A Bayesian Inference Approach to Uncertainty Quantification for Density Functional Theory

Kate Fisher, Michael Herbst, & Youssef Marzouk

June 22, 2022

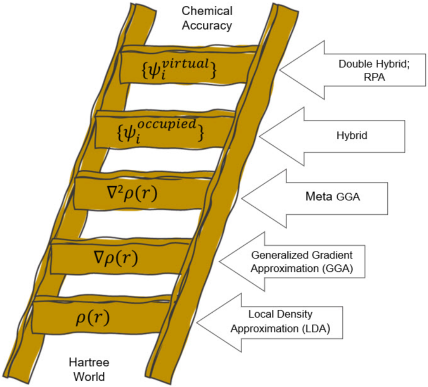
Sources of Uncertainty in DFT

Introduction

- Density Functional Theory (DFT) used as reference for molecular dynamics simulations
- Accuracy depends on chemical system, quantity of interest, and functional choice

Plan: Design a Bayesian Inference model to infer a distribution on an ensemble of DFT predictions using different approximations

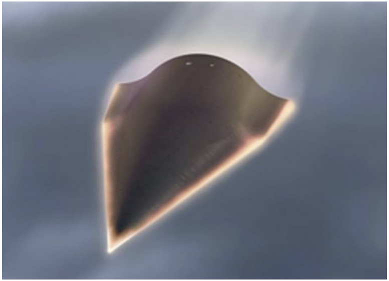
Exchange Correlation



- Kohn Sham DFT is exact, but the true exchange correlation functional, $E_{xc}[\rho]$, is unknown
- There are many approximations to $E_{xc}[\rho]$ with a range of accuracy

Long Term Applications

- Multiscale modelling of materials in extreme environments
  - Uncertainty will be be propagated to a larger scale to inform molecular dynamics simulations



- Functional Approximation design
- Multifidelity DFT predictions
  - determine the best subset of functionals and their relative accuracy
  - indicate when a high rung functional approximation is necessary

Existing Approaches to UQ for DFT

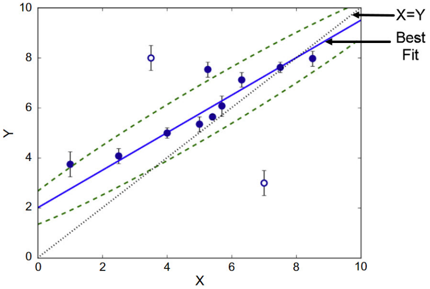
Regression

One approach to error estimation in DFT [Lejaeghere, 2020]:

Experimental Data

$$Y = a + b X + \varepsilon$$

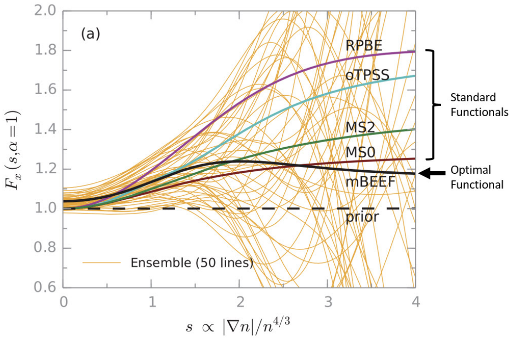
DFT Predictions



- Use a linear fit to separate predictable error (a and b) from “random” error (ε)

Bayesian Error Estimation Functionals (BEEF)

Error representation via functional ensemble [Christensen et al., Wellendorff et al., 2020]:



- Fit an optimal functional using databases
- Create an ensemble with $\sigma^2 \approx$ error of the functional against the data

Bayesian Inference Approach

Bayesian Modelling

Consider a chemical system, Y, and some quantity of interest (i.e. atomization energy) with unknown true value ν.

- Assumption:
  - Experimental measurements and theoretical predictions are distributed around ν in some pattern that can be represented by a statistical model
- Approach:
  - Relate the data to ν with statistical model
  - Obtain probability distribution for ν

Our Approach

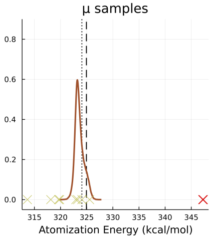
We will adapt a method used by Tebaldi et al. [2005, 2009] for UQ in climate models. The idea is to

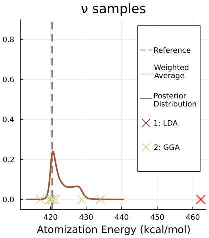
- Use predictions by multiple functionals to infer a distribution on a Quantity of Interest
- Leverage cases where high level theory is available
- Based on the spread of DFT predictions around the high level data for chemical compound X, infer a distribution on predictions for chemical compound Y

Preliminary Results

System X System Y

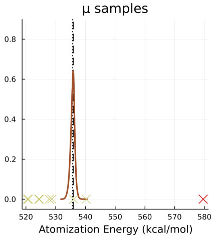
Case 1:

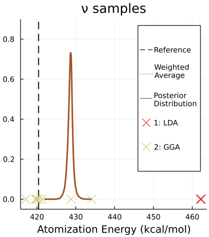




The model has some promising behavior...

Case 2:





...and limitations

Bayes' Law

Posterior Distribution

Likelihood

Prior Distribution

$$\mathbb{P}(\text{Parameters}\big|\text{ Data}) \;\propto\; \mathbb{P}(\text{Data }\big|\text{Parameters})\;\mathbb{P}(\text{Parameters})$$

In our case, the data is

$$\begin{aligned} X_0 &\equiv \text{ Reference data for chemical system X} \\ X_j &\equiv \text{ DFT prediction by j for system X} \\ Y_j &\equiv \text{ DFT prediction by j for system Y} \end{aligned}$$

where $j \equiv$ Functional j

Components of a Simple Model

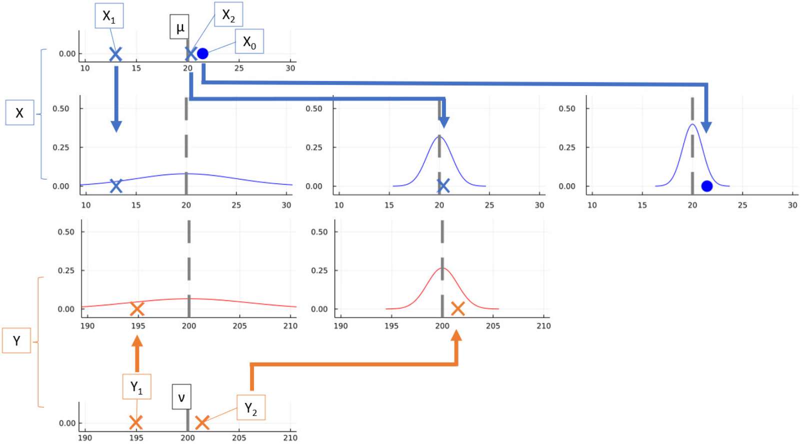
Likelihoods	$X_0 \sim \mathcal{N}(\mu, \lambda_0^{-1})$ $X_j \sim \mathcal{N}(\mu, \lambda_j{}^{-1})$ $Y_j \vert X_j \sim \mathcal{N}\left(\nu + \beta(X_j - \mu), (\phi\lambda_j)^{-1}\right)$
Priors	$\lambda_1, \; \ldots, \lambda_M \sim Ga(a_\lambda, \; b_\lambda)$ $\mu, \; \nu, \; \beta \sim$ constant, uninformative $\phi, \; a_\lambda, \; b_\lambda \sim Ga(a, \; b)$
Fixed	$a, \; b, \; \lambda_0^{-1}$

Interpretation of Parameters

Likelihoods	$X_0 \sim \mathcal{N}(\mu, \lambda_0^{-1})$ $X_j \sim \mathcal{N}(\mu, \lambda_j{}^{-1})$ $Y_j \vert X_j \sim \mathcal{N}\left(\nu + \beta(X_j - \mu), (\phi\lambda_j)^{-1}\right)$

- $\mu \rightarrow$ exact value of QOI for system X
- $\nu \rightarrow$ exact value of QOI for system Y
- $\lambda_j \rightarrow$ confidence in functional approximation j
- β, $\phi \rightarrow$ controls of correlation between X and Y

How does the model balance demands?

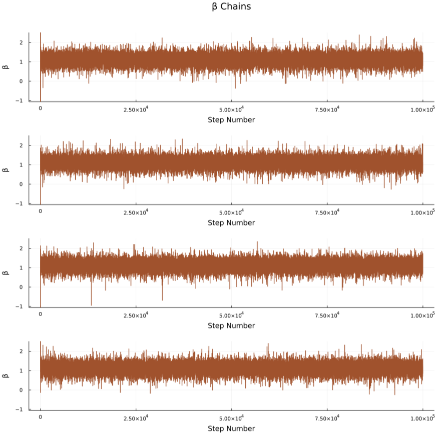


Possible Limitations

- Zero bias assumption
  - All predictions and experimental data are assumed to be centered on the exact value for the QOI
- Independence assumption
  - Functional approximations are assumed to be independently distributed about exact value
- Priors
  - There is some disagreement as to whether the Gamma prior is uninformative [Gelman, 2006]
- Simplicity of precision/confidence parameters
  - It is very likely the "best" functional approximation will be different for X and Y

Inference

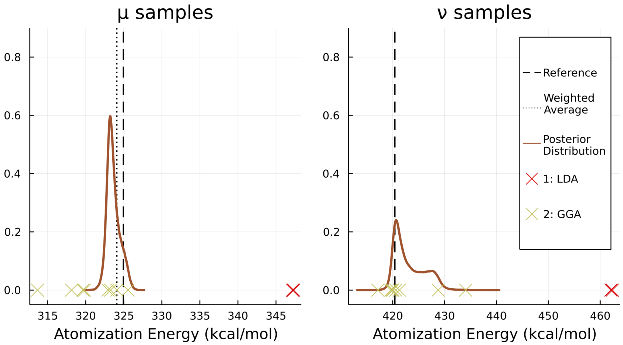
- The parameter set is small enough that posterior samples can be obtained using MCMC
  - Gibbs sampling is used for nearly all parameters
  - Exception: a_λ and b_λ are updated with Metropolis sampling



Results

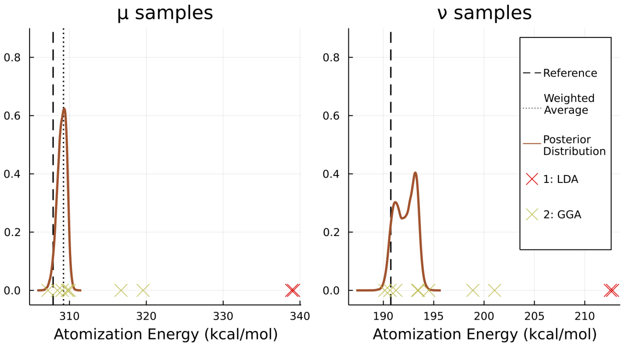
When the model works well...

X: SiH_4 (Saturated) Y: CH_4 (Saturated)



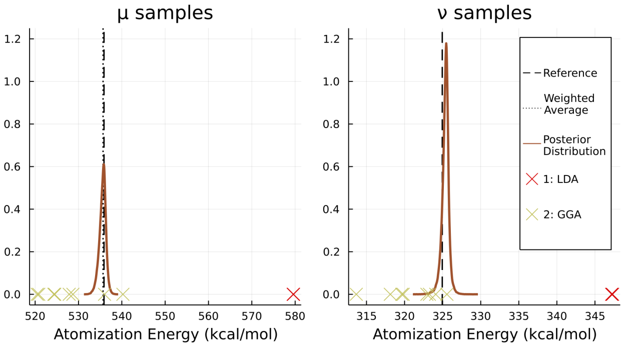
When the model works well...

X: CH_3 (Radical)   Y: CH_2 (Biradical)



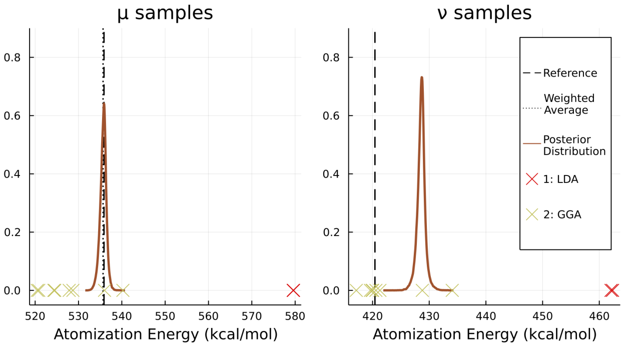
Overconfidence...

X: Si_2H_6 (Saturated) Y: SiH_4 (Saturated)



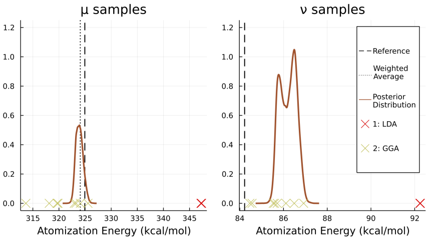
When the model is confidently wrong...

X: Si_2H_6 (Saturated) Y: CH_4 (Saturated)



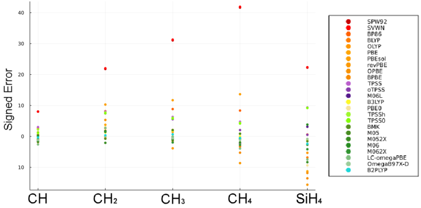
Misleading Data...

X: SiH_4
(Saturated)

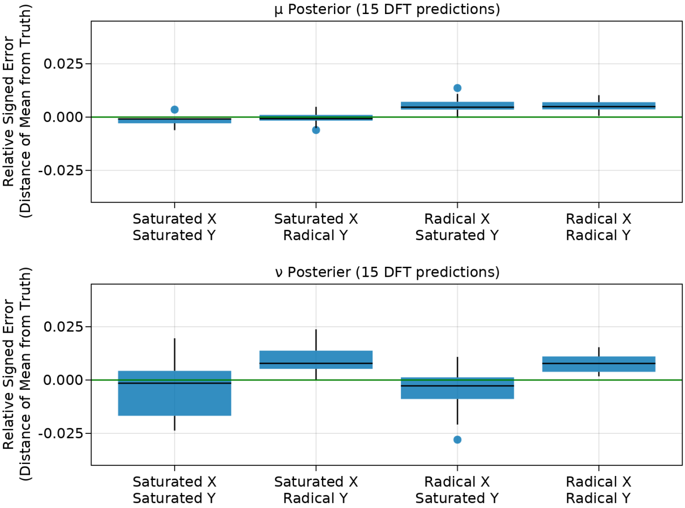


Y: CH
(Triradical)

Functional behavior for various chemical systems:



Compound Type and Error



Next Steps

Current and Future Work

The current model is limited by...

- The assumption that all DFT predictions are distributed with the same mean
  - **Plan:** We can adapt our parameter choice to capture bias in functional approximation classes
- A lack of procedure for checking the accuracy of the posterior mean and width
  - **Plan:** Develop a cross validation procedure to quantify inference success in the absence of reference data for Y
- Only a single point of reference (System X)
  - **Plan:** We can incorporate multiple reference systems and QOI into our inference model

References I

[1] R. Christensen, T. Bligaard, and K. Jacobsen.
Bayesian error estimation in density functional theory.
Uncertainty Quantification in Multiscale Materials Modeling, pages 77–91, 2020.

[2] A. Gelman.
Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).
Bayesian Analysis, 3(1):515–534, 2006.

[3] K. Lejaeghere.
The uncertainty pyramid for electronic-structure methods.
In Y. Wang and D. L. McDowell, editors, *Uncertainty Quantification in Multiscale Materials Modeling*, chapter 2, pages 41–76. Elsevier Ltd., Atlanta, 2020.

References II

[4] R. Smith, C. Tebaldi, D. Nychka, and L. Mearns.
Bayesian modeling of uncertainty in ensembles of climate models.
Journal of American Statistical Association, 104(485):97–116, 2009.

[5] C. Tebaldi and R. Knutti.
The use of the multi-model ensemble in probabilistic climate projections.
The Philosophical Transactions of Royal Society A, 365(1857):2053–2075, 2007.

[6] C. Tebaldi, R. Smith, D. Nychka, and L. Mearns.
Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles.
Journal of Climate, 18(10):1524–1540, 2005.

References III

[7] J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen, and T. Bligaard.
mbeef: An accurate semi-local bayesian error estimation density functional.
Uncertainty Quantification in Multiscale Materials Modeling, pages 77–91, 2020.

[8] L. Ying, J. Yu, and L. Ying.
Numerical methods for kohn–sham density functional theory.
Acta Numerica, 28:405–539, 2019.

Climate to Quantum

	X	Y
Climate Science	Current Temperature	Future Temperature
Quantum Chemistry	Reference Chemical Compound	Unknown Chemical Compound

⇒ Infer probability Distributions

Factorization

Let our set of parameters be $\boldsymbol{\theta}$.
With some assumptions about independence, we can factorize the likelihood and prior:

$$\begin{aligned}\mathbb{P}(\text{Data}\ \big|\boldsymbol{\theta}) &= \mathbb{P}(\mathbf{Y}\ \big|\mathbf{X}, X_0, \boldsymbol{\theta})\ \ \mathbb{P}(\mathbf{X}\ \big|X_0, \boldsymbol{\theta})\ \ \mathbb{P}(X_0\ \big|\boldsymbol{\theta}) \\ &= \prod_{j=1}^{M} \mathbb{P}(Y_j\ \big|X_j, \boldsymbol{\theta})\ \ \prod_{j=1}^{M} \mathbb{P}(X_j\ \big|\boldsymbol{\theta})\ \ \mathbb{P}(X_0\ \big|\boldsymbol{\theta})\end{aligned}$$

$$\mathbb{P}(\boldsymbol{\theta}) = \ \mathbb{P}(\theta_1)\ \ldots\ \mathbb{P}(\theta_n)$$

Mean of the Conditional for Y

We assume that predictions for Y are drawn from a conditional distribution:

$$Y_j|X_j \sim \mathcal{N}\left(\nu + \beta(X_j - \mu),\ \frac{1}{\phi\lambda_j}\right)$$

The construction of the mean:

- follows from an assumption that $[X_j, Y_j]^T$ has a multivariate Gaussian distribution
- resembles (but is not the same as) linear regression

Comparison with Regression

We can compare the inference model to a similar linear regression set up:

Slope

Error of jth prediction for X

$$(Y_j - \nu) \;=\; \beta \;(X_j - \mu) \;+\; \varepsilon_j$$

$$\varepsilon_j \;\sim\; N(0, \lambda^{-1})$$

Error of jth prediction for Y

Noise

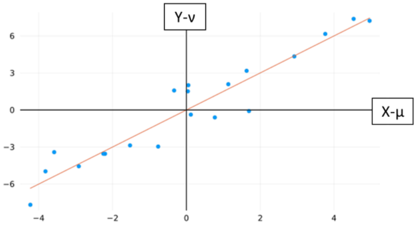
Comparison with Regression

A related regression formulation:

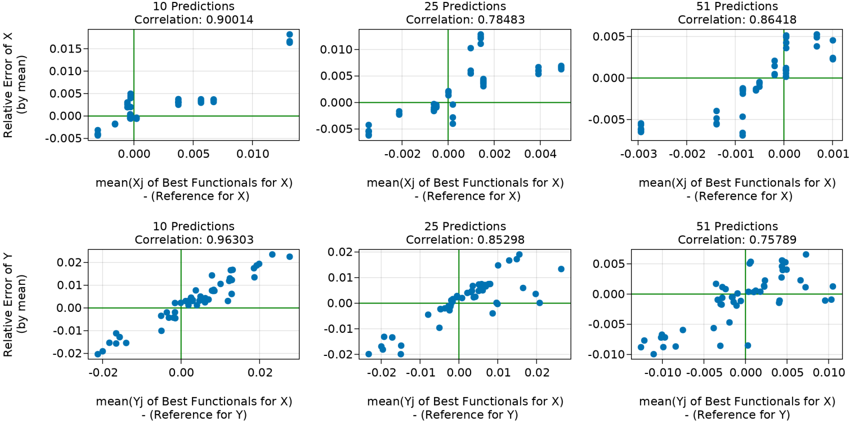
$$
\begin{aligned}
(Y_j - \nu) &= \beta(X_j - \mu) + \epsilon_j \\
\epsilon_j &\sim N(0, \lambda^{-1})
\end{aligned}
$$

- Our inference model is more flexible:
  - X_j is treated as a random variable
  - The variance of the random variables is dependent on j

Regression Example



Predictors of Inference Error: Subset DFT Mean


10 Predictions
Correlation: 0.90014

25 Predictions
Correlation: 0.78483

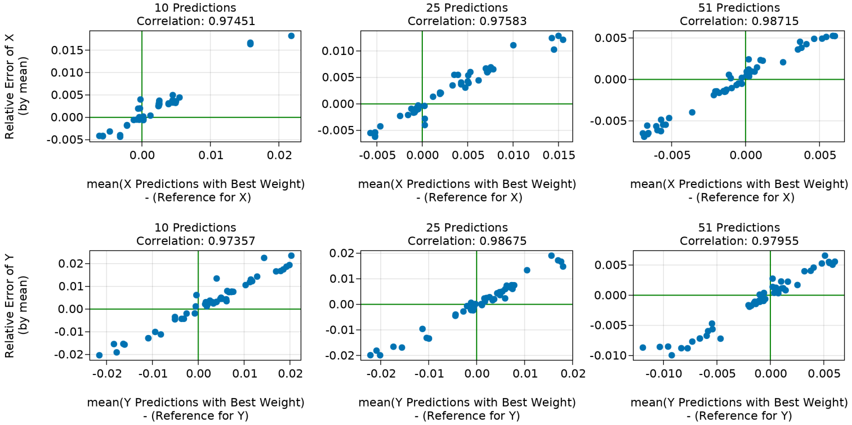
51 Predictions
Correlation: 0.86418

10 Predictions
Correlation: 0.96303

25 Predictions
Correlation: 0.85298

51 Predictions
Correlation: 0.75789

Predictors of Inference Error: Subset DFT Mean



Multireference Model

Likelihood	$\begin{bmatrix} X_j^{(1)} \\ X_j^{(2)} \\ Y_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \eta \\ \nu \end{bmatrix}, \begin{bmatrix} v_{11} & c_{12} & c_{1Y} \\ c_{12} & v_{22} & c_{2Y} \\ c_{1Y} & c_{2Y} & v_{YY} \end{bmatrix} \right)$

- We can choose the expressions for elements of the covariance matrix to model the relationships between the systems
- Prior distributions on the parameters can be used to incorporate chemical information into the inference