Outline for MOF energy histogram work

Ben Bucior, N. Scott Bobbitt, Arun Gopalan, Neda Bagheri, Randy Snurr

November 6, 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*

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

- Fill in the remainder of the figures I could see for this paper
- Debrief with Scotty and get his input on availability and possible directions for the manuscript. He might have some good figures we should adapt for the manuscript, too.
- Do we pre-set the upper and/or lower bounds of the histograms?
- Check out the long-tail(?) of the parity plot (top right, below the parity line). Could be another variable we need to account for
- Need to comment on nonlinearity somewhere, based on question from AIChE session
- What is the original distribution of y?

3.2 Potential directions

- **How general** is the method? Just hMOFs? Also ToBaCCo?
  - Could retrain on ToBaCCo and compare coefs
  - IZA database of zeolites?
  - 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, xenon, multi-site molecules?
- **Experimental collaboration:** Screening the CCDC MOFs to identify top candidates for hydrogen storage (Would likely take a few weeks at best)

- **Isotherm prediction or Langmuir rationalization:** Repeat the predictions on multiple pressures (possibly temperatures as well). What do the coefficients look like? Is it similar to the Langmuir intuition on how much variability there will be?
- **Temperature dependence/optimization:** Could also consider changing temperature. Cryo vs. room temperature adsorption?

4 Introduction

- Hydrogen storage challenge
- MOFs and MOF databases
- Screening MOFs for hydrogen storage
  - Scotty/Jiayi paper
  - Yamil, Diego works
- Use of ML in the MOF literature
  - APRDF
  - Recent work from U Connecticut
  - See also my section in the review book
  - Relation for the Coulomb descriptor and related proxies in the literature (heat of adsorption, etc.)

5 Methods

5.1 Calculating the energy grid

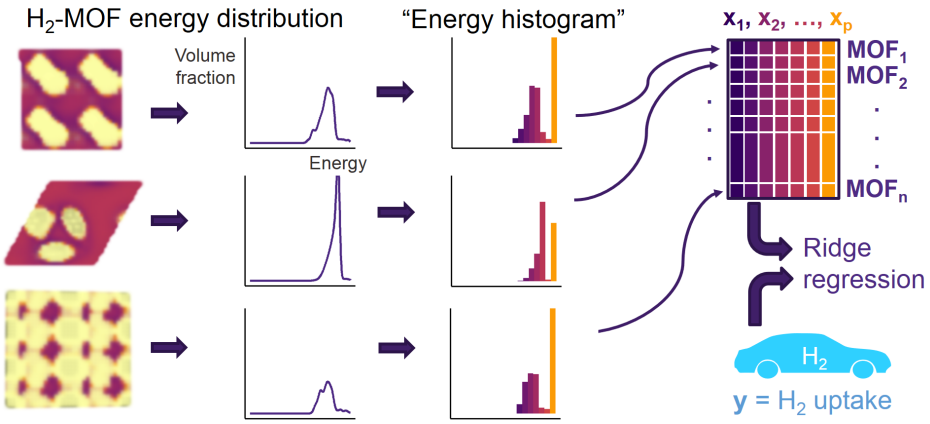


Figure 1: Calculation of the energy histogram from inputs

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
- Data preprocessing: z-score bins and remove columns with zero variance. Also filter out unphysical uptake (¡ 0 g/L) from GCMC with giant error bars
- Define equations used for model evaluation: Q2, RMSE, MAE

6 Results and discussion

6.1 LJ metric

- Inspiration from "binding fraction"
- Ask Scotty about this section. Figures of different distributions? Comparing MOFs that bind too strongly, etc?

6.2 Ridge regression

- Formalizing the results from LJ metric studies
- Meaning of betas and intuition

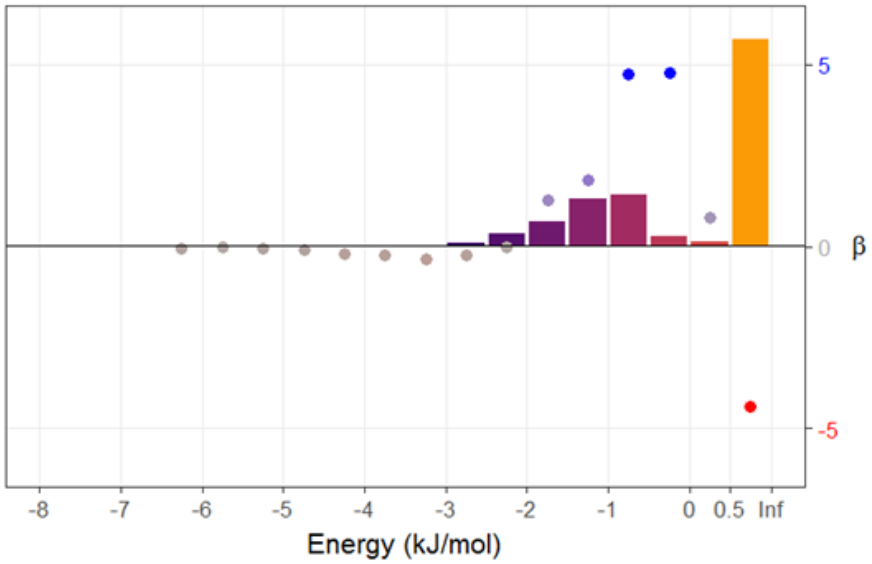


Figure 2: Regression coefficients overlaid on the histogram

- How good is the model and fit?

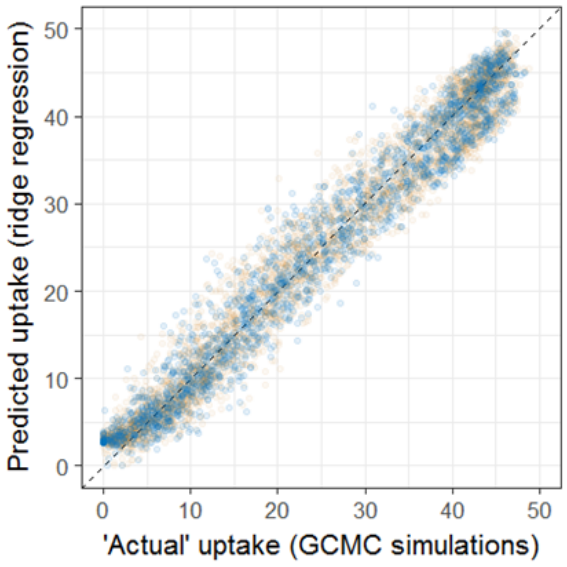


Figure 3: Parity plot for (a) training and (b) test data on the energy histograms (TODO: split into two separate figures so it's more black and white-friendly, and easier to read without an animation)

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

6.3 Screening

- Plan of testing applicability on 1000 MOFs
- Benchmarking (maybe as a table): Full GCMC vs. energy grid calculations, feature representation, and ML

Figure 4: TODO: top candidates for experimental synthesis from the CCDC MOF database

6.4 Generalizability to other gases

- See above

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
- Bin width and degree of overlap (TODO: consider adding examples, and Q2 figure)
- Lambda selection

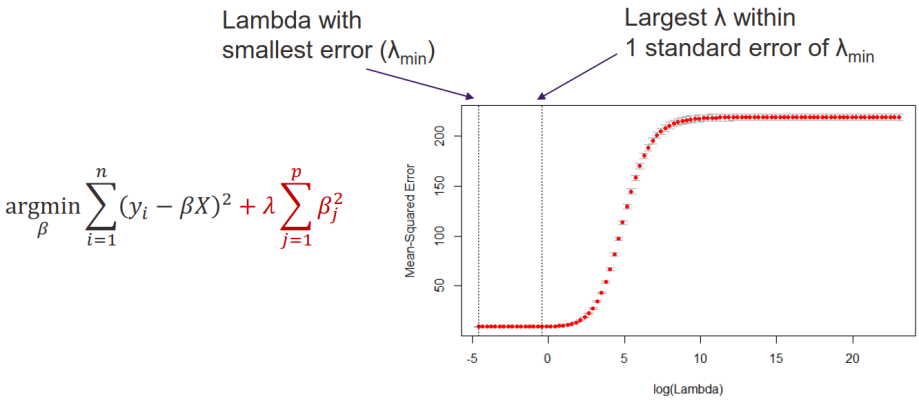


Figure 5: Determination of the regularization parameter λ for ridge regression

8.2 Model evaluation/consistency

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

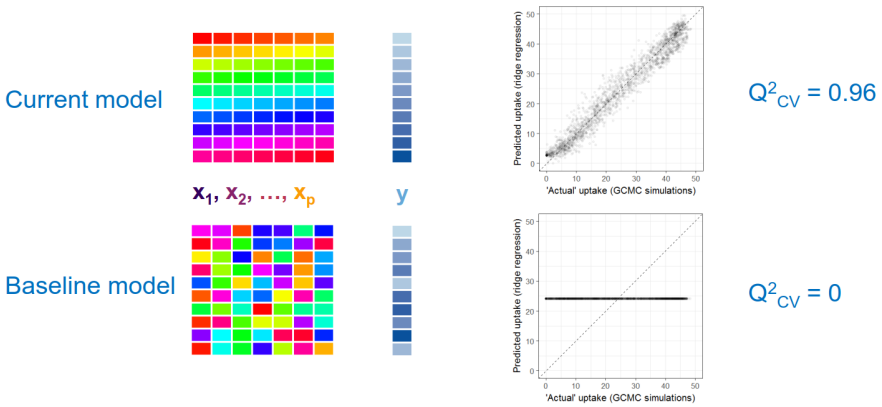


Figure 6: Comparison of Q^2 against a baseline of random data

8.3 Alternative approaches

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

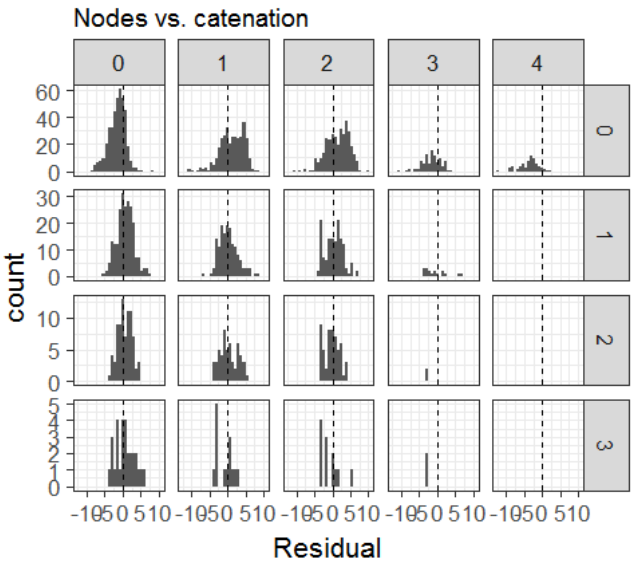
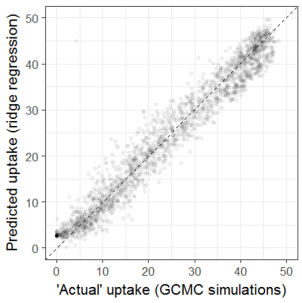
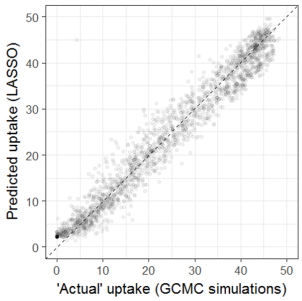


Figure 7: Consistency of model accuracy across MOF compositions. Note that Zr MOFs are less accurate (node 4), possibly due to differences in topology and undersampling relative to **pcu** MOFs.



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 8: Ridge regression and LASSO give similar results