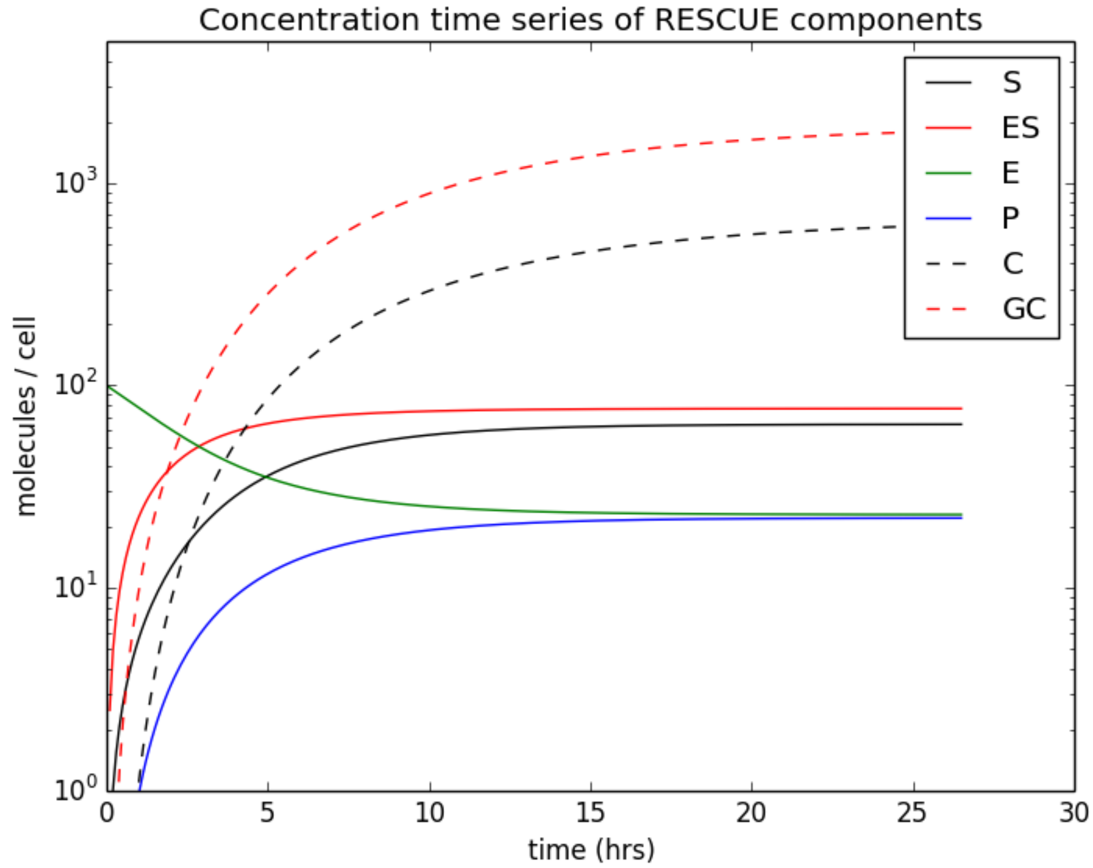

Figure 1: Model components after tuning. Simulated laboratory measurements will also be drawn from this model run later in the EnKF experiment.

EnKF algorithm

We used the stochastic filter form of the EnKF (Evensen, 1994). The procedure is as follows:

1. Generate an ensemble of model outcomes.

2. Compute the prior ensemble average $\langle \boldsymbol{\psi_b} \rangle$ (aka prior estimate), prior ensemble perturbations and the prior ensemble covariance $\boldsymbol{B}$ via:

$$\langle \boldsymbol{\psi_b} \rangle = \frac{1}{N_e} \sum_{e=1}^{N_e} \boldsymbol{\psi_{b,e}} \,,$$

$$\boldsymbol{\psi'_{b,e}} = \boldsymbol{\psi_{b,e}} - \langle \boldsymbol{\psi_b} \rangle,$$

$$\boldsymbol{B} \; = \frac{1}{N_e \; - 1} \sum_{e=1}^{N_e} \boldsymbol{\psi'_{b,e}} \, \boldsymbol{\psi'_{b,e}}^{\top},$$

where N_e is the number of ensemble members, $\boldsymbol{\psi_{b,e}}$ is the e-th prior ensemble member concentrations and $\boldsymbol{\psi_{b,e}}'$ is the perturbation of the e-th prior ensemble member from the ensemble mean.

3. Compute the Kalman gain matrix:

$$\boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{\top} + \boldsymbol{R})^{-1}.$$

4. Compute the posterior ensemble mean via

$$\langle \boldsymbol{\psi_a} \rangle = \langle \boldsymbol{\psi_b} \rangle + \boldsymbol{K}[\boldsymbol{y} - \boldsymbol{H}\langle \boldsymbol{\psi_b} \rangle]$$

This is the posterior estimate of the EnKF.

5. Generate the posterior ensemble by first computing the posterior perturbation ensemble:

$$\boldsymbol{\psi_{a,e}}' = \boldsymbol{\psi_{b,e}}' + \boldsymbol{K}[\boldsymbol{y_e}' - \boldsymbol{H}\boldsymbol{\psi_{b,e}}']$$

where $\boldsymbol{y_e}'$ is the e-th random draw from the normal distribution $N(\boldsymbol{0}, \boldsymbol{R})$. Then, add the perturbations to the posterior ensemble mean to obtain the posterior ensemble:

$$\boldsymbol{\psi_{a,e}} = \boldsymbol{\psi_{a,e}}' + \langle \boldsymbol{\psi_a} \rangle$$

where $\boldsymbol{\psi_{a,e}}$ is the e-th posterior ensemble member.

Setup of EnKF experiment

To simulate the scenario of uncertain and erroneous parameters, we will use the “erroneous” parameters that are 80% in value of the simulated truth parameters. The large uncertainty of this parameter is represented by defining standard deviations that are 30% of each parameter. 20000 sets of parameters were then randomly drawn from a multivariate normal distribution with centered on the “erroneous” parameters, with the aforementioned standard deviations, and all cross-parameter covariances set to zero. These 20000 model runs were then integrated in time until all of the concentrations have reached static equilibrium. The resulting ensemble of 20000 sets of concentrations form the prior ensemble[1].

The simulated measurements are equilibrium concentrations of GFP and GFP-mCherry. To mimic measurement errors, perturbation concentrations of GFP and GFP-mCherry were drawn from a normal distribution with zero mean and standard deviation that is 1% of the equilibrium concentrations of the simulated truth (Figure 1). These perturbations were respectively applied to the simulated truth’s equilibrium concentrations of GFP and GFP-mCherry. The result is the simulated measurements.

The EnKF algorithm used here is outlined in the previous section. Here, the simulated measurements of equilibrium GFP and mCherry-GFP concentrations were used. R is a diagonal measurement error covariance, with the squares of the standard deviations used to generate measurement perturbations as the diagonal terms.

[1] The reason why we utilized 20000 ensemble members that is to generate smooth looking histograms. In reality, similar results can be achieved with as little as 20 ensemble members. As such, the EnKF is actually computationally cheap to run.

Results of EnKF experiment

Table 1: Comparison of the absolute errors of the prior and posterior estimated concentrations

Species	Absolute prior estimate error w.r.t. truth	Absolute posterior estimate error w.r.t. truth	Absolute ratio of prior/posterior errors
Enzyme-substrate	2.5%	1.0%	2.5
Free enzyme	8.4%	3.4%	2.5
Substrate mRNA	11%	1.3%	8.6
Product mRNA	16%	1.1%	14
mCherry	28%	0.61%	46
GFP-mCherry	21%	0.37%	57

EnKF estimate is at least 50% closer to the truth than prior estimate, and more precise

Figures 2 and 3 shows the 2D histograms of the prior ensemble concentrations with respect to mCherry and GFP-mCherry. From Figures 2 and 3, we notice that the center locations of the posterior ensemble clusters are closer to the truth than that of the prior ensemble. This indicates that the EnKF is able to generate estimates for all concentrations that are better than those of the prior ensemble. The errors of the estimates from the prior and posterior ensembles with respect to the magnitude of the truth are displayed in Table 1.

As can be seen from Table 1, there is a dramatic reduction of ensemble-estimated errors after applying the EnKF for all quantities. The improvements are the most dramatic for the ensemble-averaged mCherry and GFP-mCherry as those quantities were measured. While the improvements estimates of the mRNA and enzyme concentrations were not as dramatic, those improvements are significant. The magnitude of the errors are clearly reduced by 50% or more with the EnKF.

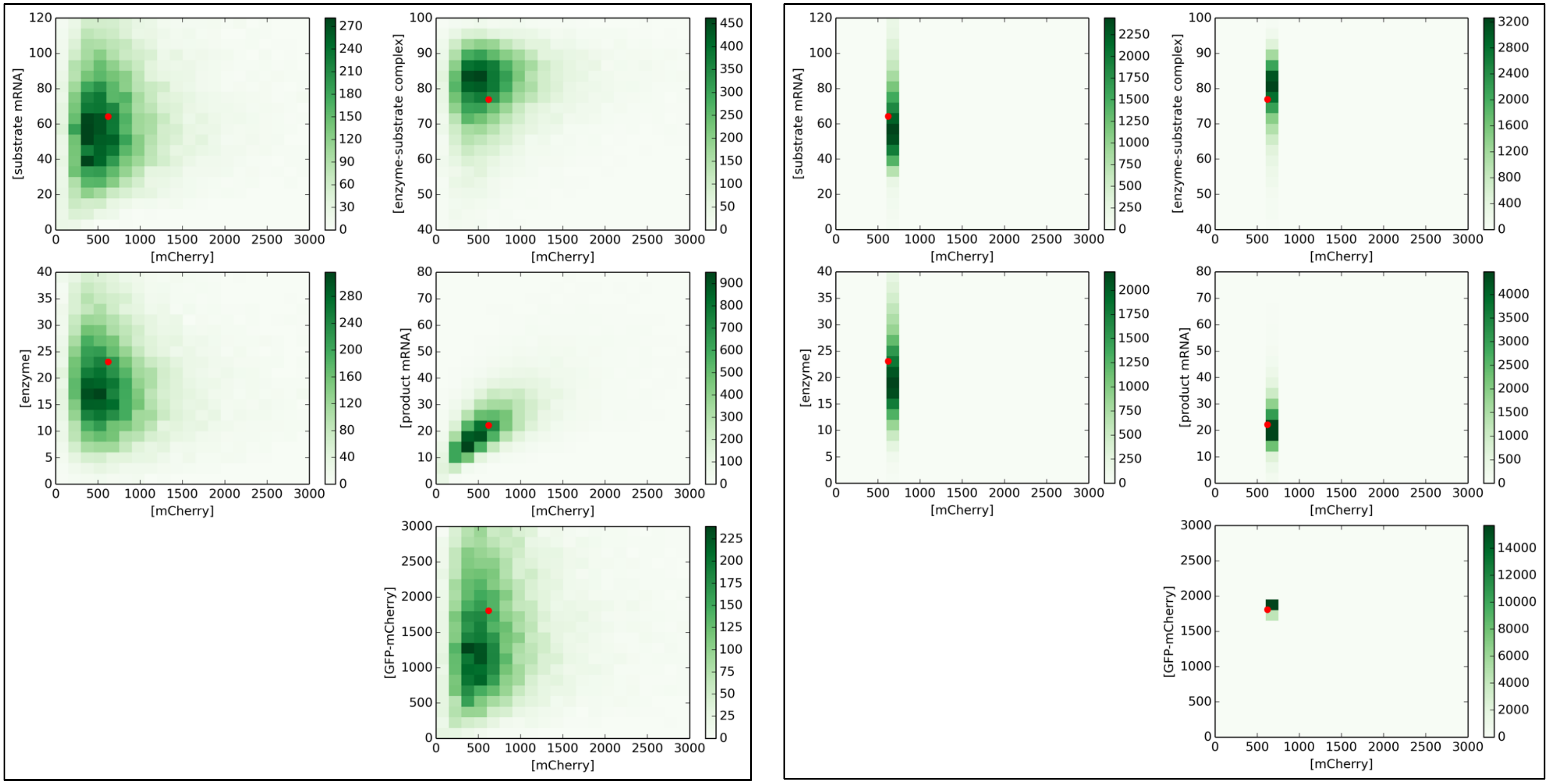

Figure 2: Histograms showing the prior ensemble (left) and posterior ensemble (right) equilibrium concentrations (left) with respect to mCherry concentrations. The red dot indicates the equilibrium concentrations from the simulated truth. Note that the x-axis and y-axis are the same across panels

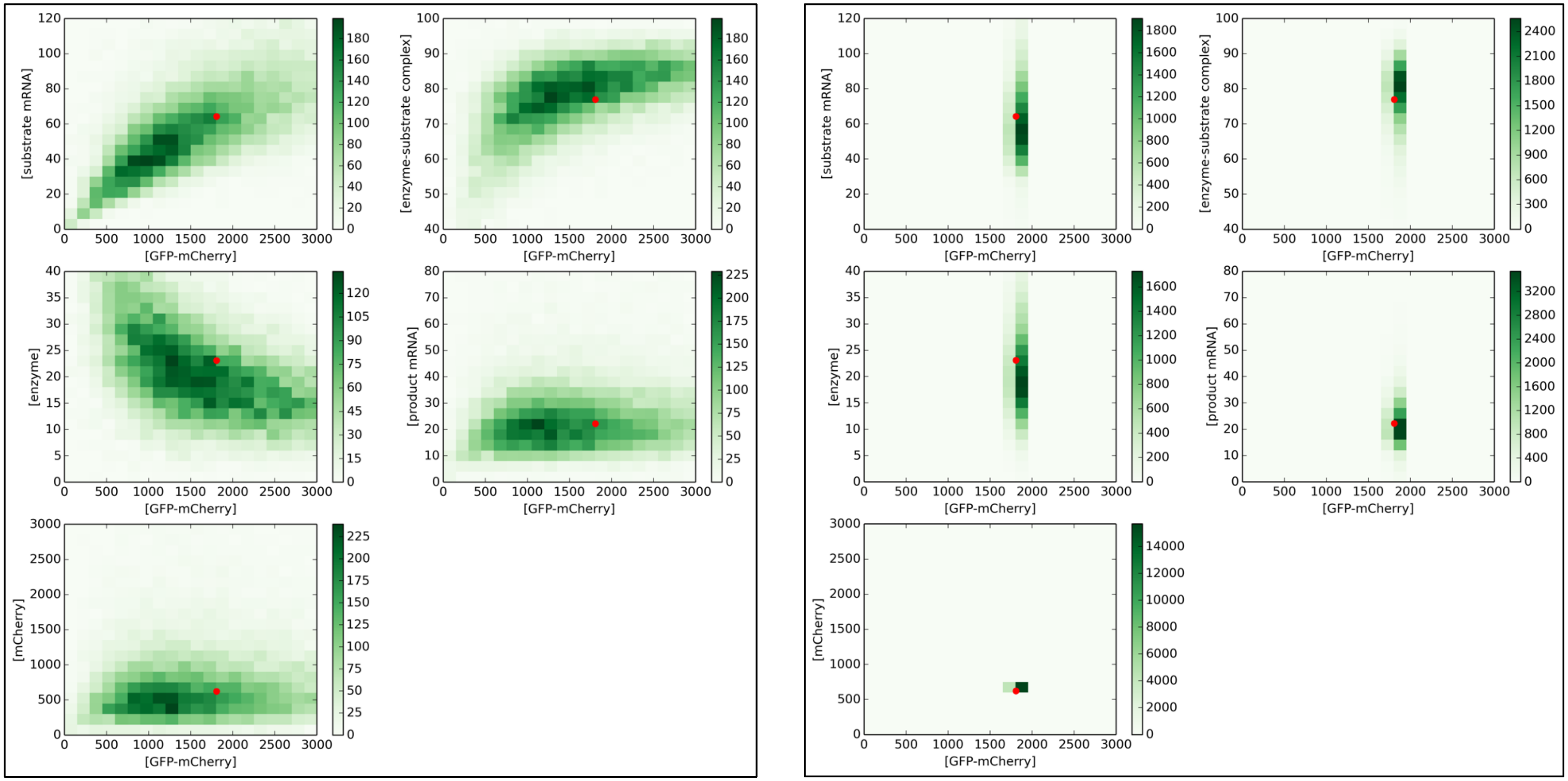

Figure 3: Histograms showing the prior ensemble (left) and posterior ensemble (right) equilibrium concentrations (left) with respect to GFP-mCherry concentrations. The red dot indicates the equilibrium concentrations from the simulated truth. Note that the x-axis and y-axis are the same across panels

It is also interesting to note that the improvements for the mRNA estimates are the most dramatic after those of mCherry and GFP-mCherry. This is because the concentrations of the substrate and product mRNA most directly impact the concentrations of the mCherry and GFP-mCherry. As such, there should be very strong statistical connections between substrate and product mRNA, and the concentrations of the mCherry and GFP-mCherry. These strong connections explain why the EnKF is very effective at updating these hidden quantities.

The estimates of enzyme-substrate and free enzyme received the least improvement. This is not surprising, these things only impact the concentrations of mCherry and GFP-mCherry indirectly through the substrate and product mRNA. As such, a weaker statistical connection is to be expected, resulting in a weaker improvement in these quantities.

Summary

In summary, we constructed a simple model of the RESCUE system that simulates the evolution of intracellular substrate mRNA, product mRNA, mRNA editing enzyme, enzyme-substrate mRNA, GFP-mCherry protein that is generated from the substrate mRNA, and, GFP protein that is generated from the product mRNA. We tested the potential of using the EnKF to infer the hard-to-measure mRNA concentrations using simulated laboratory measurements of GFP and GFP-mCherry. It was found that with the EnKF, the ensemble estimated concentrations of both mRNA species improved dramatically.

Appendix A: How the EnKF works

The EnKF is essentially a Bayesian inference process that operates via Bayes' theorem:

$$P(\boldsymbol{\psi} \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid \boldsymbol{\psi})\, P(\boldsymbol{\psi}).$$

The proportionality constant is simply a normalization constant. Here $\boldsymbol{\psi}$ is a vector containing a list of model concentrations and $\boldsymbol{y}$ is a vector containing laboratory measurements. According to the Bayesian interpretation, $P(\boldsymbol{\psi})$ is the probability density function that $\boldsymbol{\psi}$ is true, before looking at the laboratory measurements (prior pdf). $P(\boldsymbol{y} \mid \boldsymbol{\psi})$ is the probability density function of measuring the values in $\boldsymbol{y}$, given that $\boldsymbol{\psi}$ is true. The result of combining the two pdfs through Bayes' theorem is the probability that $\boldsymbol{\psi}$ is true, given the laboratory measurement $\boldsymbol{y}$ (posterior pdf).

Under the assumption that all three pdf's are Gaussian, the three pdfs can be written as:

$$P(\boldsymbol{\psi}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\psi} - \langle\boldsymbol{\psi_b}\rangle)^\top \boldsymbol{B}^{-1}(\boldsymbol{\psi} - \langle\boldsymbol{\psi_b}\rangle)\right)$$

$$P(\boldsymbol{y} \mid \boldsymbol{\psi}) \propto \exp\left(-\frac{1}{2}[\boldsymbol{h}(\boldsymbol{\psi}) - \boldsymbol{y}]^\top \boldsymbol{R}^{-1}[\boldsymbol{h}(\boldsymbol{\psi}) - \boldsymbol{y}]\right)$$

$$P(\boldsymbol{\psi} \mid \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\psi} - \langle\boldsymbol{\psi_a}\rangle)^\top \boldsymbol{A}^{-1}(\boldsymbol{\psi} - \langle\boldsymbol{\psi_a}\rangle)\right).$$

In the EnKF, the prior pdf are represented by the outcomes of an ensemble of simulations. $\langle\boldsymbol{\psi_b}\rangle$ is the ensemble-averaged concentrations, and $\boldsymbol{B}$ is the covariance matrix derived from the ensemble of concentrations. $\boldsymbol{h}$ is an operator that ingests the model results and extracts data that corresponds to the laboratory measurement, and $\boldsymbol{R}$ is the error covariance of the laboratory measurements. For most purposes, $\boldsymbol{R}$ is assumed to be a diagonal matrix (i.e., the errors are not correlated between laboratory measurements of different quantities).

When the EnKF is applied, all of the members of the prior ensemble gets updated to generate a new ensemble, which represents the posterior pdf. As in the case of the prior pdf, the vector $\psi_a$is the ensemble-average of the posterior ensemble and $A$is the posterior covariance matrix derived from the posterior ensemble.

To get a feeling of how the EnKF estimates the hidden concentrations, we can explicitly solve for $\boldsymbol{\psi_a}$, the posterior estimate of all concentrations. It can be shown that [see Kalnay (2002)]:

$$\langle \boldsymbol{\psi_a} \rangle - \langle \boldsymbol{\psi_b} \rangle = \delta \langle \boldsymbol{\psi_a} \rangle = \boldsymbol{B}\boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R})^{-1} [\boldsymbol{y} - \boldsymbol{h}(\langle \boldsymbol{\psi_b} \rangle)]$$

where

$$\boldsymbol{H} \equiv \frac{\partial \boldsymbol{h}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} |_{\boldsymbol{\psi} = \langle \boldsymbol{\psi_a} \rangle}.$$

For the simplicity of subsequent discussions, we will assume that h is a linear operator. In other words,

$$\boldsymbol{h}(\boldsymbol{\psi}) = \boldsymbol{H}\,\boldsymbol{\psi}$$

$$\frac{\partial \boldsymbol{h}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \boldsymbol{H} \quad \forall \quad \boldsymbol{\psi} \in \mathbb{R}^n$$

In the limit of a single measurement, the estimate reduces to

$$\delta \langle \boldsymbol{\psi_a} \rangle \propto Cov(\boldsymbol{\psi}, \boldsymbol{H\psi})[\boldsymbol{y} - \boldsymbol{H}\langle \boldsymbol{\psi_b} \rangle]$$

In other words, the difference between the prior ensemble's estimate of the concentration (i.e., the ensemble mean) and the measured concentrations are broadcasted to all of the concentrations contained in $\boldsymbol{\psi}$. This broadcasting is done through the prior ensemble covariance between all concentrations and the quantity to be measured. This broadcasting through the prior covariance is the reason why the EnKF can be used to estimate hidden concentrations.

Aside from that, the estimate from applying the EnKF is also more precise than the prior ensemble. This can be deduced by solving for the posterior covariance matrix [see Kalnay (2002)]:

$$\boldsymbol{A} = \boldsymbol{B} - \ \boldsymbol{B}\boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\top + \boldsymbol{R})^{-1} \boldsymbol{H}\boldsymbol{B}$$

For ease of illustration, we will consider a one-variable model. The posterior covariance matrix reduces to a posterior variance (${\sigma_a}^2$) and likewise for the prior covariance matrix (${\sigma_b}^2$). Thus:

$${\sigma_a}^2 = {\sigma_b}^2 (1 - \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_o}^2}) < {\sigma_b}^2$$

where ${\sigma_o}^2$ is the observation error variance. From this equation, it is clear that the estimate from the EnKF is more precise than that of the prior ensemble.

Appendix B: RESCUE modelling equations

Equilibrium curve in phase space

When a system is out of equilibrium, the new equilibrium state can be obtained by considering the chemical equilibrium equation and the conservation of mass:

$$[E]\,[S] \;= K_{eq}[ES]$$
$$[E] \;+\; [ES] \;=\; E_T$$
$$[S] \;+\; [ES] \;=\; S_T$$

During the equilibration process, E_T and S_T are constants (mass conservation). Solving this set of equations yields the following:

$$[ES] \;= \frac{1}{2}(E_T + S_T + K_{eq}) \;\pm \frac{1}{2}\sqrt{(E_T + S_T + K_{eq})^2 - 4\,E_T\,S_T}$$

Clearly, there can only be one possible concentration of $[ES]$, if given a value of E_T and S_T. The unrealistic solution is one which would definitely result in a negative concentration if plugged into the conservation equations. From the conservation equations, for concentrations to always be non-negative, the conditions are

$$0 \leq [ES] \leq E_T \quad \text{and} \quad 0 \leq [ES] \leq S_T.$$

Since E_T , S_T and $K_{eq}$are always positive, then it is clear that

$$[ES] \;= \frac{1}{2}(E_T + S_T + K_{eq}) \;- \frac{1}{2}\sqrt{(E_T + S_T + K_{eq})^2 - 4\,E_T\,S_T}$$

In our simulations, if we assume equilibrium to always hold, the enzyme-substrate concentrations must always fall on this curve. To recast this into a form suitable for time-integrations, we apply the chain rule. For ease of writing, we define:

$$f(E_T, S_T) \;\equiv \frac{1}{2}(E_T + S_T + K_{eq}) - \frac{1}{2}\sqrt{(E_T + S_T + K_{eq})^2 - 4\,E_T\,S_T}$$

Then,

$$\partial_t\,[ES] = \partial_{E_T}\,f(E_T, S_T)\;\,\partial_t\,E_T \;\;+\;\; \partial_{S_T}\,f(E_T, S_T)\,\partial_t\,S_T$$

The ∂_xdenotes a partial derivative with respect to variable x.

As such, to determine the evolution of the enzyme-substrate concentration, we need the time-derivatives of the total enzyme and total substrate concentrations.

Evolution of total enzyme due to ES ⇒ P + E

Assuming no creation or destruction of cas13-gRNA, then,

$$\partial_t \, E_T = 0$$

I.e.,

$$\partial_t \, [ES] = \; \partial_{S_T} \, f(E_T, S_T) \, \partial_t \, S_T$$

Evolution of total substrate

The substrate is generated constitutively and decays over time. Furthermore, the ES ⇒ P + E reaction removes ES. Hence, the total substrate's rate of change is:

$$\partial_t \, S_T = g_S - d_P[S] - K_P[ES]$$

g_S and d_S are the constitutive production and decay constants. The last term is the loss of substrate from ES ⇒ P + E.

Evolution of modified mRNA (P)

$$\partial_t \, [P] = K_P[ES] - d_P[P]$$

d_P is the decay constant for the modified mRNA. It is likely that $d_P = d_S$.

Production and decay of reporter proteins

$$\partial_t \, [GC] = \gamma_{GC}[S] - d_{GC}[GC]$$
$$\partial_t \, [C] = \gamma_C[P] - d_C[C]$$