Extending OpenKIM with an Uncertainty Quantification Toolkit for Molecular Modeling

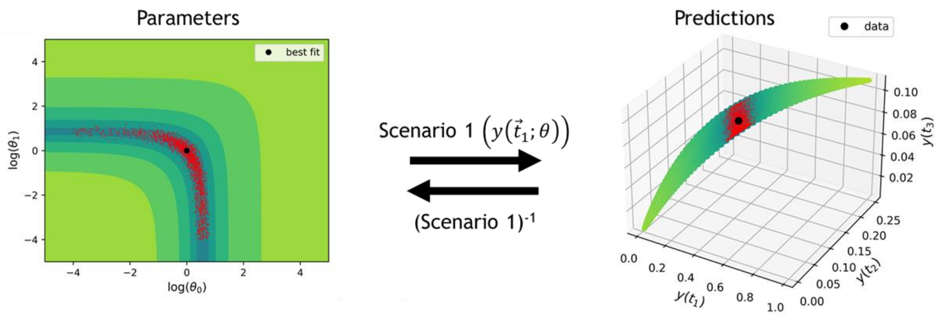
Yonatan Kurniawan, Cody Petrie, Mark Transtrum,

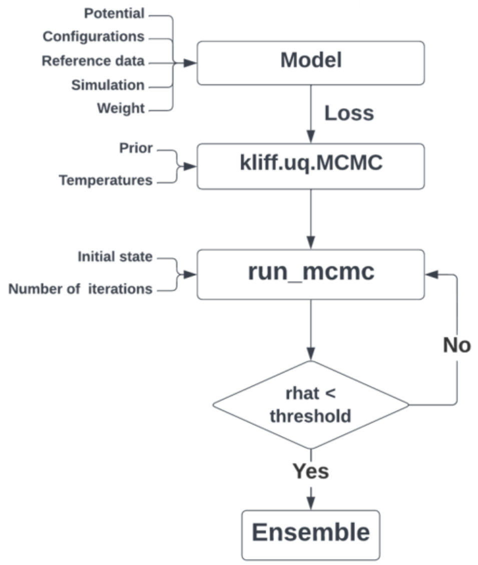
Ellad Tadmor, Ryan Elliott, Daniel Karls, Mingjian Wen

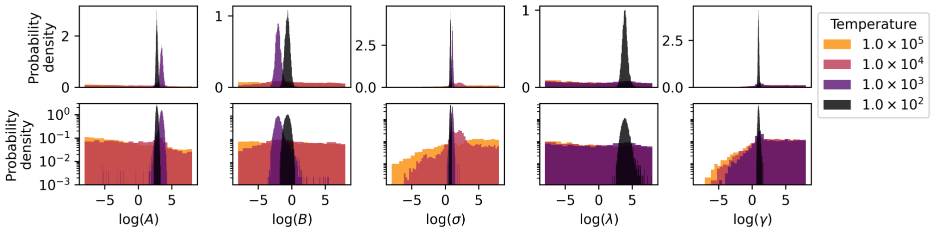
Contact: kurniawanyo@outlook.com

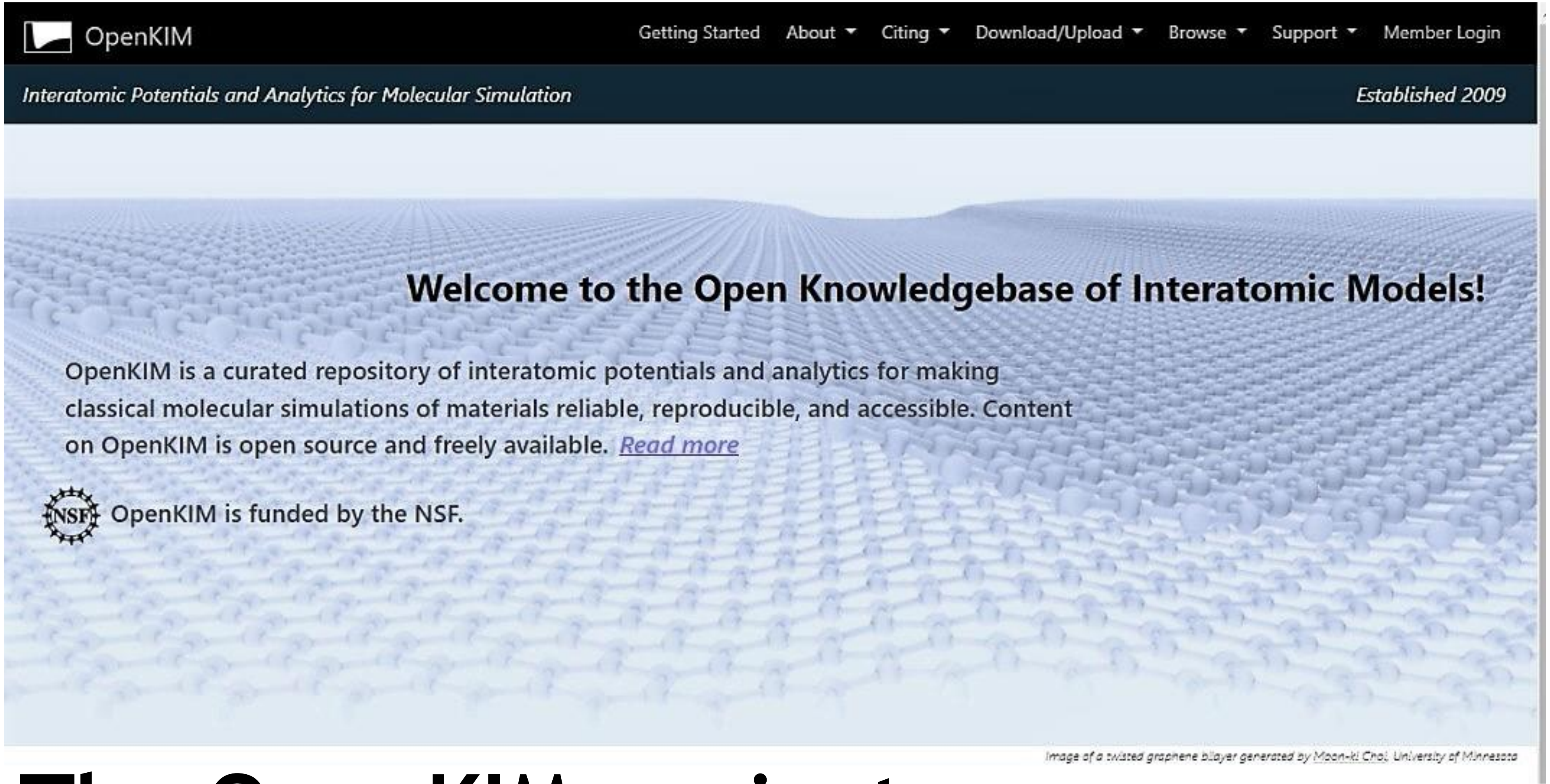
Outline

1. The OpenKIM project
2. Introduction to uncertainty quantification
3. UQ extension to KLIFF
4. Demonstration: Study of SW potential
5. Conclusion and Future work







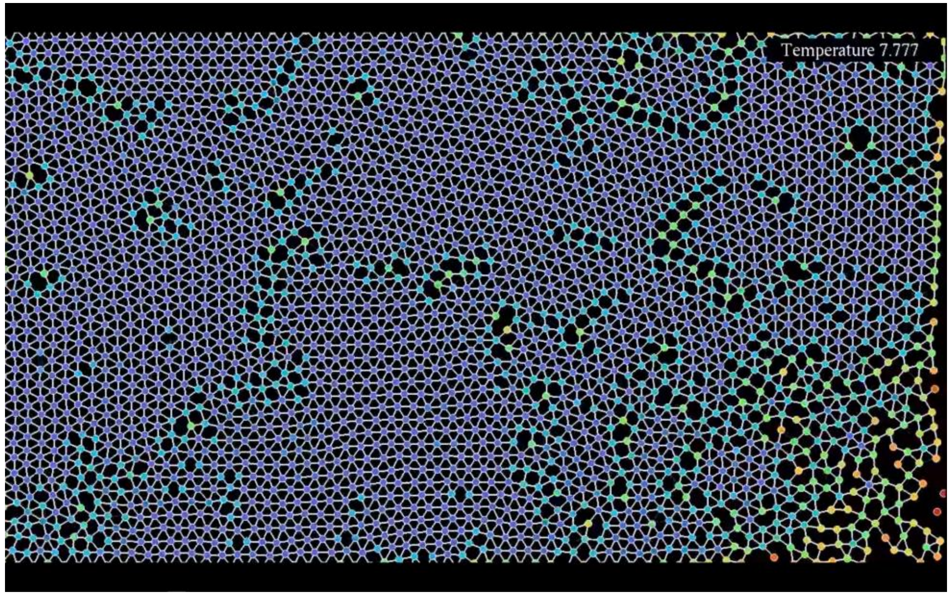


The OpenKIM project

https://openkim.org/

Interatomic potential

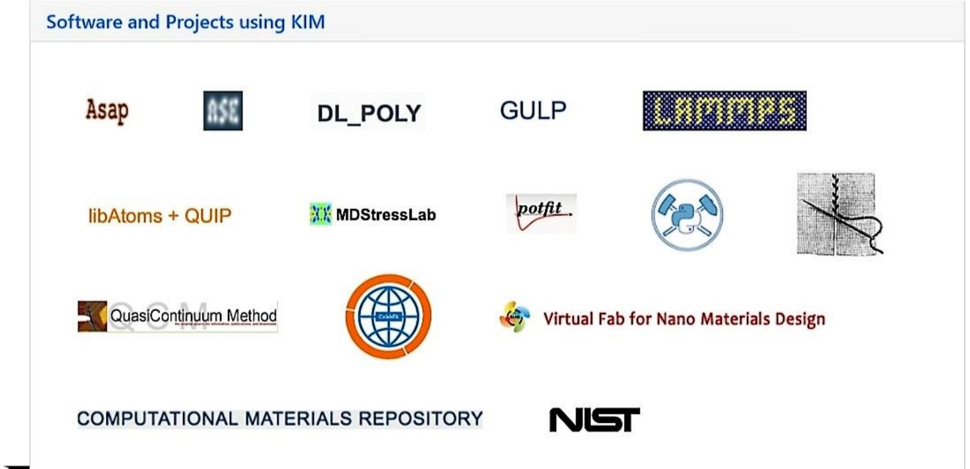
- In atomistic scale simulation, the atoms are treated as classical particles.
- Interatomic potential (IP) approximates interaction energy between atoms.
- IPs are developed for specific applications, resulting in plethora of potentials.
- The functional forms of these potentials have limited scope, miss some physics, and thus introduce model errors.

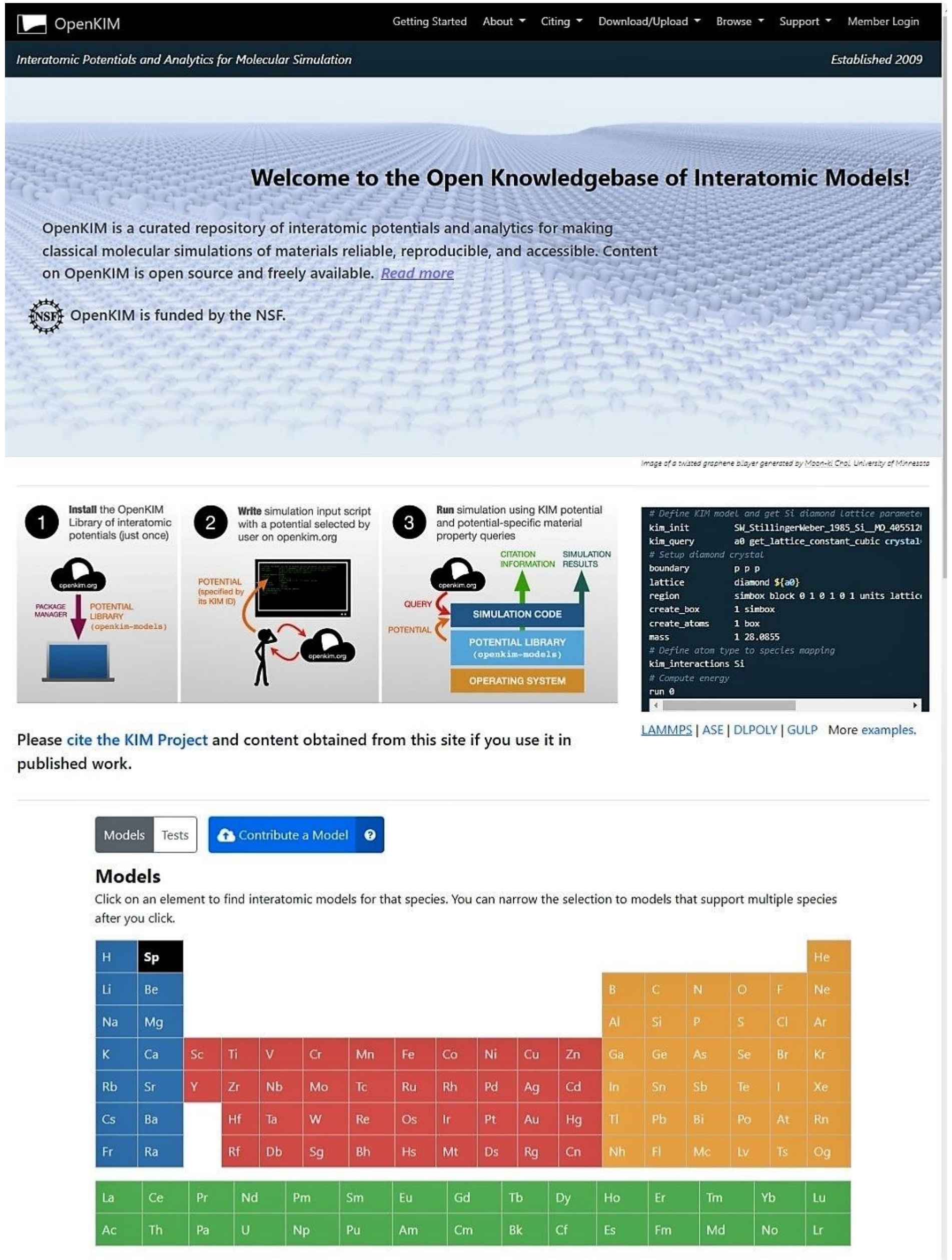


(Berglund, *Freezing and melting at the molecular scale: a representation with atomic bonds* 2021 https://www.youtube.com/watch?v=LdTDIpRx0XQ)

OpenKIM repository

- OpenKIM project aims to collect and standardize the computational implementation of IPs.
- Collected IPs are archived in OpenKIM repository (openkim.org/).
- KIM API allows seamless integration of these IPs with many simulation programs.





https://openkim.org/

BYU

KIM-based Learning-Integrated Fitting Framework

- KLIFF is a general-purpose fitting framework for IPs.
- KLIFF employs the force-matching algorithm [1].
- The IPs are trained to match atomic forces of several configurations from first-principle simulation.
- The trained IPs conform to the KIM API.

https://kliff.readthedocs.io/

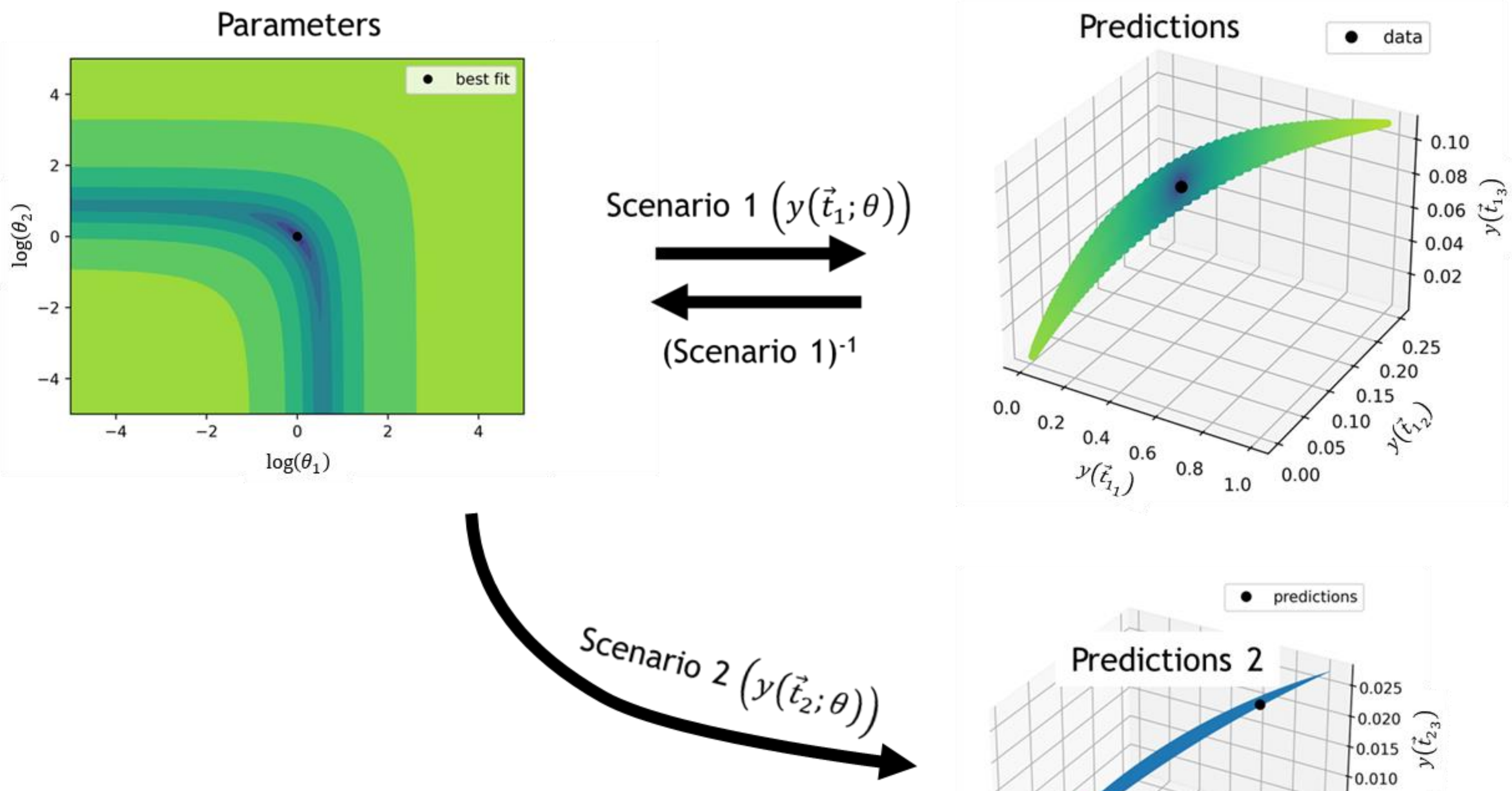
[1] F. Ercolessi and J. B. Adams, “Interatomic Potentials from First-Principles Calculations: The Force-Matching Method,” *EPL*, vol. 26, no. 8, p. 583, Jun. 1994, doi: 10.1209/0295-5075/26/8/005.

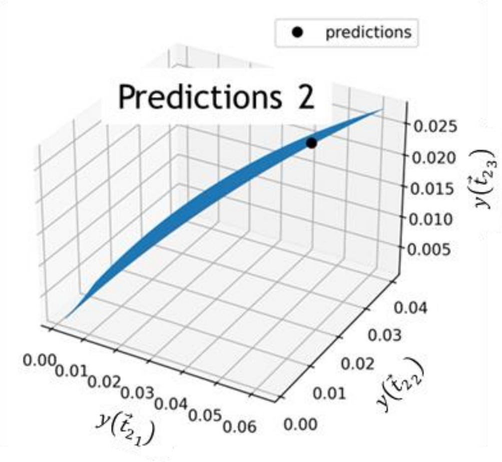
Contribution

- We integrate an uncertainty quantification framework into KLIFF.

Goal

- We want to facilitate UQ studies for IPs.
- We hope that this integration can lead to more transparent and reproducible UQ analysis for IPs.

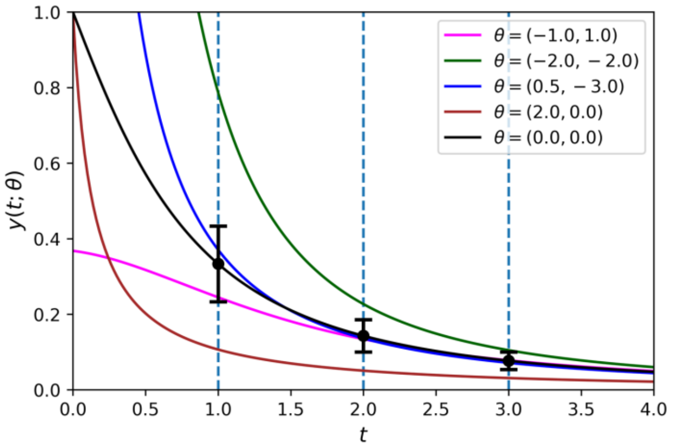


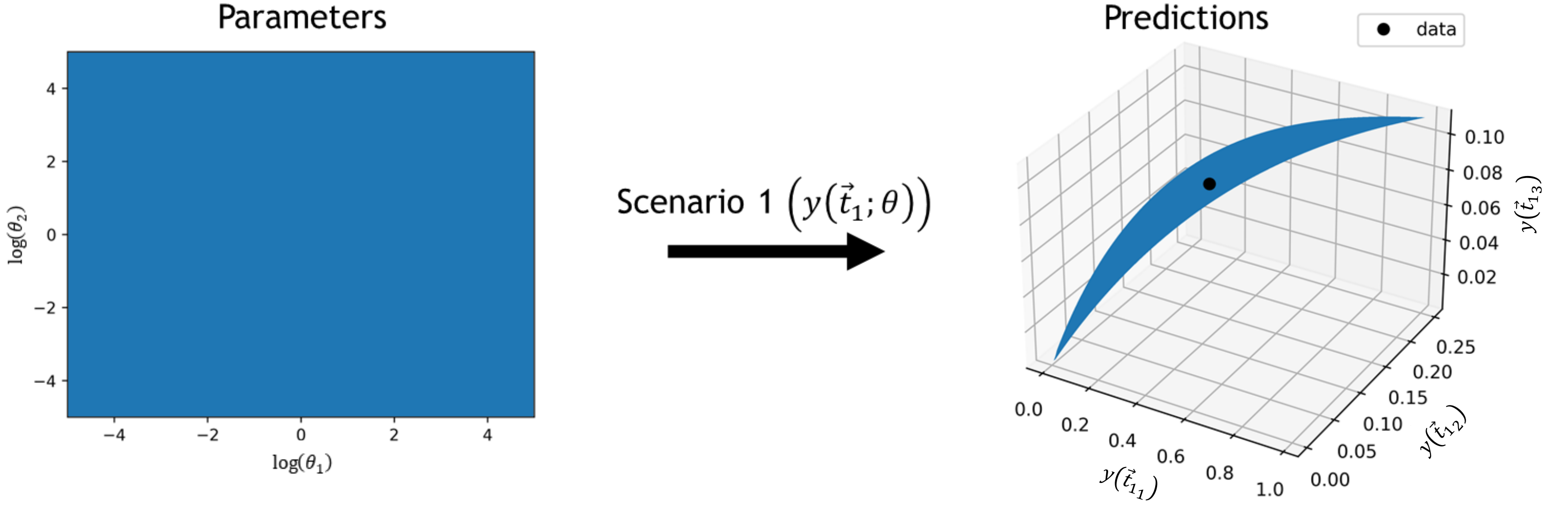


Introduction to uncertainty quantification

Geometry of a model

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$

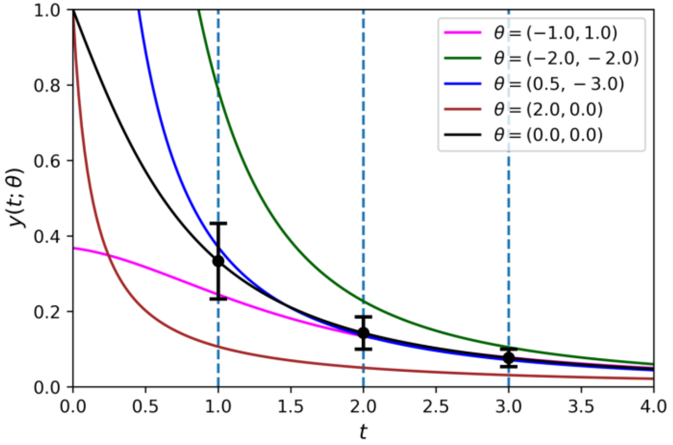




- Model is a mapping from a parameter space to a prediction space.
- The model manifold is the range of the model map.

Loss function

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$



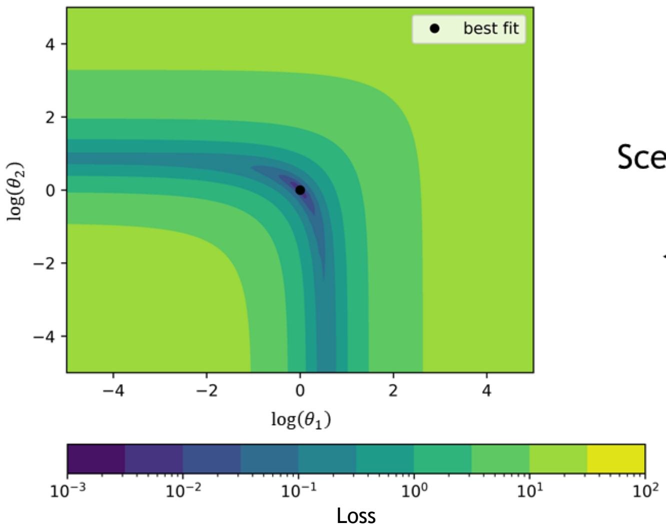
Assumptions:

$$d_m = y(t_m;\theta) + \xi_m$$
$$\xi_m \sim \mathcal{N}(0, \sigma_m)$$

Loss function:

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$

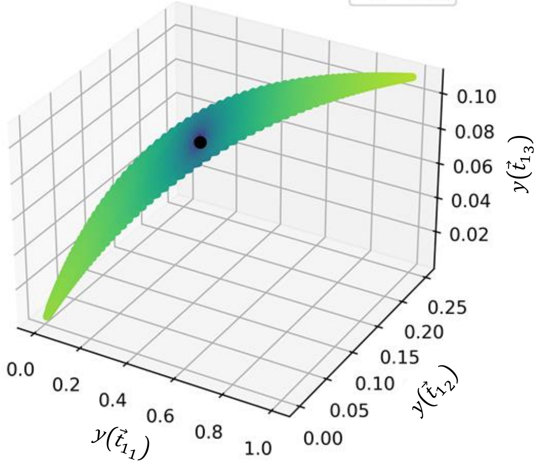
Parameters



Scenario 1 $\left(y(\vec{t}_1;\theta)\right)$

(Scenario 1)$^{-1}$

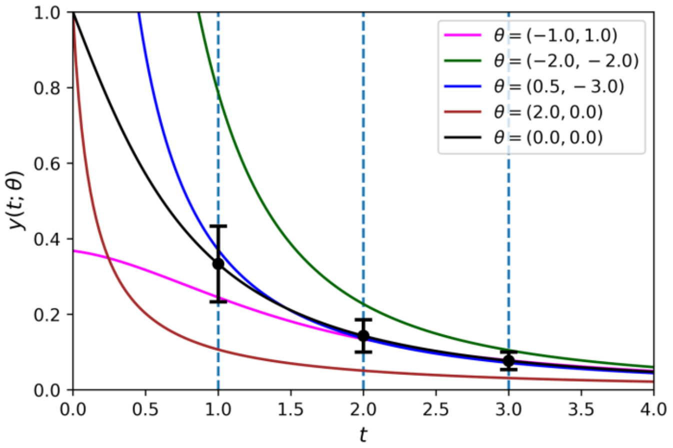
Predictions



- Loss function measures the quality of model predictions compared to the observed data.
- The best fit parameters minimize the loss function.

Uncertainty quantification

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$

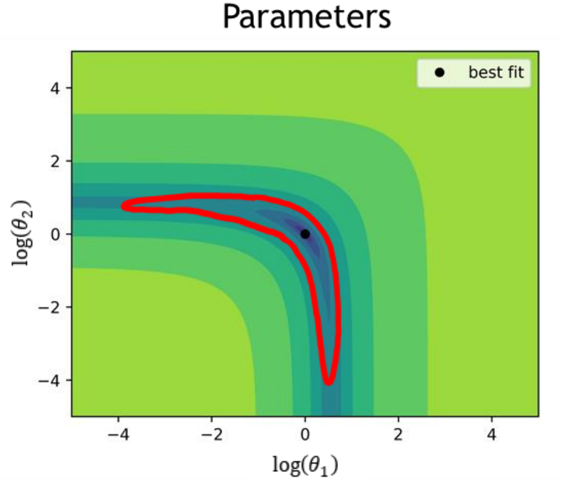


Assumptions:

$$d_m = y(t_m;\theta) + \xi_m$$
$$\xi_m \sim \mathcal{N}(0, \sigma_m)$$

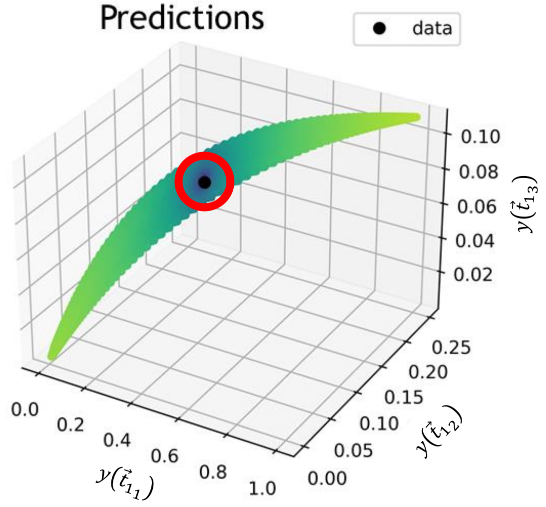
Loss function:

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$



Scenario 1 $\left(y(\vec{t}_1;\theta)\right)$

(Scenario 1)$^{-1}$

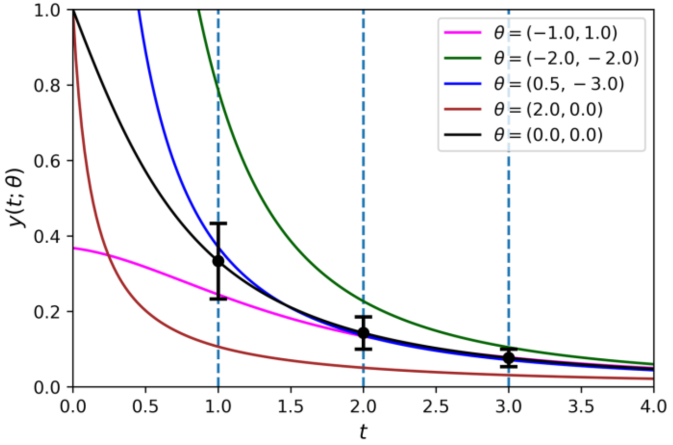


Distribution of data

Distribution of parameters

Uncertainty quantification

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$

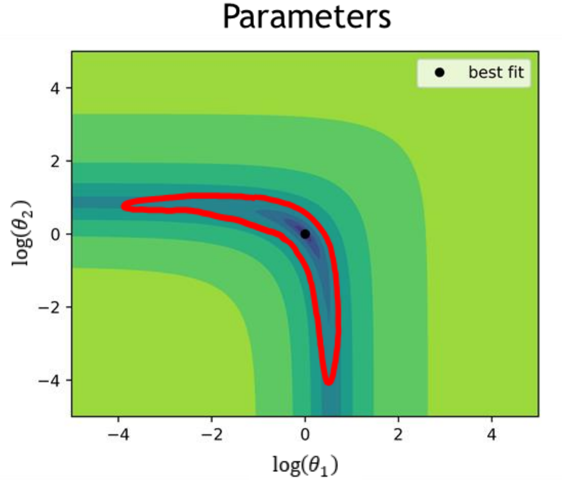


Assumptions:

$$d_m = y(t_m;\theta) + \xi_m$$
$$\xi_m \sim \mathcal{N}(0, \sigma_m)$$

Loss function:

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$



Distribution of data

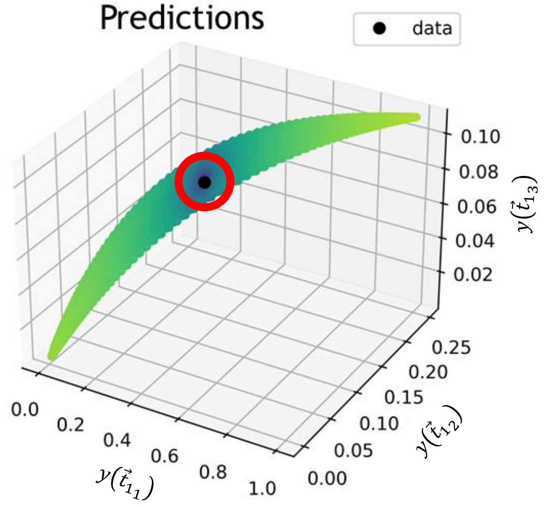
Distribution of parameters

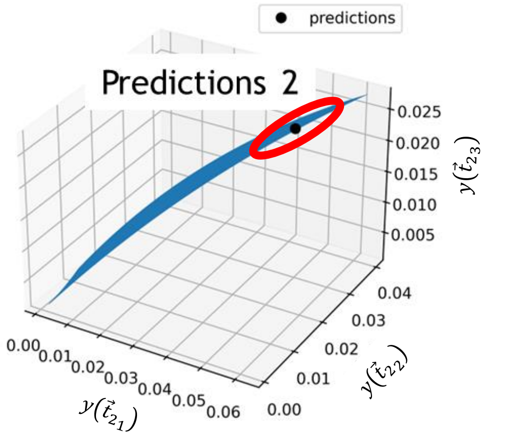
Distribution of predictions

Scenario 1 $\left(y(\vec{t}_1;\theta)\right)$

(Scenario 1)$^{-1}$

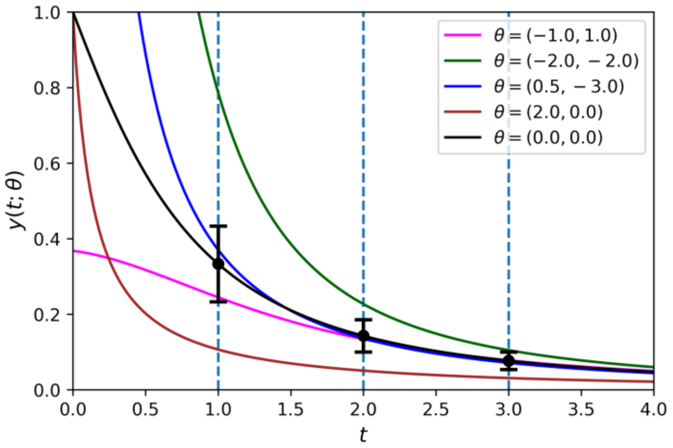
Scenario 2 $\left(y(\vec{t}_2;\theta)\right)$





Markov Chain Monte Carlo

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$

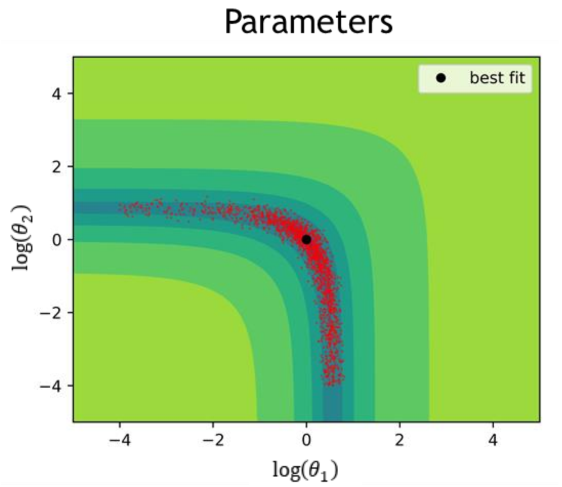


Assumptions:

$$d_m = y(t_m;\theta) + \xi_m$$
$$\xi_m \sim \mathcal{N}(0, \sigma_m)$$

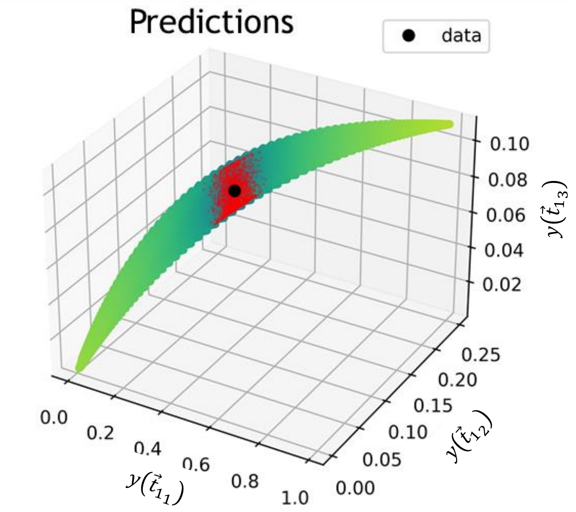
Loss function:

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$



Scenario 1 $\left(y(\vec{t}_1;\theta)\right)$

(Scenario 1)$^{-1}$



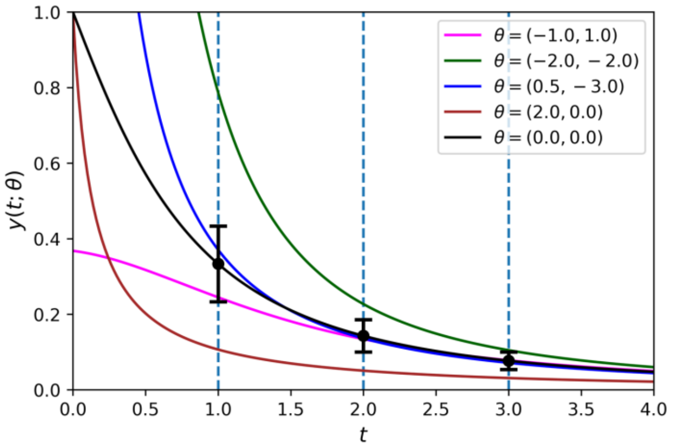
- Bayes' rule:

$$P(\theta|\vec{d}) \propto \mathcal{L}(\theta|\vec{d}) \times \pi(\theta),$$
$$\mathcal{L}(\theta|\vec{d}) \propto \exp(-L(\theta))$$

- Use MCMC algorithm to sample the posterior $P(\theta|\vec{d})$.

Markov Chain Monte Carlo

Model: $y(t;\theta) = \dfrac{1}{t^2 + \theta_1 t + \theta_2}$

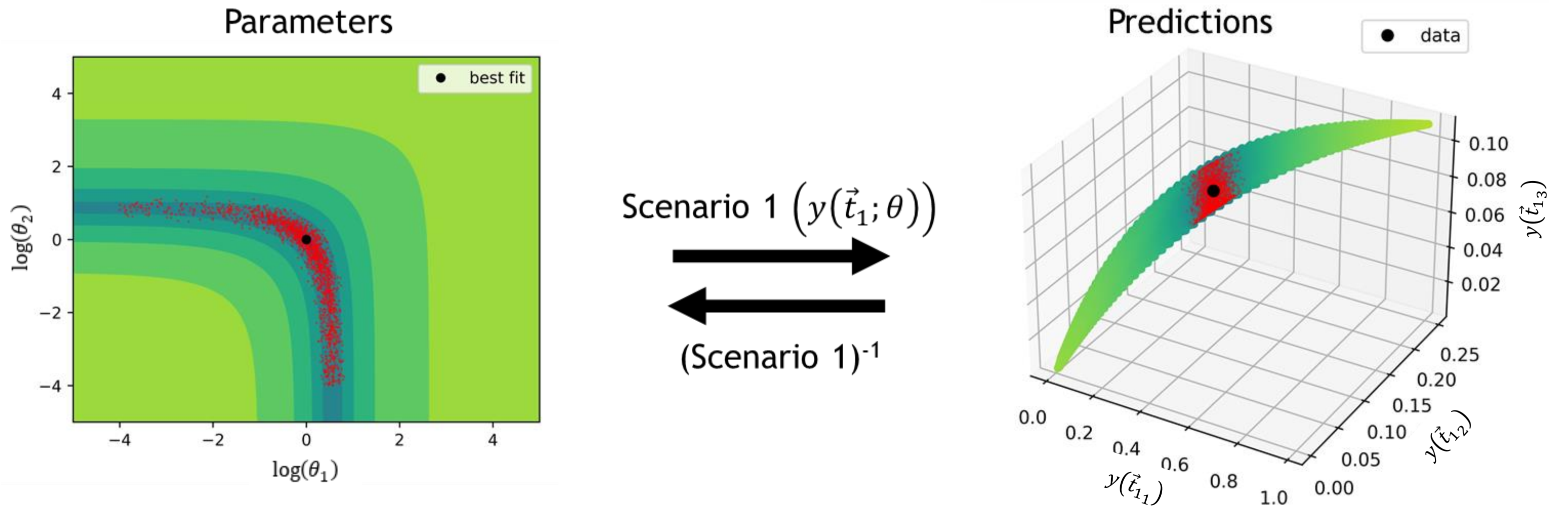


Assumptions:

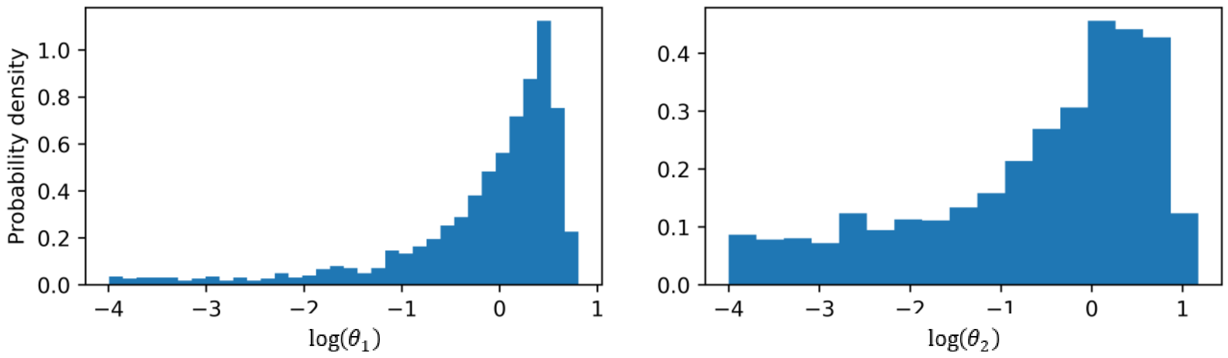
$$d_m = y(t_m;\theta) + \xi_m$$
$$\xi_m \sim \mathcal{N}(0, \sigma_m)$$

Loss function:

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$

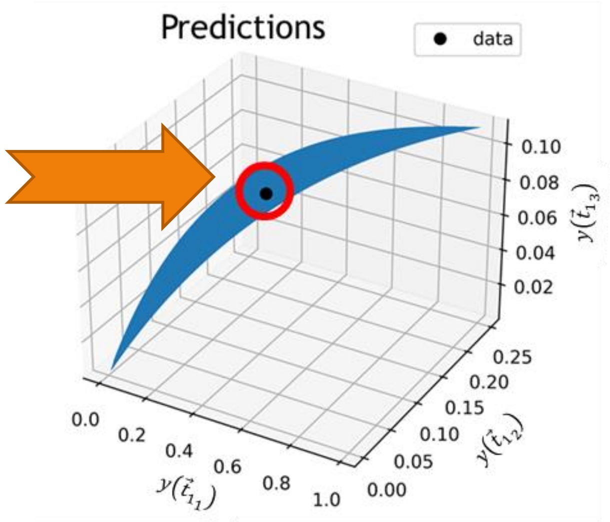


- Distribution of the parameters is inferred from the resulting samples.



Model inadequacy

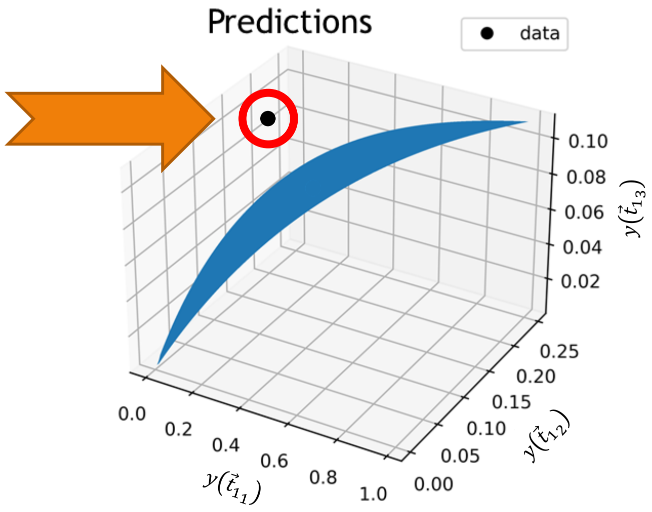
- $\mathcal{L}(\theta|\vec{d}) \propto \exp(-L(\theta))$ assumes the model can reproduce the data within the error bar.
- The high-density circle/sphere intersects the manifold.



Adequate model

Model inadequacy

- In some cases, this assumption is invalid.
- The data is far from the manifold; the high-density circle/sphere doesn't intersect the manifold.
- We need to fix the UQ formulation to include model inadequacy.



Inadequate model

Model inadequacy

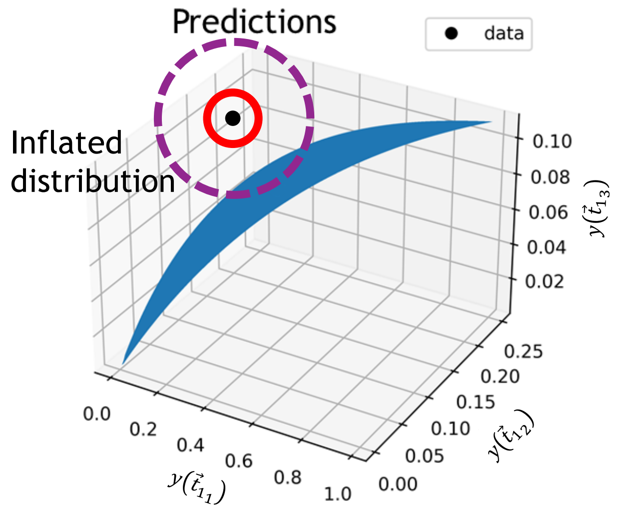
- Suggestion: Inflate the likelihood [2]:

$$\mathcal{L}(\theta|\vec{d}) \propto \exp\left(-\frac{L(\theta)}{T_0}\right), \qquad T_0 = \frac{2L_0}{N}$$

$$L(\theta) = \frac{1}{2}\sum_m \left(\frac{d_m - y(t_m;\theta)}{\sigma_m}\right)^2$$

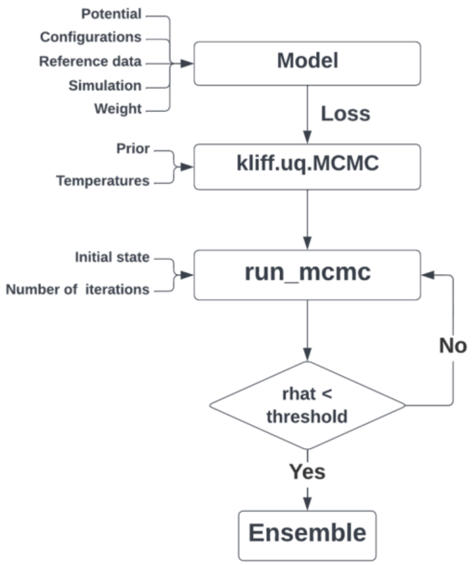
$L_0 \equiv$ minimum loss

$N \equiv$ number of parameters.



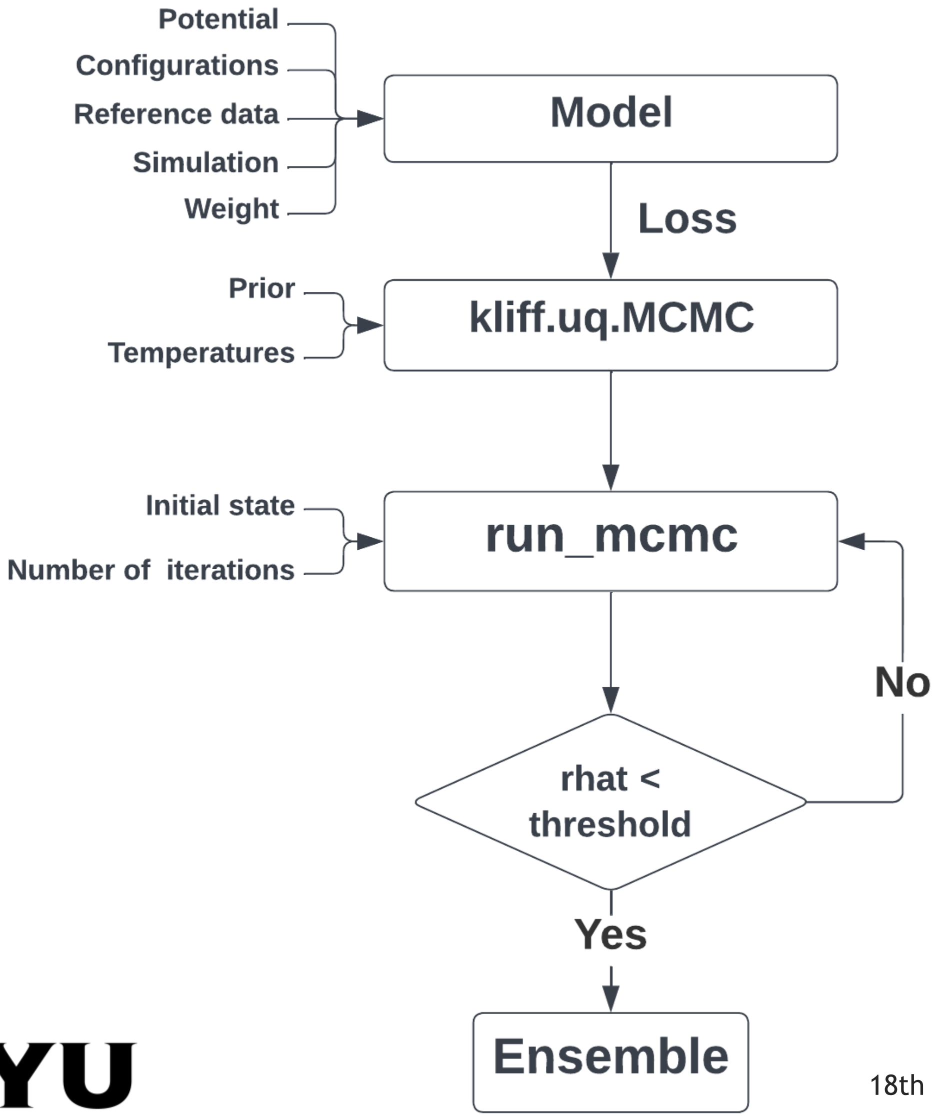
Inadequate model

[2] P. Pernot and F. Cailliez, "A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy," *AIChE Journal*, vol. 63, no. 10, pp. 4642-4665, 2017, doi: 10.1002/aic.15781.



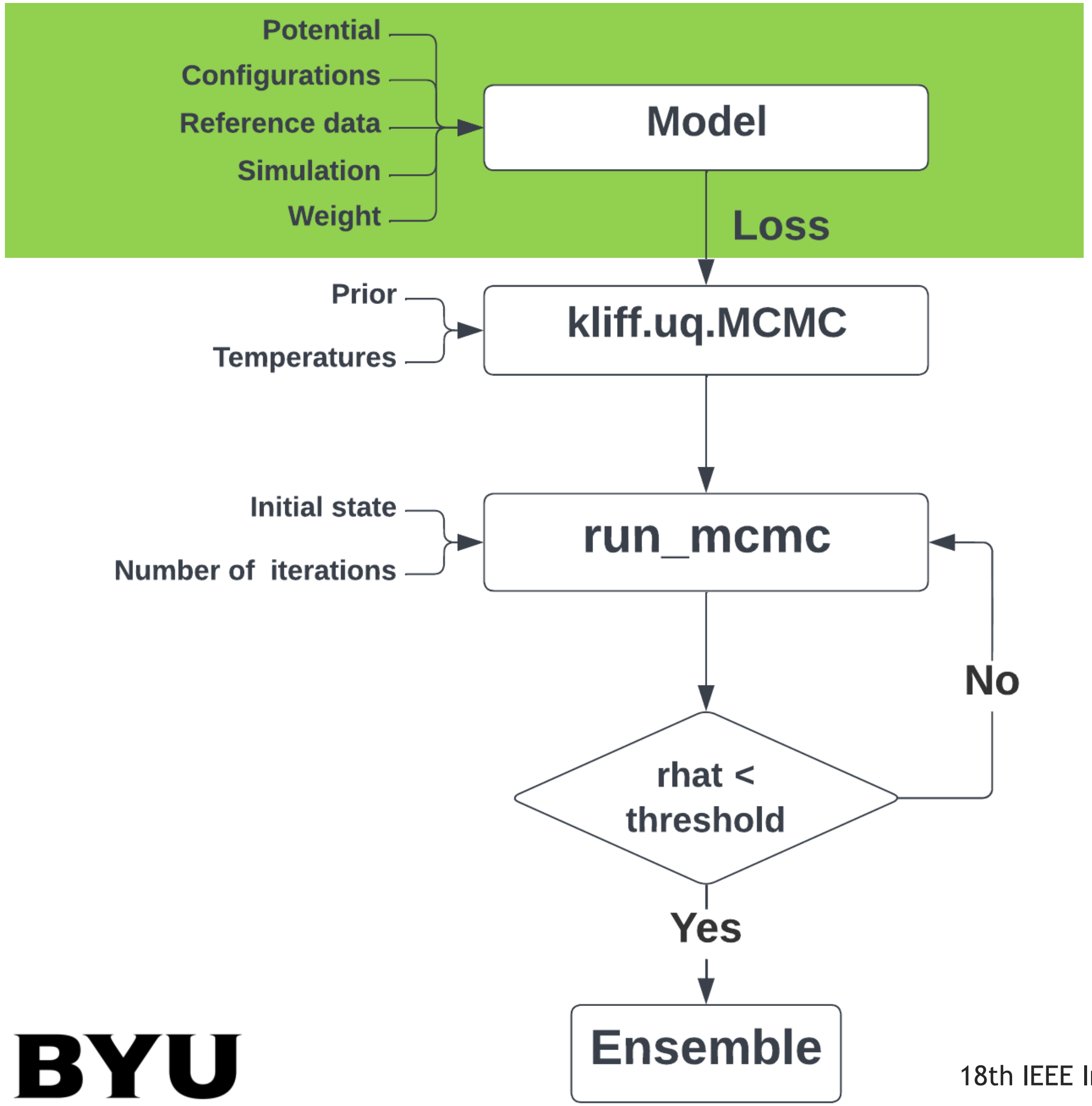
UQ extension to KLIFF

Implementation and workflow



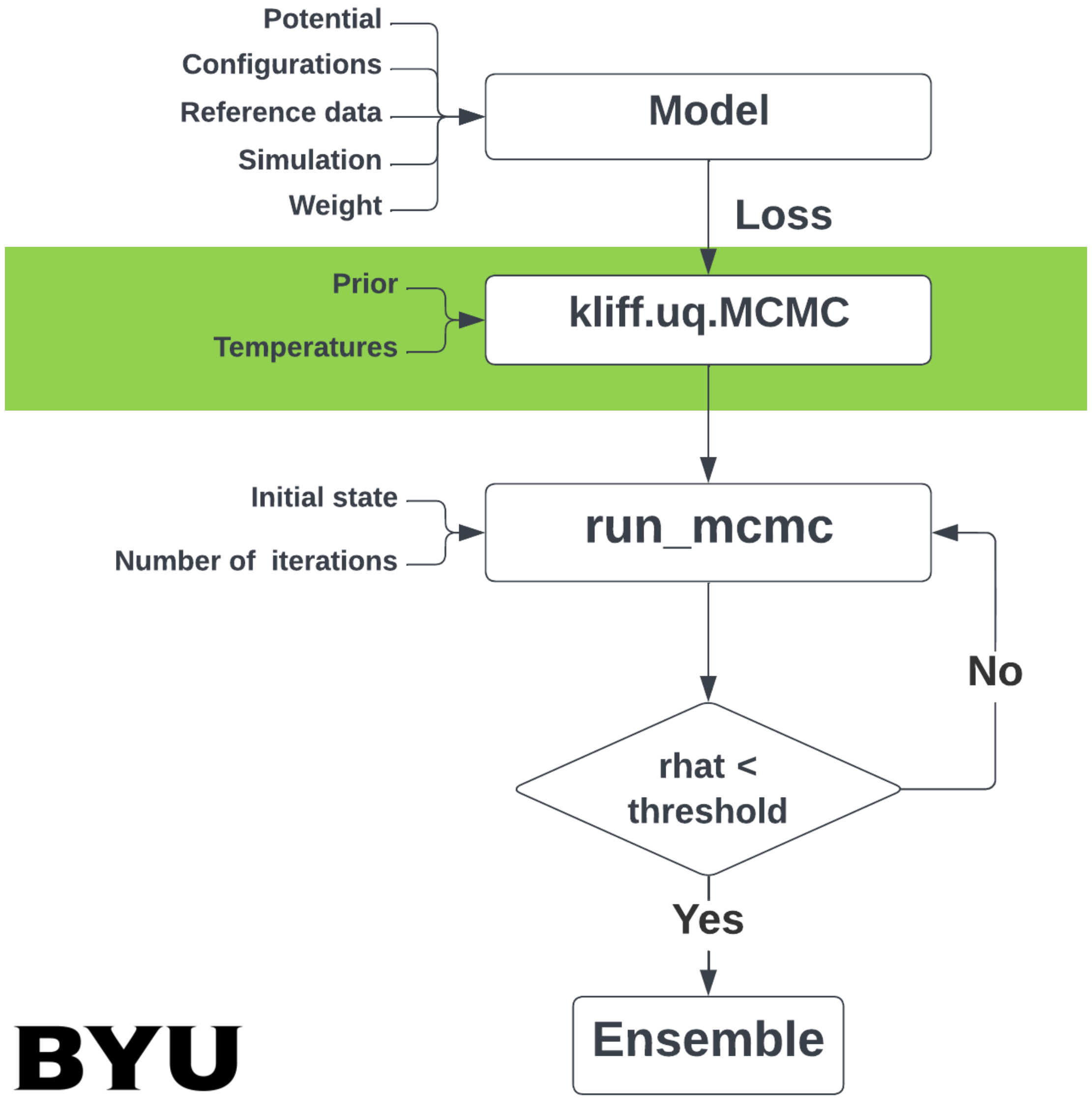
- We extend KLIFF to include uncertainty quantification functionality.
- This integration can:
  - Facilitate UQ studies for IPs.
  - Lead to more transparent and reproducible UQ analysis for IPs.
- KLIFF uses MCMC method.
- Other UQ methods will be implemented in the future.

1. Defining the model and loss function



- This functionality has been implemented previously and is not part of this integration.
- For more detail, visit https://kliff.readthedocs.io/.

2. Instantiating the posterior sampler

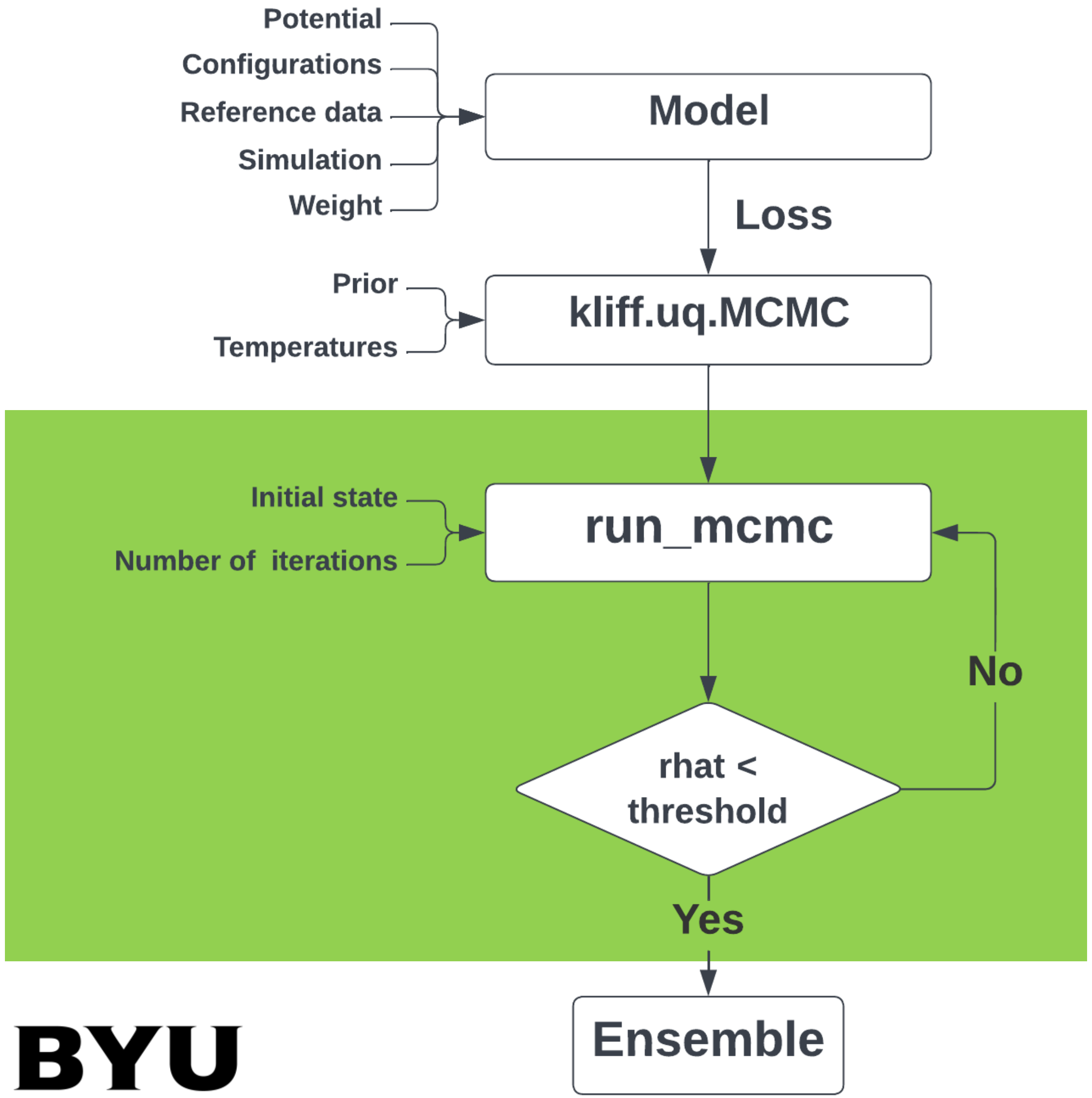


- We use ptemcee [3, 4] to perform parallel-tempered MCMC:
  - Simulating multiple different sampling temperatures, each with multiple chains/walkers.
- Parallel tempering improves convergence.
- Parallel tempering also allows us to explore how sampling results evolve with different scale of model error.
- Recommendation: Set the temperature ladder to be logarithmically spaced from 1.0 to few times larger than T_0.

[3] W. Vousden, “Willvousden/ptemcee: A parallel-tempered version of emcee.,” *GitHub*. [Online]. Available: https://github.com/willvousden/ptemcee. [Accessed: 14-Sep-2022].
[4] W. D. Vousden, W. M. Farr, and I. Mandel, “Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations,” *Monthly Notices of the Royal Astronomical Society*, vol. 455, no. 2, pp. 1919–1937, Jan. 2016, doi: 10.1093/mnras/stv2422.

BYU

3. Running MCMC & monitoring convergence



- Convergence is monitored by computing the Potential Scale Reduction Factor (PSRF) [5]:

$$\hat{R}^p = \frac{K-1}{K} + \frac{J+1}{J} \lambda_{max} \left(W^{-1} B/K\right)$$

$$\frac{B}{K} = \frac{1}{J-1} \sum_{j=1}^{J} (\bar{\psi}_j - \bar{\psi})(\bar{\psi}_j - \bar{\psi})^T$$

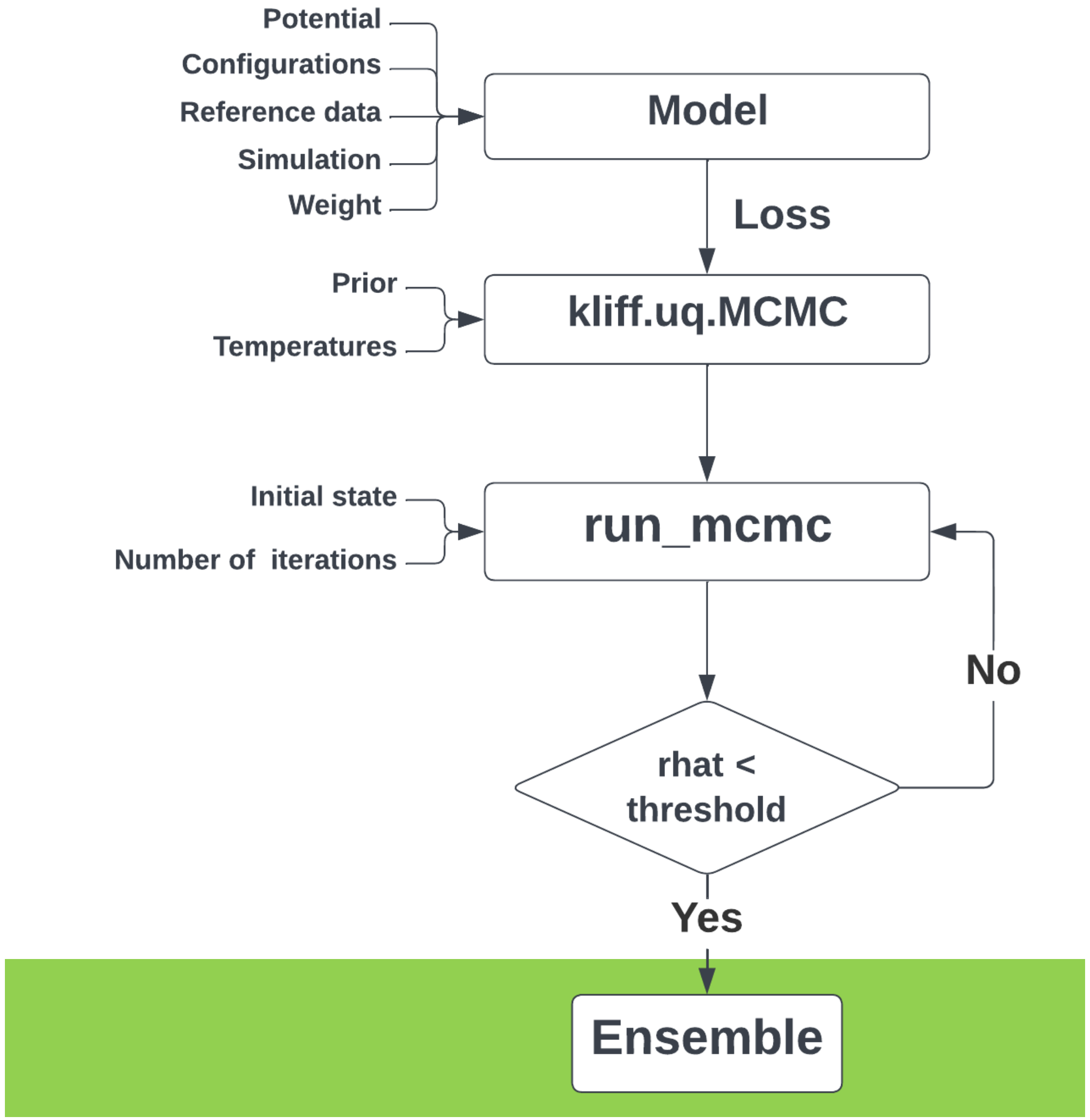
$$W = \frac{1}{J(K-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} (\psi_{jk} - \bar{\psi}_j)(\psi_{jk} - \bar{\psi}_j)^T$$

- When samples have converged, $\hat{R}^p \rightarrow 1$ (common threshold is 1.1).

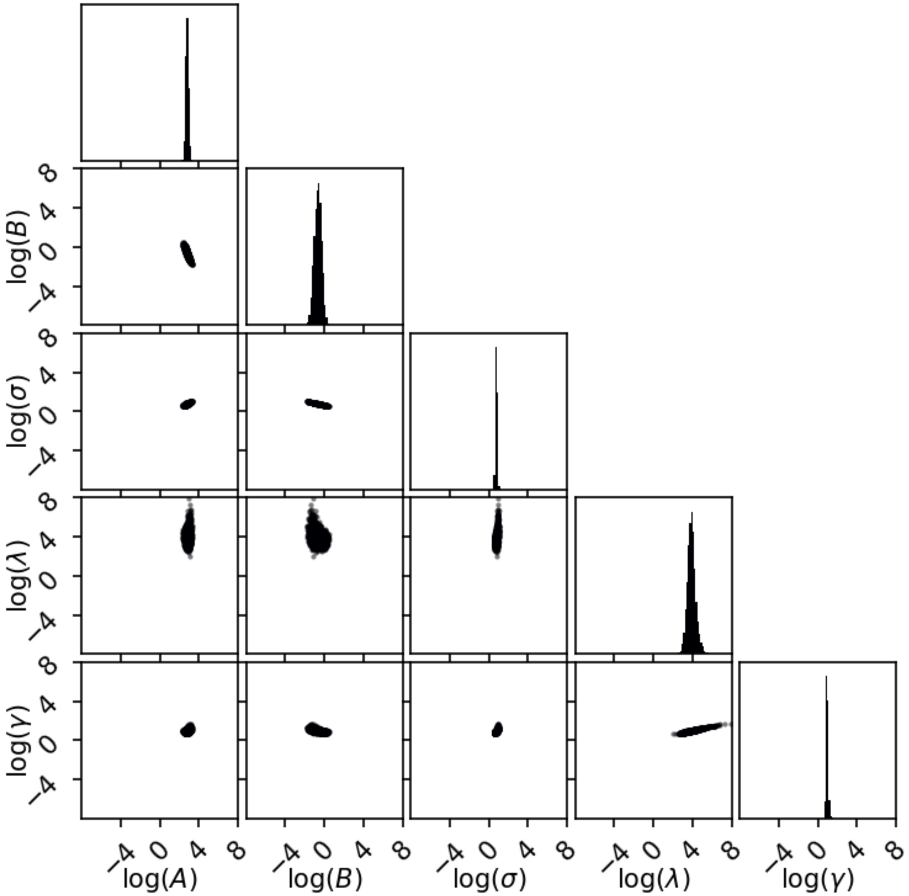
[5] S. P. Brooks and A. Gelman, "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 434–455, Dec. 1998, doi: 10.1080/10618600.1998.10474787.

BYU

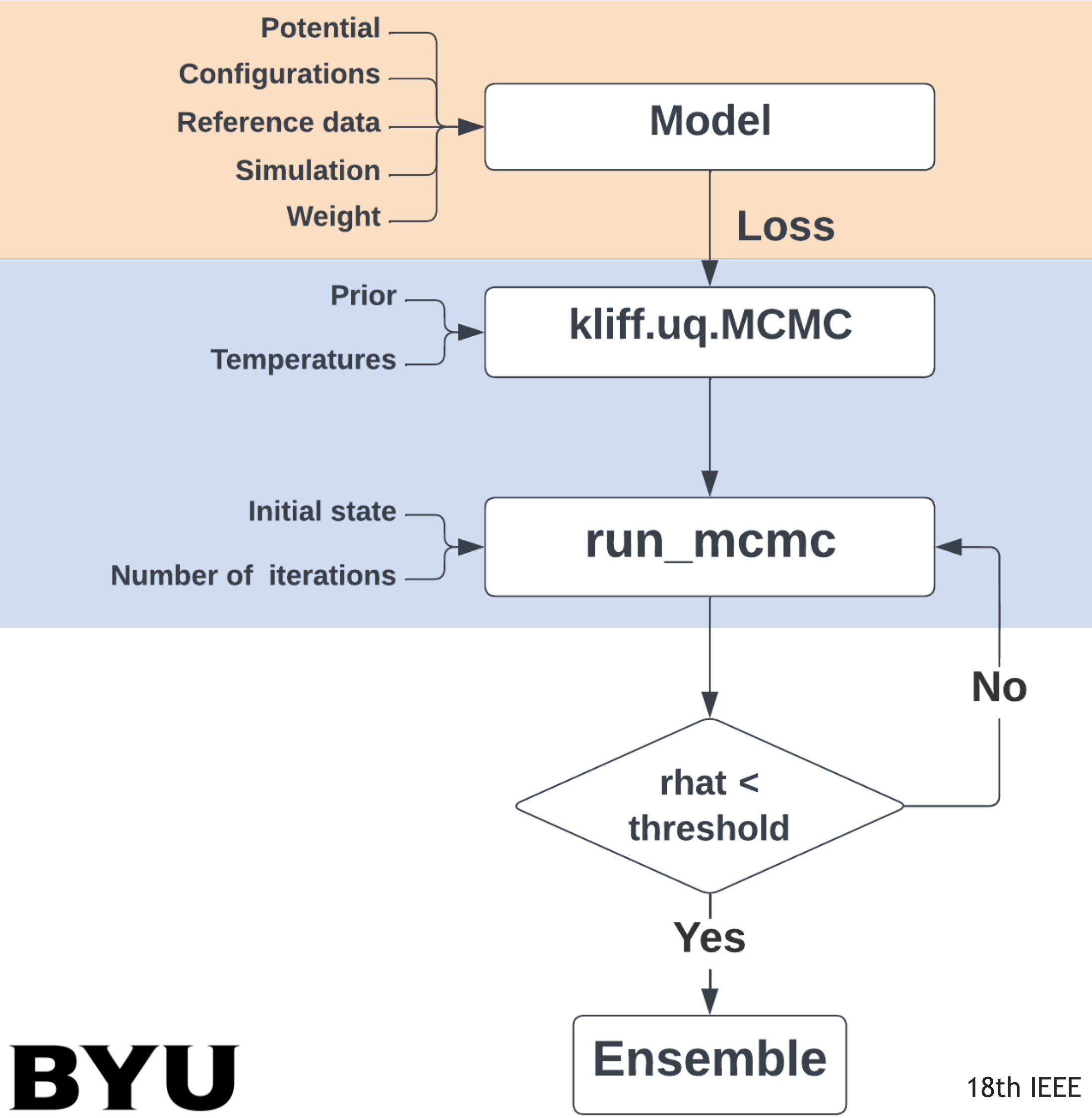
4. Retrieving the ensemble



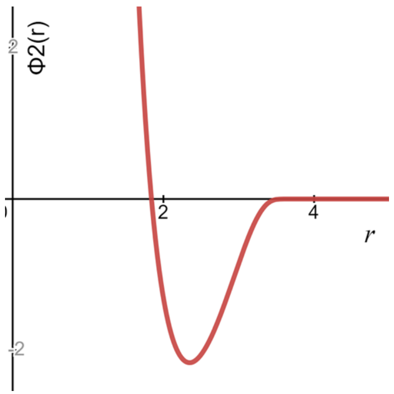
- The distribution of the parameters is inferred from the ensemble.

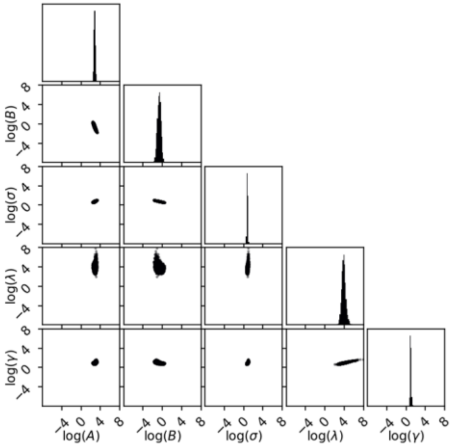


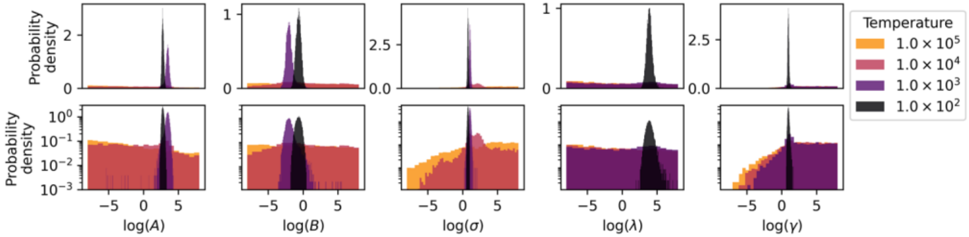
Parallel computing



- Parallelization can be done in 2 places:
  - Loss evaluation (over configurations)
  - MCMC sampling (over walkers)
- Suggestion:
  - Use OpenMP-style parallelization for loss evaluation.
  - Use MPI-style parallelization for MCMC sampling.
- For more detail, visit https://kliff.readthedocs.io/.







Demonstration: Study of Stillinger-Weber potential

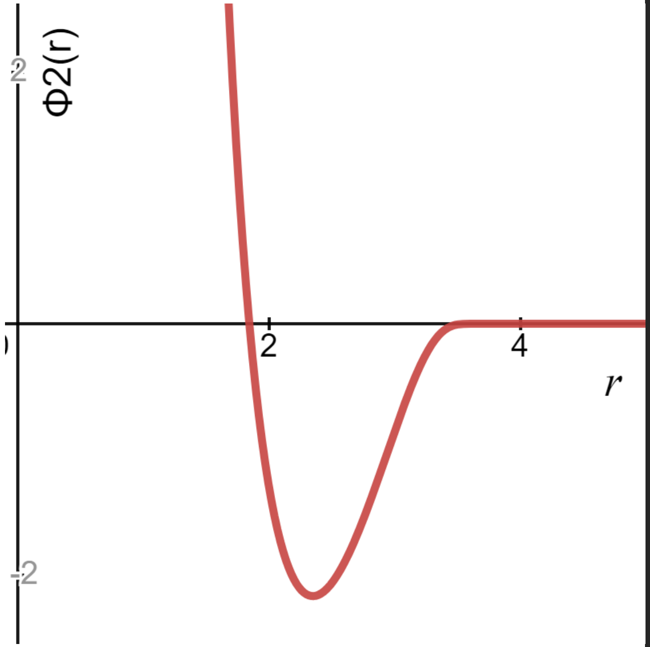
Stillinger-Weber potential

Access to the example scripts.

- Model: Stillinger-Weber potential [6]

$$\phi_2(r_{ij}) = A\left[B\left(\frac{\sigma}{r_{ij}}\right)^p - \left(\frac{\sigma}{r_{ij}}\right)^q\right]\exp\left(\frac{\sigma}{r_{ij} - r^{cut}}\right)$$

$$\phi_3(r_{ij}, r_{ik}, \beta_{jik}) = \lambda\left[\cos(\beta_{jik}) - \cos(\beta^0)\right]^2 \times \exp\left(\frac{\gamma}{r_{ij} - r^{cut}} + \frac{\gamma}{r_{ik} - r^{cut}}\right)$$



- Parameters: $\log(A), \log(B), \log(\sigma), \log(\lambda), \log(\gamma)$
- Training data: energy and force of Silicon atoms in several configurations (weights ∝ data values).
- Best fit:

$A = 15.2792223$ eV
$B = 0.6032372$
$\sigma = 2.09420085$ Å

$\lambda = 45.47927476$ eV
$\gamma = 2.51306949$ Å

$p = 4$
$q = 0$
$\cos(\beta^0) = -0.33333333$
$r^{cut} = 3.77118$ Å

BYU

[6] A. K. Singh, F. H. Stillinger, and T. A. Weber, "Stillinger-Weber potential for Si due to Stillinger and Weber (1985) v006," OpenKIM, https://doi. org/10.25950/dd263fe3, 2021.

MCMC setup

Access to the example scripts.

MCMC Setup

- Posterior distribution:

$$P(\theta|\vec{d}) \propto \mathcal{L}(\theta|\vec{d}) \times \pi(\theta),$$

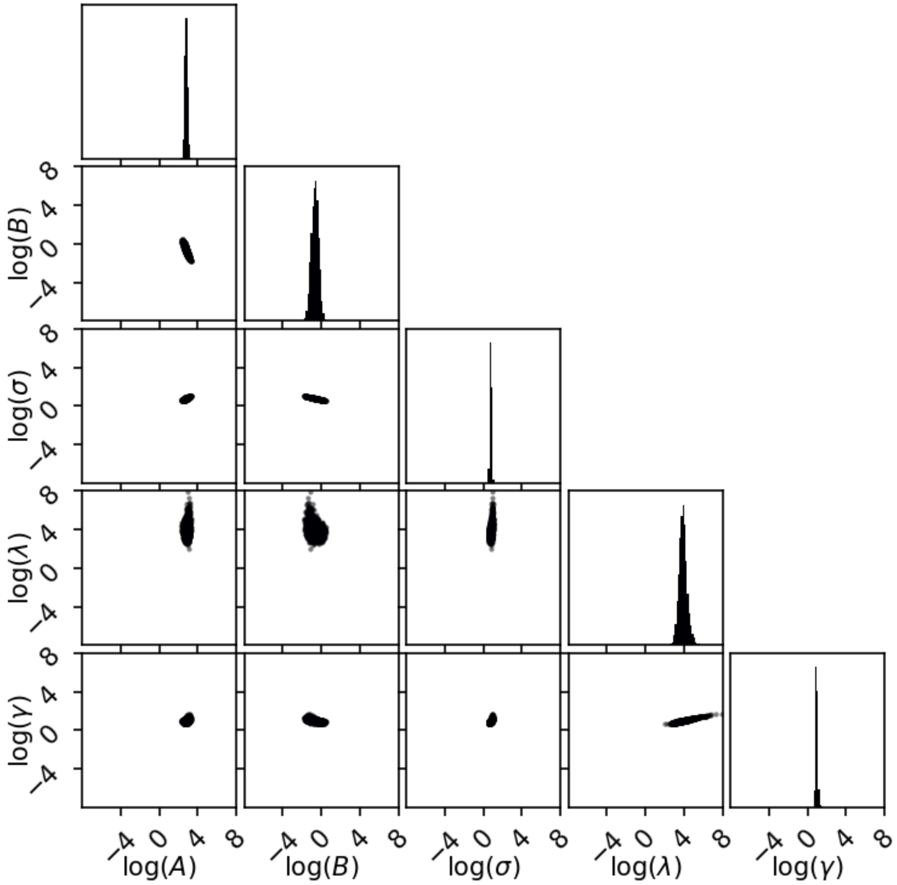
$$\mathcal{L}(\theta|\vec{d}) \propto \exp(-L(\theta)/T)$$

- Temperatures:
  - $T_0 = 1.324$
  - $T \in [1, 10^7]$
- Prior: $\log(\theta) \sim \mathcal{U}(-8, 8)$
- Run MCMC for $150{,}000$ steps
  - Burn-in: $10{,}000$
  - Thinning factor: 200
- Convergence test: $\hat{R}^p \leq 1.046$

BYU

Presenting the samples

Access to the example scripts.

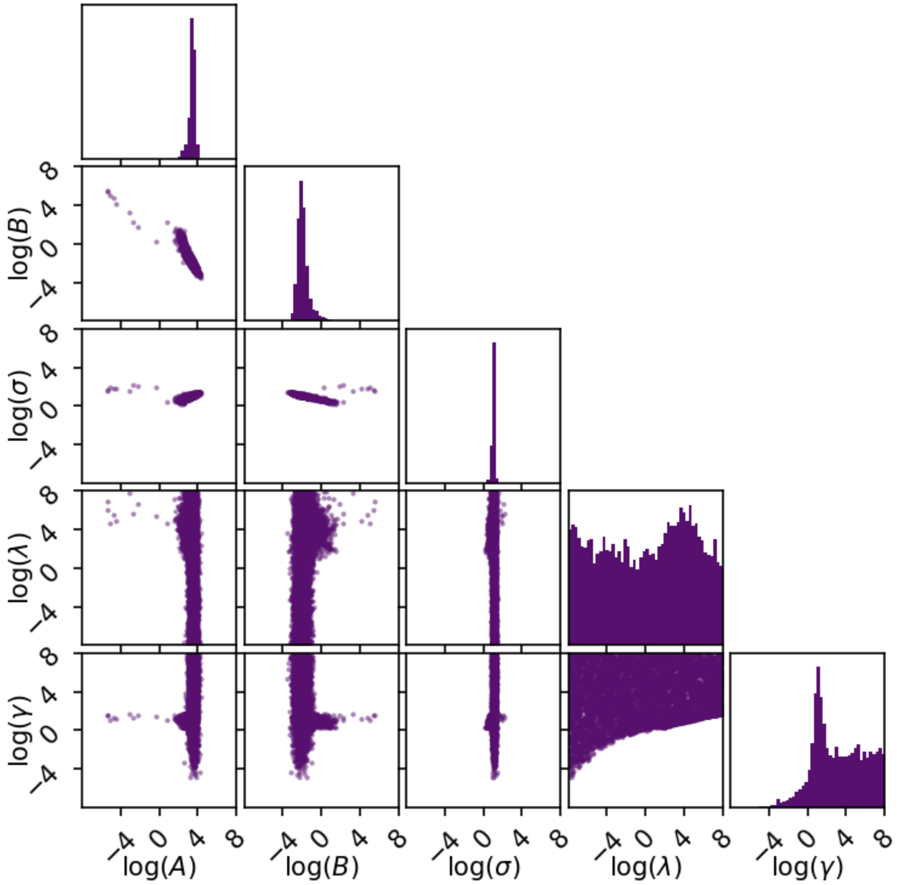


- Sampling temperature $T = 10^2$
- What's plotted:
  - Main diagonal: Marginal distribution for each parameter.
  - Below diagonal: 2D projection of the samples in parameter space.
- At lower temperature, the distributions are concentrated.

BYU

Parameter evaporation

Access to the example scripts.



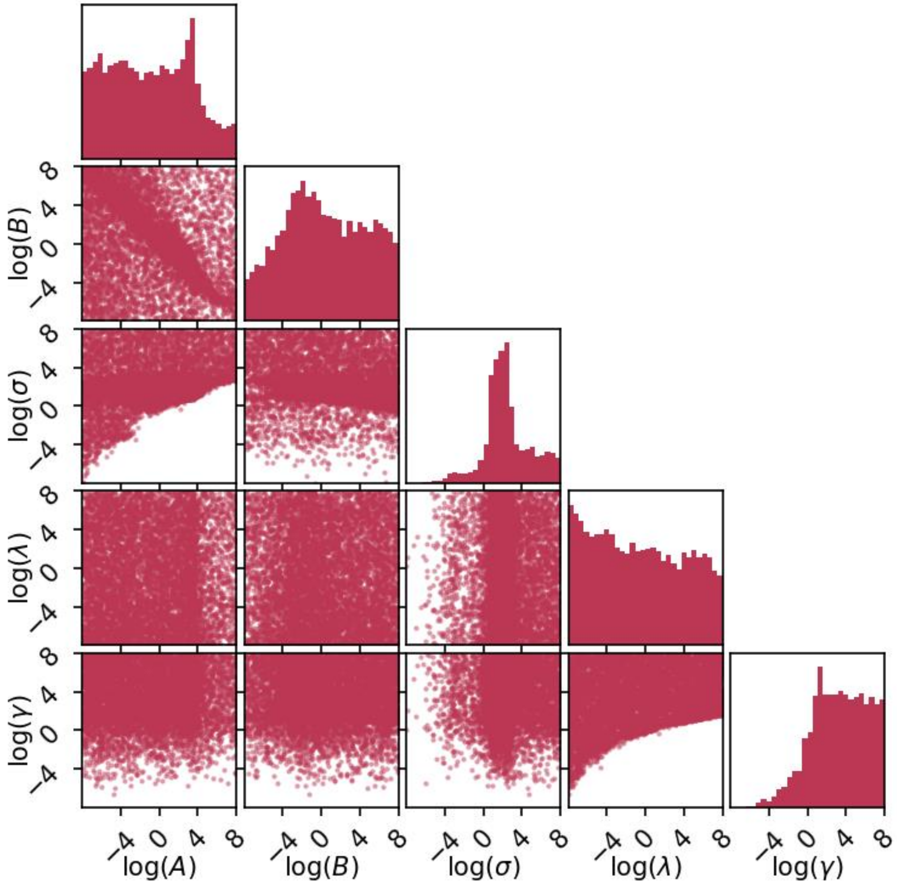
- Sampling temperature $T = 10^3$
- The distribution becomes wider as we increase the temperature.
- Parameter evaporation occurs: the walkers tend to run to sub-optimal parameter values [7, 8].
- Evaporated parameters are unconstrained by the data.

[7] M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Geometry of nonlinear least squares with applications to sloppy models and optimization," *Phys. Rev. E*, vol. 83, no. 3, p. 036701, Mar. 2011, doi: 10.1103/PhysRevE.83.036701.
[8] R. Gutenkunst, "Sloppiness, Modeling, and Evolution in Biochemical Networks," Cornell University, Ithaca, New York, 2007. Accessed: May 14, 2021. [Online]. Available: https://ecommons.cornell.edu/handle/1813/8206

BYU

Access to the example scripts.

Parameter evaporation



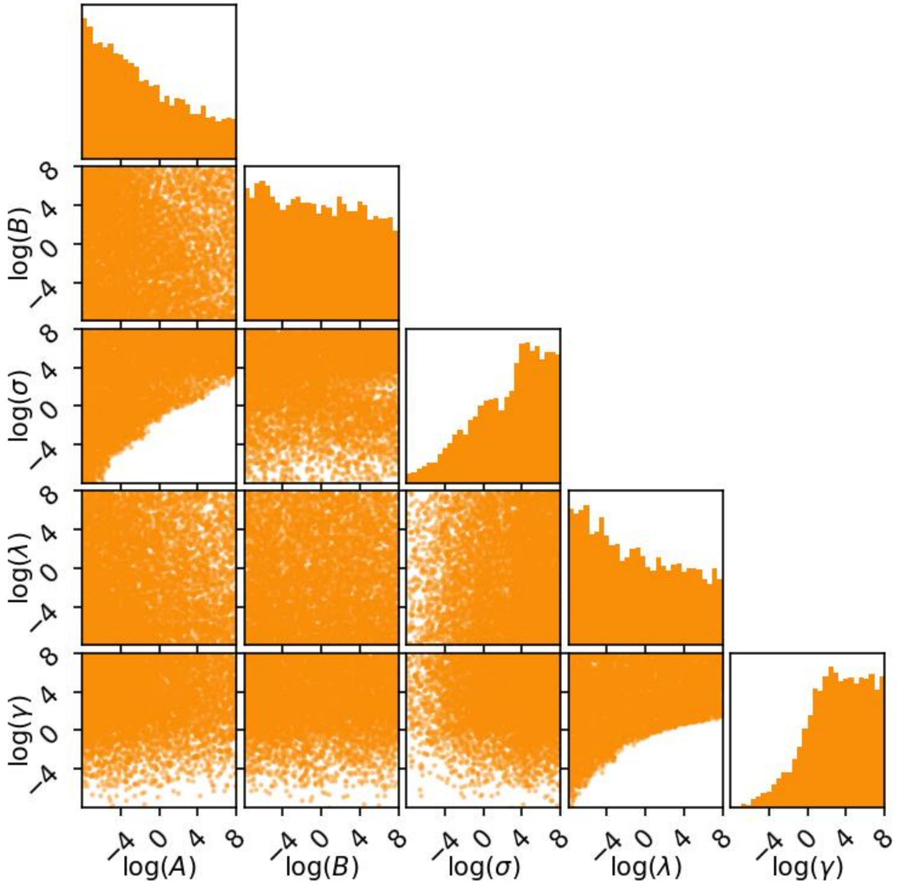
- Sampling temperature $T = 10^4$
- The distribution becomes wider as we increase the temperature.
- Parameter evaporation occurs: the walkers tend to run to sub-optimal parameter values [7, 8].
- Evaporated parameters are unconstrained by the data.
- Parameter evaporation becomes more apparent at higher temperatures.

[7] M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Geometry of nonlinear least squares with applications to sloppy models and optimization," *Phys. Rev. E*, vol. 83, no. 3, p. 036701, Mar. 2011, doi: 10.1103/PhysRevE.83.036701.
[8] R. Gutenkunst, "Sloppiness, Modeling, and Evolution in Biochemical Networks," Cornell University, Ithaca, New York, 2007. Accessed: May 14, 2021. [Online]. Available: https://ecommons.cornell.edu/handle/1813/8206

BYU

Access to the example scripts.

Parameter evaporation



- Sampling temperature $T = 10^5$
- The distribution becomes wider as we increase the temperature.
- Parameter evaporation occurs: the walkers tend to run to sub-optimal parameter values [7, 8].
- Evaporated parameters are unconstrained by the data.
- Parameter evaporation becomes more apparent at higher temperatures.
- We can use this result as a guide to collect more training data.

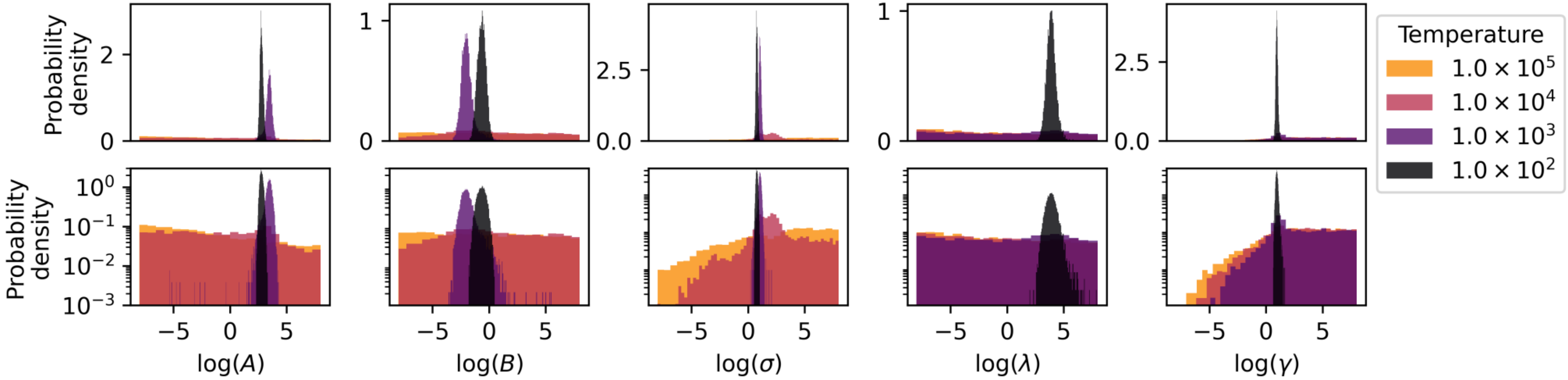
[7] M. K. Transtrum, B. B. Machta, and J. P. Sethna, “Geometry of nonlinear least squares with applications to sloppy models and optimization,” *Phys. Rev. E*, vol. 83, no. 3, p. 036701, Mar. 2011, doi: 10.1103/PhysRevE.83.036701.
[8] R. Gutenkunst, “Sloppiness, Modeling, and Evolution in Biochemical Networks,” Cornell University, Ithaca, New York, 2007. Accessed: May 14, 2021. [Online]. Available: https://ecommons.cornell.edu/handle/1813/8206

BYU

Access to the example scripts.

Comparison of the marginal distributions

Marginal distribution of the parameters at several sampling temperatures



- Compare the distributions at $T = 10^2$ and $T = 10^3$:
  - λ and γ evaporate.
  - The expectation value of A, B, and σ are shifted away from the best fit.
- How we should treat parameter evaporation is an open question.

BYU

Conclusion

- We enhance KLIFF with UQ framework.
- This implementation can facilitate more UQ studies and lead to more transparent and reproducible UQ analysis for IPs.
- We demonstrate it to study SW potential for silicon system.
- The result indicates parameter evaporation.
  - The data cannot constrain the evaporated parameters and future predictions.
  - The sampling result is highly dependent on the sampling temperature and prior.

- Suggestions:
  - Check for robustness of the result to several choice of prior.
  - Use the result to inform what other training data are needed.

- Future work:
  - Integrate other UQ methods.
  - Work on accelerating MCMC.

Access to the example scripts.

Acknowledgement

- This work has been funded by the NSF under grant CMMT-1834332
- OpenKIM
- Collaborators:
  - Ellad B. Tadmor
  - Ryan S. Elliott
  - Daniel S. Karls
  - Mingjian Wen
- Contact: kurniawanyo@outlook.com

OpenKIM

Access to the example scripts.