ACI
FALL 2023 CONVENTION

BAYESIAN MACHINE LEARNING FOR MODELING AND DESIGN OF ULTRA-HIGH PERFORMANCE CONCRETE

CHRIS CHILDS, AARON MILLER, WILLIE NEISWANGER,
BARNABAS POCZOS, LAUREN STEWART, KIMBERLY KURTIS AND NEWELL WASHBURN

OCTOBER 29TH, 2023

Georgia Tech

Carnegie Mellon University

THE IMPORTANCE OF FEATURE ENGINEERING IN APPLIED MACHINE LEARNING

- Features are variables used to define a system and build models to predict its properties
- In building predictive models, we seek features that:
  - Provide accurate predictions with a sparse feature set
  - Generalize across the response surface to make new predictions that are accurate and interesting

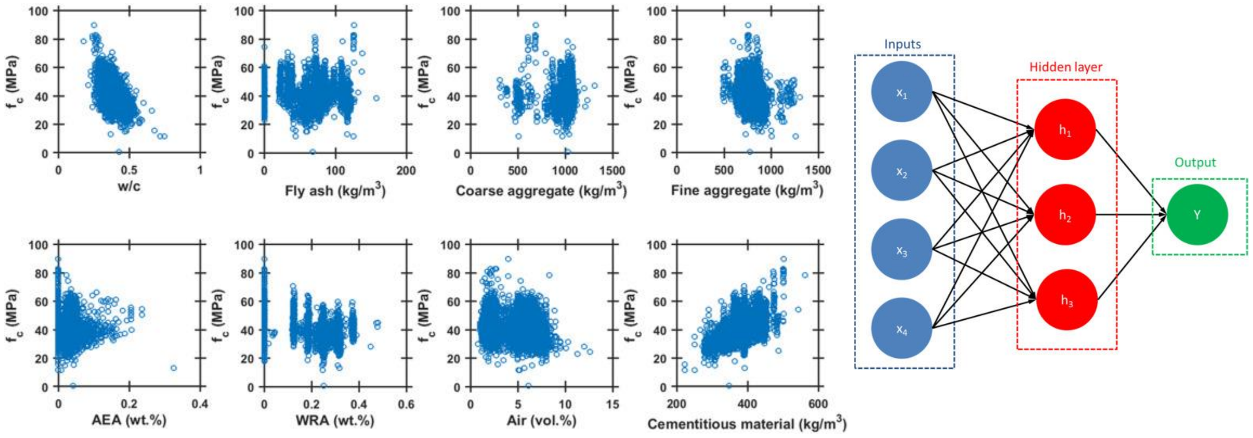
- **For cement and concrete, what are good features?**

“**Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering.**”
— Prof. Andrew Ng, Stanford U.

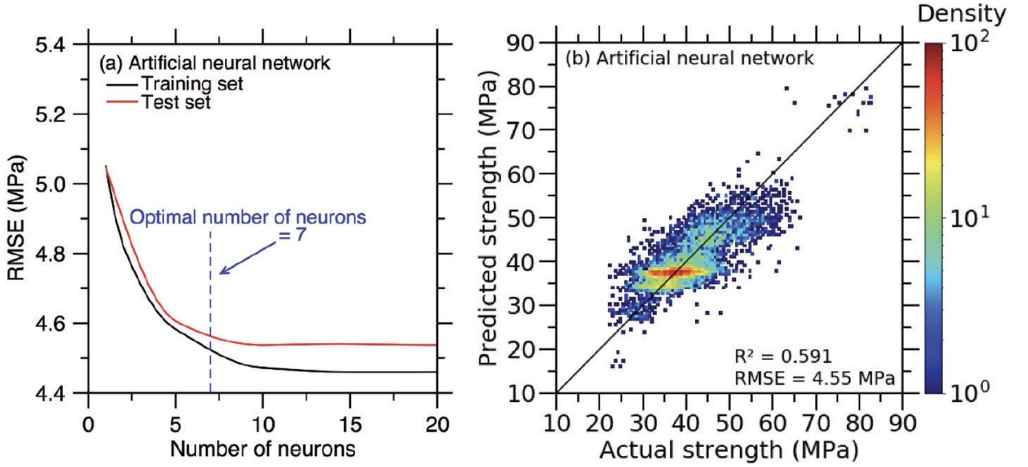
EXAMPLE: CONCRETE STRENGTH PREDICTIONS PARAMETERIZED BY COMPOSITION

- Compositional models relate properties to the amounts of specified components
  - Parameterized by "What is it?'
- The Yeh datasets of compressive strength contain ~1000 concrete samples with eight compositional variables
  - w/c, fly ash, coarse aggregate, fine aggregate, air-entraining agent, water-reducing agent, air, cementitious
- Sant, Bauchy and co-workers performed a benchmark machine learning study on the Yeh dataset
- Accuracy is moderate
  - Parity plot R^2 = 0.591
  - Feature selection is not used in compositional models
  - What are the prospects for models based on 50 samples?

Yeh. Cem. Concr. Res. (1998)



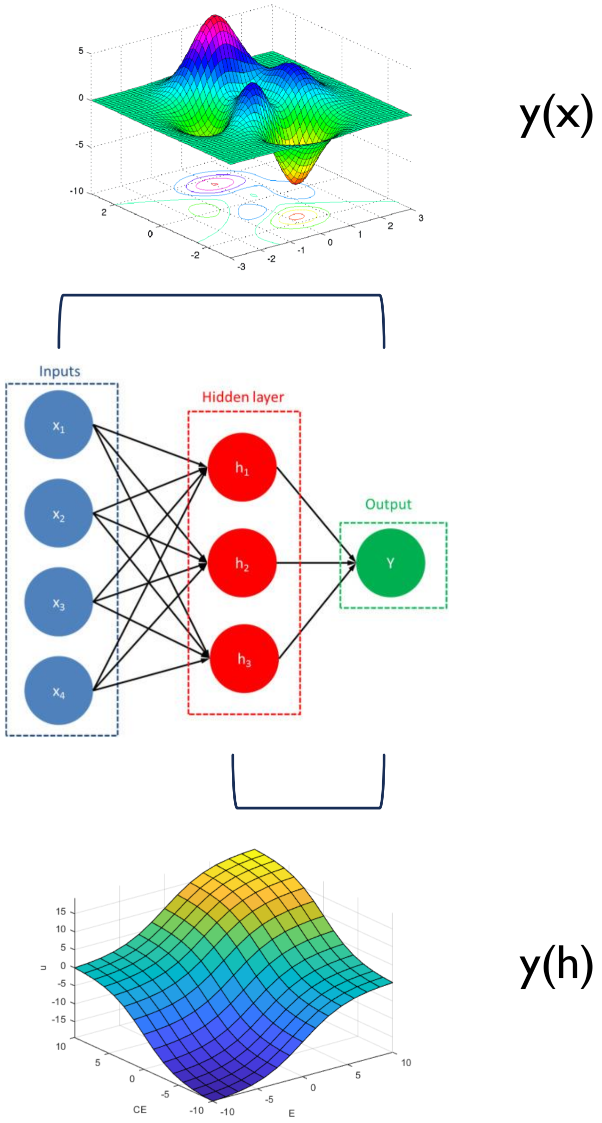
Sant & co. Cement Concr. Res. (2019)



Sant, Bauchy & co. ACI Mater. J. (2020)

EXPLICIT HIDDEN VARIABLE MODELS OF COMPLEX SYSTEMS

- Hidden (latent) variables are not controlled directly but hypothesized to govern the properties of complex systems
- We can estimate their values using additional experiments, mathematical modeling, domain knowledge
- Why build hidden variable models?
  - More accurate for small datasets
  - Meaningful feature selection
  - Simpler response surfaces
  - More interpretable
  - More generalizable
- Challenges with latent variable models:
  - Require additional data, empirical models, meaningful domain knowledge, etc.
  - Bias



OUR GOALS IN MACHINE LEARNING OF CEMENTITIOUS SYSTEMS

- Accurate predictions of workability, set time, strength development, and durability for complex mixes
- Accurate predictions from small datasets
- Simultaneous optimization of performance metrics for a given set of constituents
- A tool for technological innovation and mix design
- Mechanistic understanding
- Accurate uncertainty estimates

Wolfgang Tillmans, "Concrete Column"

BAYESIAN MODELS

- Bayes’ theorem provides a relationship between evidence (data) and a hypothesis (model)
- Prior distribution represents expectations of the range of results (confidence intervals) without any observations
- The posterior distribution represents updated confidence intervals, which are narrower around data points but revert to the prior at values far from these

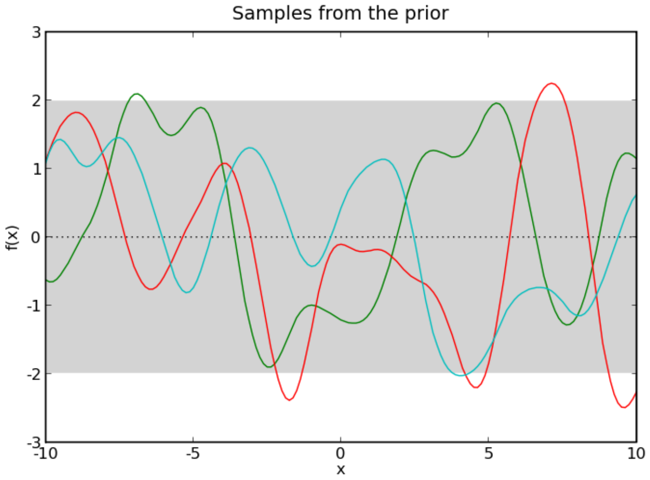
probability a hypothesis is true given the evidence

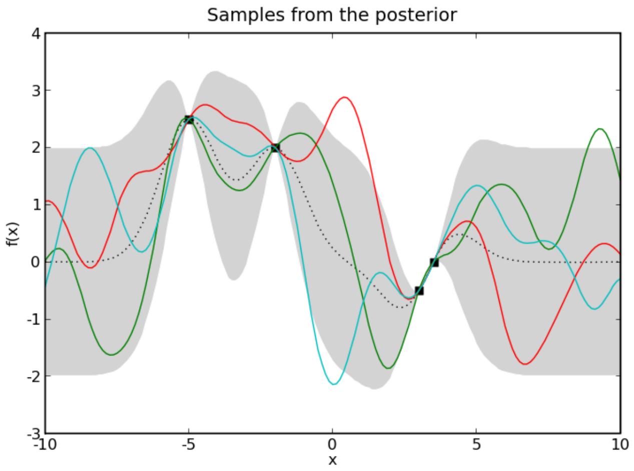
probability a hypothesis is true (before any evidence is present)

probability of seeing the evidence if the hypothesis is true

probability of observing the evidence

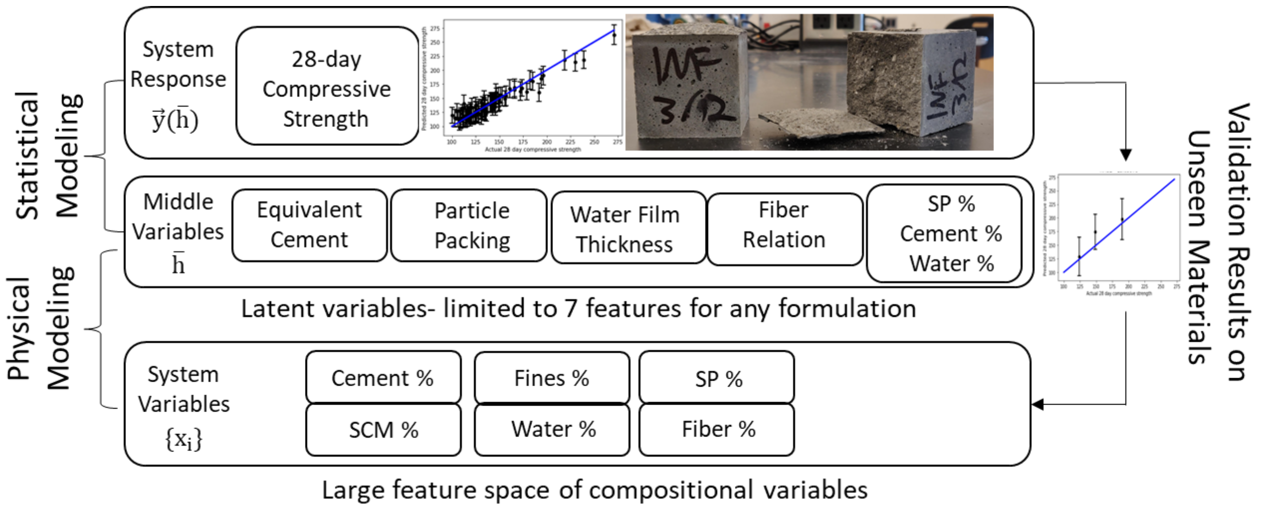
$$P(H/E) = \frac{P(H)\,P(E/H)}{P(E)}$$





BAYESIAN MODELS OF ULTRA-HIGH PERFORMANCE CONCRETE

- UHPC has a compressive strength >150 MPa
- Complex formulations utilize a diversity of supplementary cementitious materials and fibers
- How do predictions compare when the model is parameterized by compositional variables (bottom later) vs. latent variables (middle layer)?



Latent variables explored for UHPC compressive strength:

Cement + 0.5(Amount of Class F Fly Ash)*
+ 0.8(Amount of Class C Fly Ash)*
+1.2(Amount of Silica Fume)*
+ 1.2(Amount of Metakaolin)*
+ X(Amount of Slag)

equivalent cement

$$K = \sum_{i=1}^{n} \left(\frac{\frac{\Phi_i}{\Phi_i^*}}{1 - \frac{\Phi_i}{\Phi_i^*}} \right)$$

particle packing

$$WFT = \frac{u'_w}{A_m}$$

water film thickness

%SP

superplasticizer dose

$$k_{fr} = e^{0.034 * p_s}$$

fiber relation

UHPC TRAINING DATA

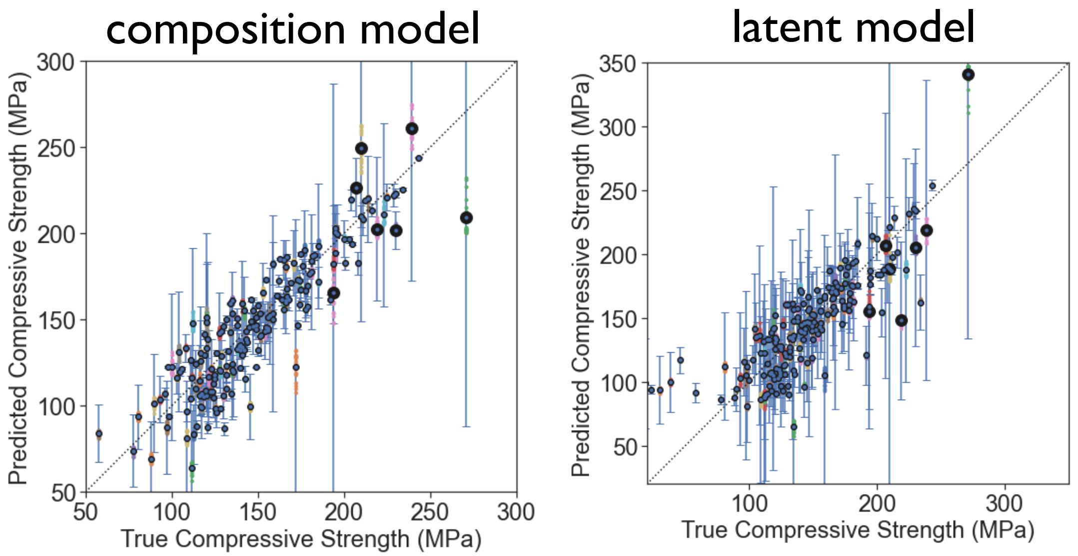
- Compressive strength results from ~100 UHPC blends were taken from literature sources to train a Bayesian machine learning model
  - Particles were only parameterized by size
  - Fibers were not differentiated and only represented by loading

Data Source	Tafraoui et. al[28]	Ghafari et. al[20]	Berry et. al[29]	Wille et. al[39]
# Samples	7	50	41	7
SCMs	Silica Fume Metakaolin	Silica Fume	Fly ash Silica Fume	Metakaolin
Fine Aggregates D50)	Sand- 230 µm Quartz- 11 µm	Sand- 400 µm Quartz- 7 µm	Sand- 500 µm	Sand- 110 µm Sand- 500 µm Glass- 5 µm
Fibers	Steel fibers, 13 mm in length and .16 mm in diameter	Two types of steel microfibers with diameters/lengths of 0.2/0.15 mm and 13/10 mm.	No Fibers Utilized	Smooth, Hooked, and Twisted Fibers ranging from 0.12-0.3mm in diameter and 6-30mm in length.

BAYESIAN RIDGE REGRESSION

- Ridge regression is a form of least squares fitting with a hyperparameter λ that serves as a regularizer on the L2 norm of the model coefficients w
  - Regularizer acts to penalize model complexity
- In Bayesian ridge regression, we trained an ensemble of 20 separate models on these literature data – each with a randomly chosen value of λ
- Results in a distribution of models from which a posterior distribution of parameters can be estimated
  - Greater accuracy
  - Greater generalizability
  - **Uncertainty estimate on each point (= error bars)**
- The parity plots for BRR indicate that parameterization of the UHPC model by compositional variables is more accurate than by latent variables
  - 20.6 MPa vs. 25.7 MPa RMSE, respectively

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

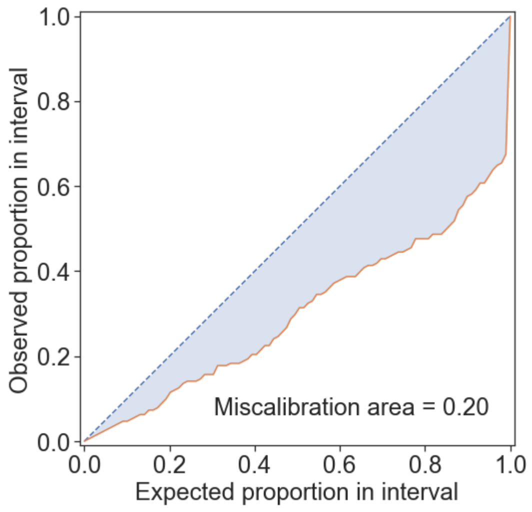


	MSE (MPa)	RMSE (MPa)
Composition	424	20.6
Latent	660	25.7

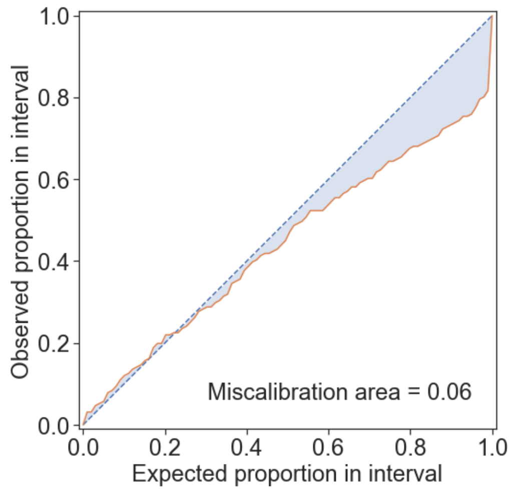
ASSESSING CONFIDENCE INTERVALS WITH MISCALIBRATION AREA

- Standard Bayesian models that provide uncertainty estimates often do not offer accurate confidence intervals
- Miscalibration area is a Bayesian error metric that provides a measure of how accurate the confidence intervals of a model predictions are
- Derived from plots of observed vs. expected proportion of population at a given confidence interval
  - Negative deviation indicates that a model is overconfident in its uncertainty estimate
  - The total (unsigned) areal deviation is reported as the miscalibration area
- While the RMSE for the compositional model is 25% lower than that of the latent model, the miscalibration area is ~3-times larger
  - **Suggests that the latent model has much greater generalizability**

Composition



Latent

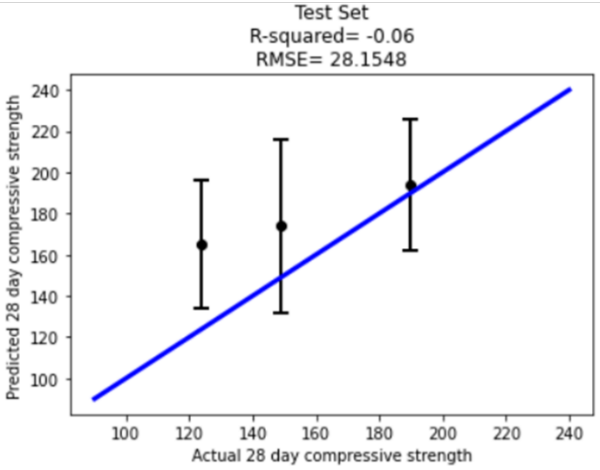


	MSE	RMSE (MPa)	Miscalibration Area
Bottom Layer	424	20.6	0.20
Middle Layer	660	25.7	0.06

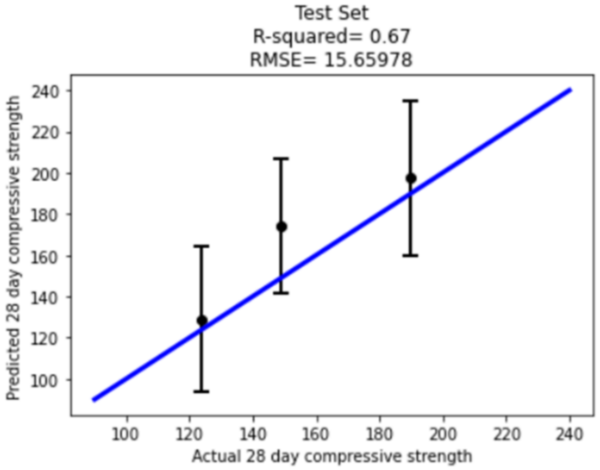
VALIDATION OF THE UHPC MODEL

- UHPC blends were formulated with sand having D50 of 600 µm
- Greater than 500 µm sand used in training the algorithm
- Three new compositions were predicted by the algorithm parameterized by latent variables
- The model parameterized by composition was much less accurate (R2 = -0.06) than that parameterized by latent variables (R2 = 0.67)

Composition



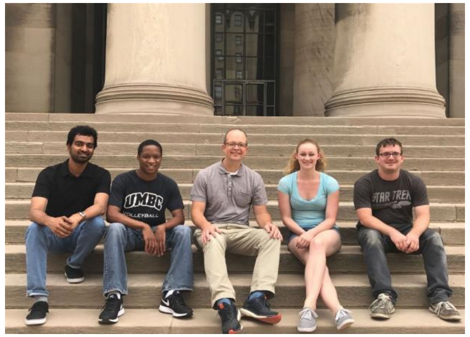
Latent



CONCLUSIONS AND ACKNOWLEDGMENTS

Prof. Barnabas Poczos (CMU)

Prof. Kim Kurtis (Georgia Tech)



- Even small datasets (10s of samples, not 1000s) on complex chemical systems can be modeled using machine learning techniques
- Latent variables are a powerful tool in machine learning of complex cementitious systems

Dr. Christopher Childs (CHEM)
Dr. Aditya Menon (MSE)
Dr. Jennifer Bone (BME)
Joseph Pugar (MSE)
Calvin Gang (CHEM)
Dr. Kedar Perkins (CHEM)

Tia Kirby (CEE)
Renee Rios (CEE)
Jiangnan Zheng (CEE)
Cheng Zhang (CEE)
Christine Huang (CHEM)
Ogulcan Canbek (GT)

arpa·e Advanced Research Projects Agency • ENERGY

NSF covestro BASF We create chemistry IMERYS ARGOS