Outline for MOF energy histogram work

Ben Bucior, Scotty Bobbitt, Arun Gopalan, Neda Bagheri, Randy Snurr

November 9, 2017

1 Notes

- The Guest Editors Andrew Ferguson and Johannes Hachmann for the special issue chaired my session at AIChE.
- Themed issue "Machine Learning and Data Science in Materials Design" for the journal *Molecular Systems Design & Engineering*
- Filling in the outline with rough figures and ideas for the flow

2 Title

- (scratch work. Let's come back to this later.)
- (from AIChE) Identifying New Descriptors for Gas Storage in Nanoporous Materials

3 Requirements and future directions

3.1 Quick tasks for Ben

- Figure out in meeting: Do we pre-set the upper and/or lower bounds of the histograms?
- Fitting models on the 2 bar and 100 bar cases. Do the coefficients sum to the combined model?
  - Answer: yes. I calculated a "discrepancy" term between the original ridge model and by subtracting the original high and low pressure models. The max discrepancy between betas is 0.3 ("Inf" bin), which isn't unexpected considering we regularize the models separately.

    Also from the plot, we can observe properties of the individual fitted models. Here, we can see that the highly attractive region actually has similar beta values. The difference really comes into play at the mildly attractive regions, where adsorption is not sufficiently strong to bind H2 at 2 bar. Also the "Inf" bin might be attributed to the differences in β_0, 41 vs. 17 g/L.

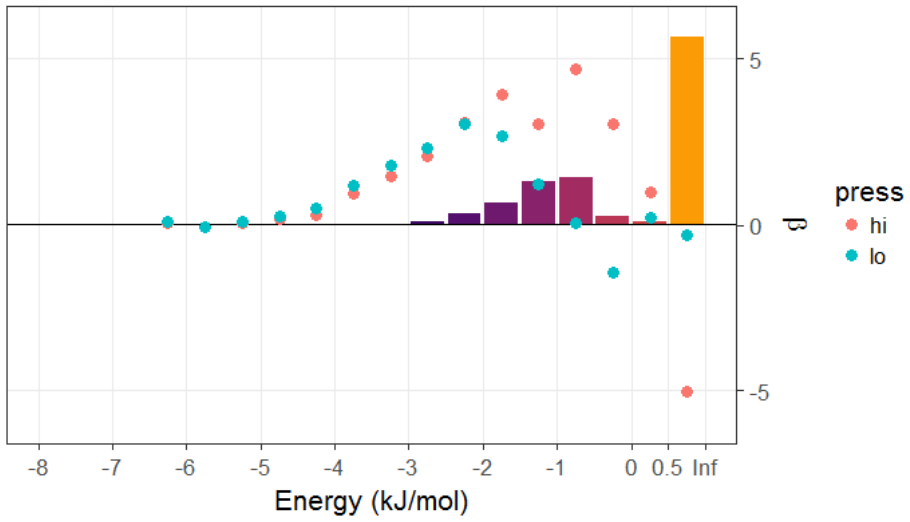


Figure 1: Difference between the beta coefficients at 2 bar and 100 bar (add this as a subfigure to the other beta plot?)

- Take a look at the journal guidelines for authors and their LaTeX templates. See also Overleaf?

3.2 Potential directions

- **How general** is the method? Just hMOFs? Also ToBaCCo?
  - Could retrain on ToBaCCo and compare coefs
  - IZA database of zeolites? (not just MOFs)
  - CCDC MOFs - test model on 1000 MOFs, check agreement, then screen the rest of them?
- Identify old sources of **GCMC data for reuse** (email Diego or others?)
- **Other gases:** methane
- **Temperature dependence/optimization:** Cryo vs. room temperature adsorption?
  - Then, along with the story, we could see if the magnitude/direction of ridge regression coefficients match the intuition of needing stronger adsorption at higher temperature.

3.3 Other (unlikely) directions

- **Experimental collaboration:** Screening the CCDC MOFs to identify top candidates for hydrogen storage (Would likely take a few weeks at best, so skip)
- **Isotherm prediction or Langmuir rationalization:** Repeat the predictions on multiple pressures. What do the coefficients look like? Comparison against isotherm prediction work?
- Other gases: xenon, multi-site molecules (large and complicated. Let's save Xe/Kr for the isotherm prediction paper)

4 Introduction

- Hydrogen storage challenge
- MOF background
- Screening MOFs for hydrogen storage
  - Generating databases of hypothetical MOFs
  - Automating screening
  - More computationally efficient methods for screening
  - Scotty/Jiayi paper (agnostic binding fraction, JPCC)
  - Yamil, Diego works
- Use of ML in the MOF literature
  - APRDF
  - Recent work from U Connecticut on textural properties + chemistry
  - Revisit my screening section in the review book for more ideas on papers to cite
  - Use of potentially related features in the literature, like the Coulomb matrix, heat of adsorption, etc.
- Current paper not only shows potential for accelerating MOF screening but also learning what is driving the capacity of these materials
  - Accelerate beyond brute force GCMC
  - LJ metric and ridge regression get to the heart of what makes a good material (avoids problem of void fraction peak). Higher metric = higher capacity

5 Methods

TODO: Flip sections 5.1 and 5.2. Rewrite them as:

1. Data collection (GCMC, DB, grid, z-scoring, etc.)
2. Data analysis (ridge, etc.–error handling?)
3. Data sharing (consider making the GCMC and PES data publicly available. Also could make a quick web interface where users can make predictions on their own MOF)

5.1 Calculating the energy grid

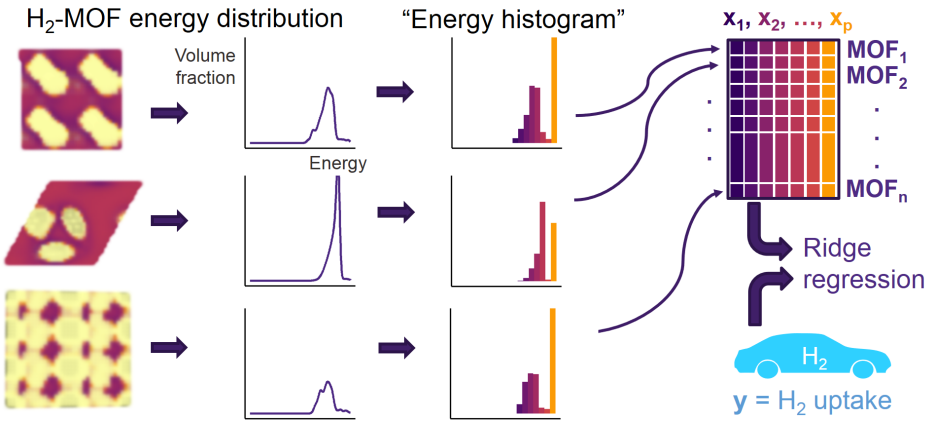


Figure 2: Calculation of the energy histogram from inputs

5.2 Molecular simulations

- RASPA
- Force field parameters for host and guests
- Number of cycles and/or source of data
- Structures
  - Wilmer's hMOFs
  - Likely ToBaCCo, so we can also look at structural diversity
  - Cleaning up the CCDC MOF subset using their solvent removal scripts, etc.

5.3 Data processing

- Ridge regression
  - Equations and loss function
  - R package `glmnet`
    - Finding the greatest lambda within 1 SE of the lambda that minimizes model error

e'll be citing a known package

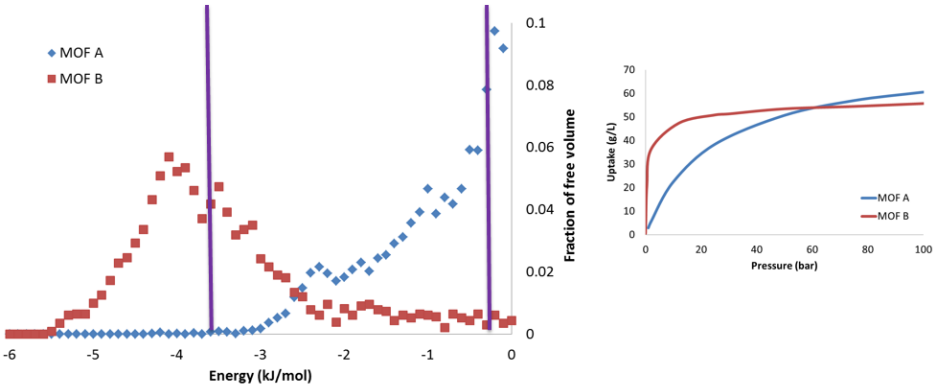
- Data preprocessing: z-score bins and remove columns with zero variance. Also filter out unphysical uptake (< 0 g/L) from GCMC
- Define equations used for model evaluation: Q2, RMSE, MAE [move to results, as necessary]

6 Results and discussion

6.1 LJ metric [we'll see where this fits best in the flow]

- Inspiration from "binding fraction"
  - A similar idea, but calculated using interaction energy instead of geometry
  - Set two cutoffs on the energy distribution, and integrate between them to get a single value.

- Describe graphene sheet calculation, showing the 7 Ang and the other cutoff, and how they correspond with our energies empirically derived



MOF	LCD (Å)	2 bar (g/L)	100 bar (g/L)	Del. Cap. (g/L)	LJ Metric
MOF A	14.8	4.1	51.1	47	0.65
MOF B	6.8	27.4	47.4	20	0.08

Figure 3: Case study of two MOFs (IDs?) demonstrating that deliverable capacity also depends on the absolute uptake of the delivery pressure. Note that MOFs A and B have similar saturation loadings, but considerably different deliverable capacity due to uptake at the low delivery pressure (REDO)

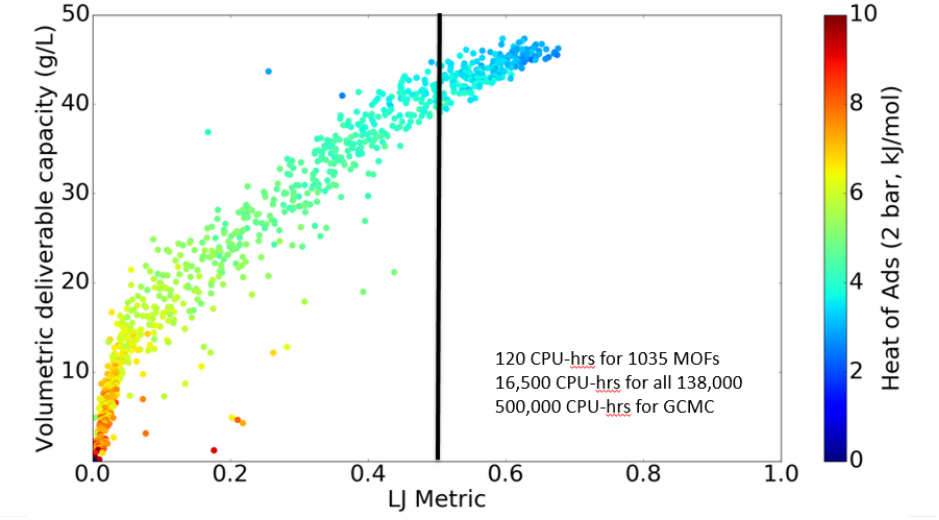


Figure 4: The LJ metric generally indicates deliverable capacity. Each point represents results for one of $\approx$ 1000 hMOFs tested. [ADD DC VS. VF PLOT as another subfigure]

6.2 Ridge regression

- Formalizing the results from LJ metric studies in a predictive model without empirical parameter determination

- Meaning of betas and intuition

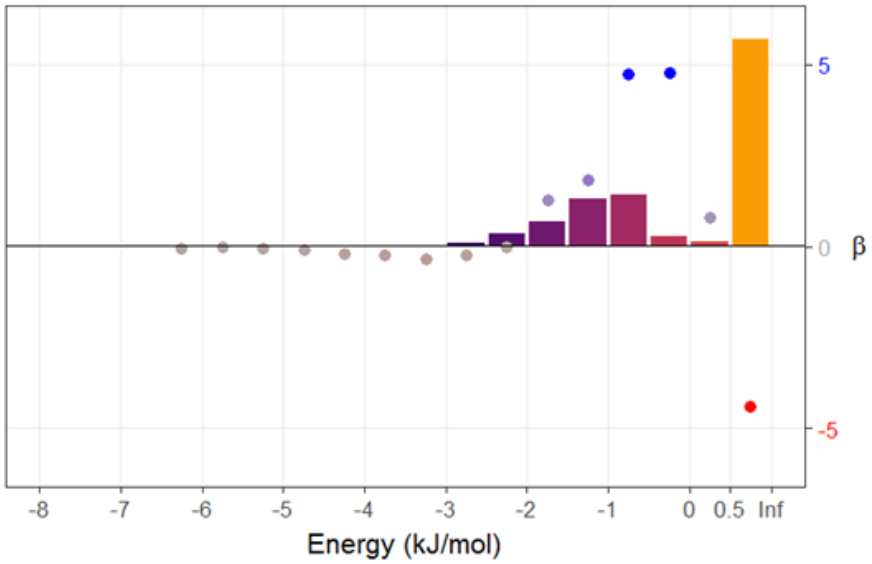


Figure 5: Regression coefficients overlaid on the histogram

- How good is the model and fit?
  - Parity plot

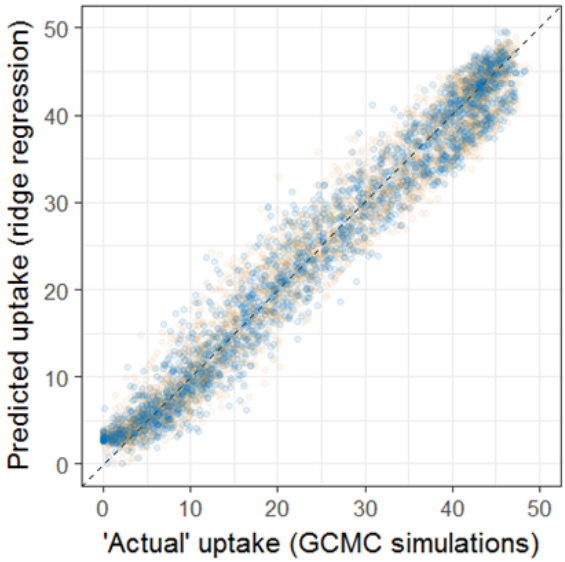


Figure 6: Parity plot for (a) training and (b) test data on the energy histograms (TODO: split into two subfigures so it's easier to read without an animation.)

  - Q2 of 0.96
  - RMSE of 3 g/L, MAE/MUE of 2.4 g/L

- Why ridge regression?

- Interpretability with a simple model
- Likely some kind of link with a multi-site Langmuir model (assuming that all sites at a given energy are identical)
- Someone at AIChE asked why our model didn't include nonlinearity. The reason is that we don't need that to make great predictions. (some of my initial work used random forest, which would avoid manual feature transformations)

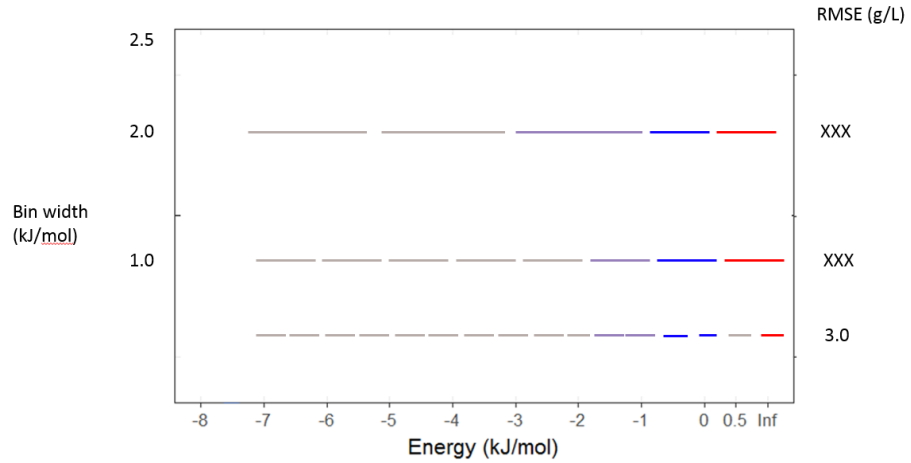


Figure 7: DEMO Model coefficients are robust to histogram bin strategy: showing the locations of the bins, color coded by the beta, to see if the model gives consistent results for certain sections of energy (TODO after meeting)

6.3 Screening other databases

- Plan of testing applicability on 1000 MOFs
- Benchmarking (maybe as a table): Full GCMC vs. energy grid calculations, feature representation, and ML
- TODO: pick out the top 500 CSD MOFs and test them with GCMC

Figure 8: TODO: top candidates for experimental synthesis based on screening the CCDC MOF database (MAYBE?)

6.4 Generalizability to other gases

- Methane. See above

6.5 Other figures

The last two sections will naturally lend themselves to additional figures as well.

7 Acknowledgements

Data Science Initiative, NMGC, etc.

8 Supporting Info

8.1 Hyperparameter tuning

- Grid spacing: convergence of a few different sample MOF histograms
  - Also might be good to have a figure overlaying sampling points on top of a continuous background, to exemplify the convergence testing
  - TODO make a figure more rigorously
- Bin width and degree of overlap (TODO: consider adding examples, and Q2 figure)
- Lambda selection

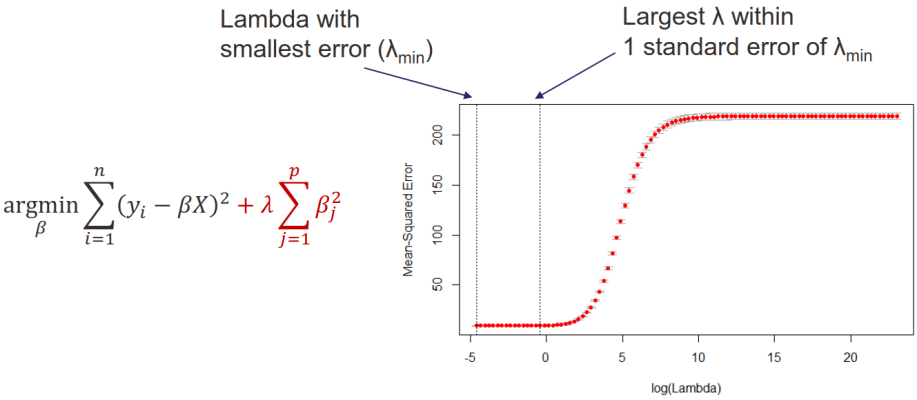


Figure 9: Determination of the regularization parameter λ for ridge regression

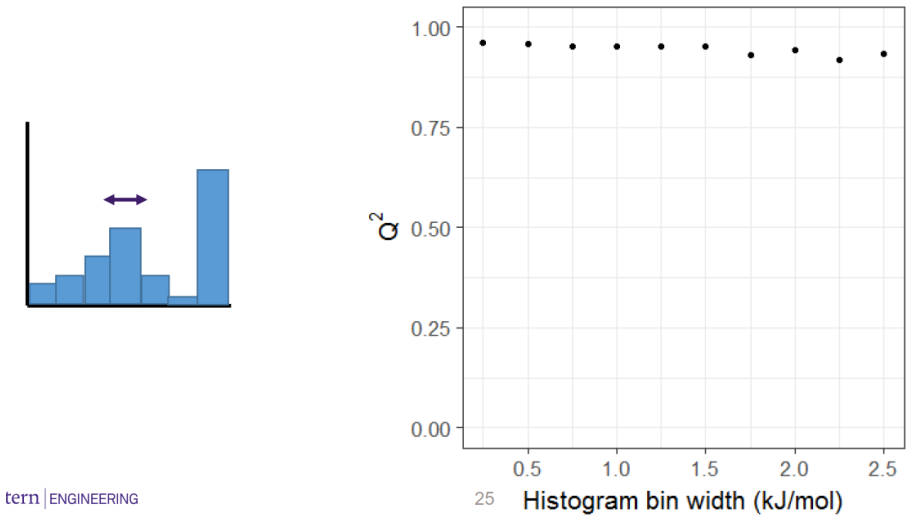


Figure 10: Model is robust to selected histogram bin width (perhaps repeat with RMSE in addition to Q2)

8.2 Model evaluation/consistency

Also consider adding a figure on "Consistency across nodes/linkers and/or other DBs" (see Zr results)

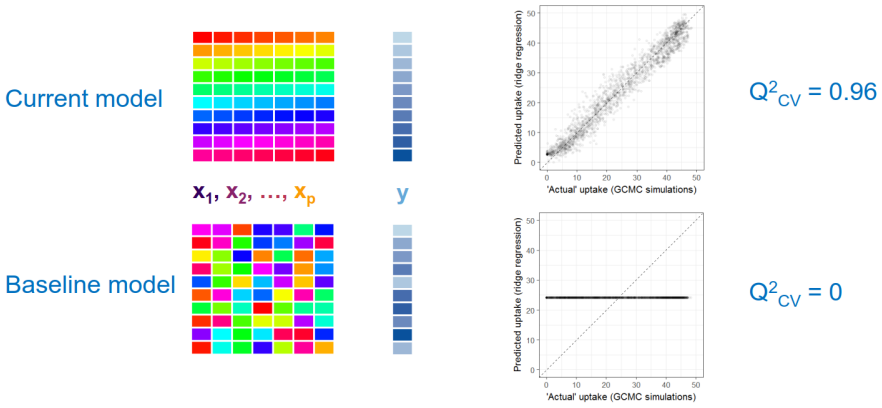


Figure 11: Comparison of Q^2 against a baseline of random data

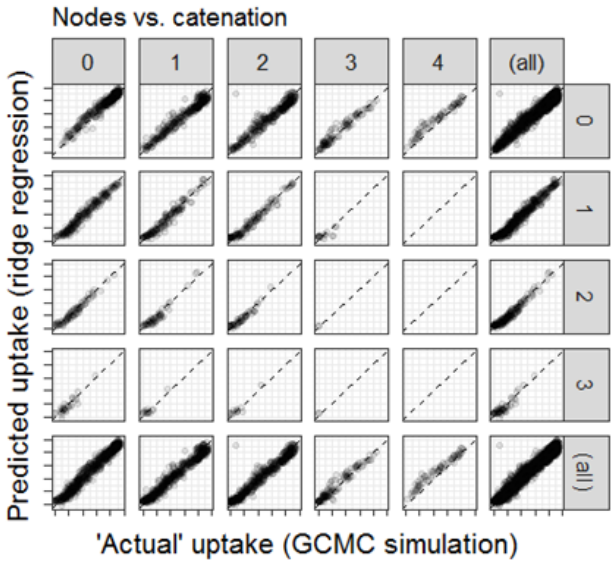


Figure 12: Consistency of model accuracy across MOF compositions. Note that Zr MOFs are less accurate (node 4), possibly due to differences in topology and undersampling relative to **pcu** MOFs.

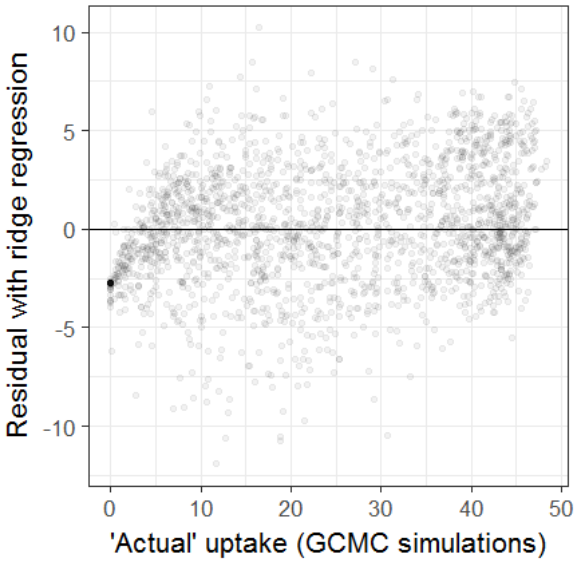


Figure 13: Residual plot for the testing data. One audience member in my talk once concerned about a long tail (?) on the top right of my parity plot, below the line of parity, possibly requiring another variable for correction. But I'm having trouble seeing it in the residual plot, so perhaps it was an optical illusion against the 45 degree line.

[TODO: also fit the residuals to a Gaussian to verify that there isn't a problem with top right]

8.3 Alternative approaches

- Benchmarking against traditional descriptors (textual properties like void fraction and density)

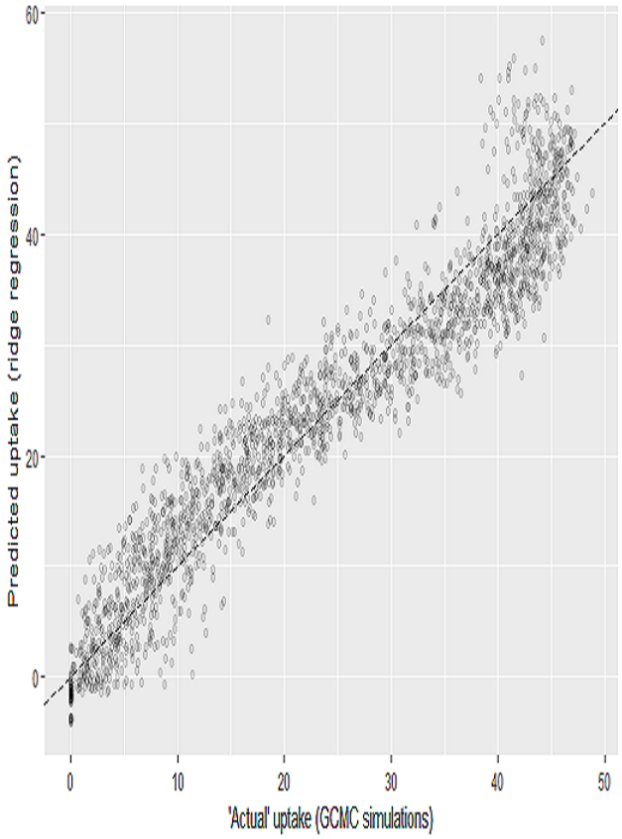
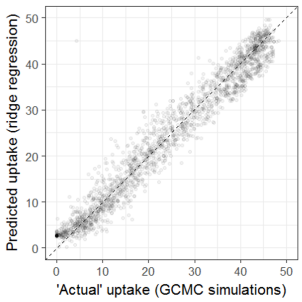
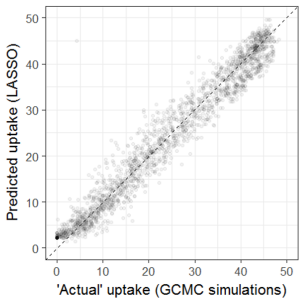


Figure 14: Textural properties (void frac, grav SA, vol SA, DPD, LCD, max catenation, and actual catenation) have poorer predictivity. On the test set, RMSE=4.3, MAE=3.5, and note the uneven residuals. (from 10/06, need to replot)

- LASSO figure and coefficients



Beta	Ridge	LASSO
(Intercept)	24.11	24.11
4	-0.06	
5	-0.01	
6	-0.05	
7	-0.10	-0.06
8	-0.22	-0.16
9	-0.26	-0.18
10	-0.36	-0.35
11	-0.24	-0.10
12	-0.03	
13	1.26	1.25
14	1.81	1.73
15	4.71	4.86
16	4.75	4.62
17	0.79	0.69
Inf	-4.41	-4.65



Performance	Ridge	LASSO
RMSE	2.92	2.92
Rsquared	0.96	0.96
MAE	2.31	2.31

Figure 15: Ridge regression and LASSO give similar results

8.4 Some newer figures on normality and transferability

How well does the model work on a set of new 500 CCDC MOFs? The left subfigure speaks for itself.

The discrepancy could be a number of causes: first we need to check the sigma/epsilon parameters for the energy calculations. The original 10k hMOFs used an epsilon of 11.25 K and sigma of 2.68 Angstroms, but the actual sigma was about 36.7 K. After correction, the fit is considerably better (right subfigure). For the remainder of the error, consider that we mostly trained on **pcu** MOFs, whereas the CCDC MOFs are considerably more diverse.

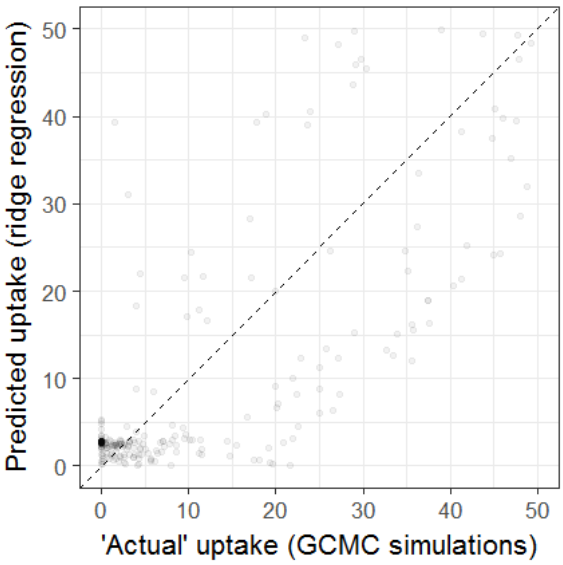


Figure 16: Parity plot for the CCDC MOFs, using the initial ridge regression model trained on 4000 hMOFs.

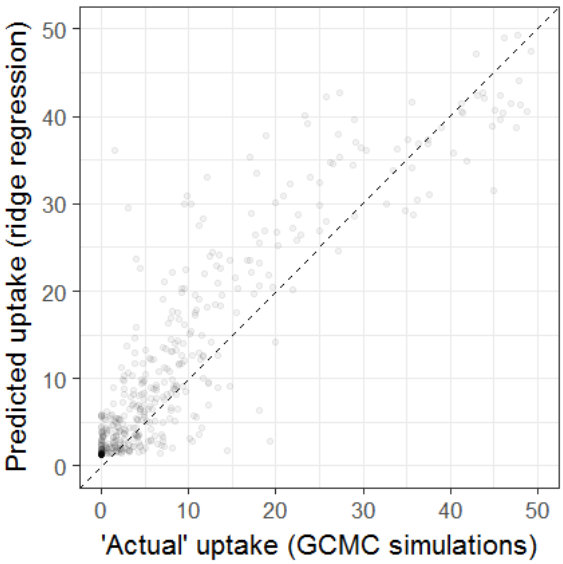


Figure 17: Parity plot for the CCDC MOFs, after scaling the energies fed to the hMOF ridge regression model to match epsilon.

Let's take a look at the distribution of z-scores. Do high or low performing MOFs tend to have similar z-scores in certain bins? Are the z-scores reasonable (especially for "important" bins)?

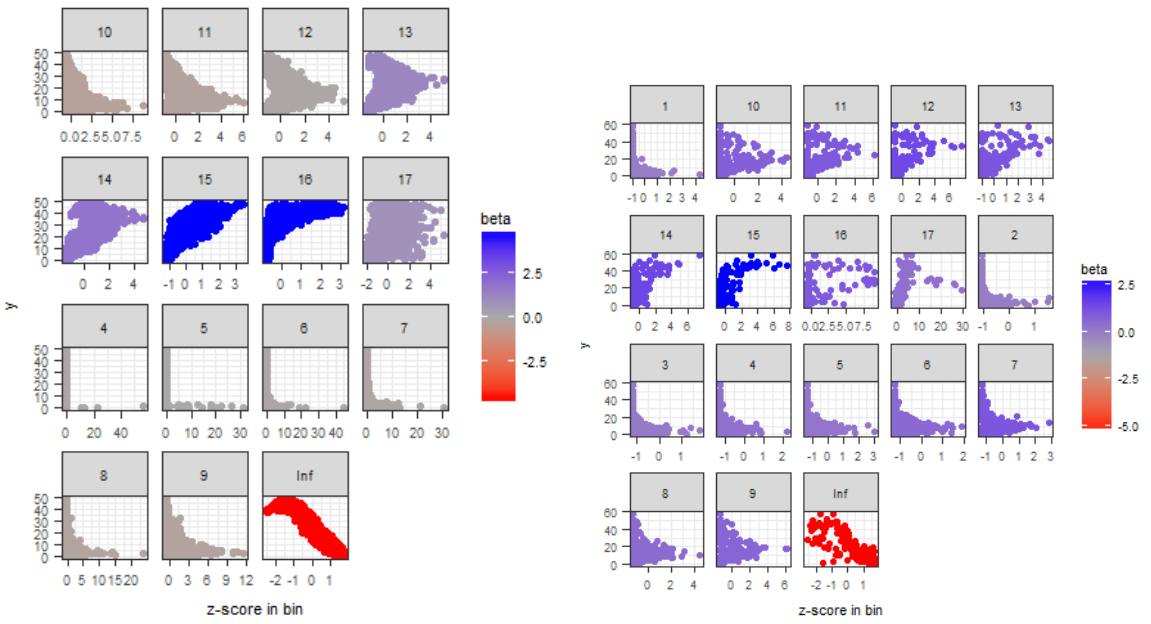


Figure 18: Distribution of z-scores (x variables) in each bin, colored by the beta of the trained model. Extreme z-scores are more common in the bins with low weights. Left: original hMOF data, right: CCDC vs. actual GCMC, with the new betas (from scaled hMOF energies)