## **Meeting Notes for the Quarter**

Taken by Kaylyn with Julia filling in for 5/24/21

## **6/7/21 Weekly Meeting Notes**

Finish presentation by Wednesday.

In charge of main room pitch: Julia

Do we want a waiting room for our presentation? Yes, Julia will be in charge of the waiting room.

Still to do:

* More unit tests (Tuesday), goal is 90% coverage
* Update documentation (Tuesday) and finalize Pep8 compliance
* Best practices for mechanistic model?
  + I’ll talk to Julia about this today.
  + Maybe mention the notebook she’s been using?
* What do we say about our analysis?
  + Talk to everyone about this.
* Future Goals
  + Completing a full factorial analysis
  + Number of impurities (how will this change the model?)
  + Fix the Langmuir isotherm model
  + Total guesses in the mechanistic model

Reconvene in a bit to talk about the presentation outline.

## **6/1/21 Weekly Meeting Notes**

## By Friday (or this weekend):

## Datasets with 2 impurities--Julia

## K Fold Cross Validation--Diego

## Visualization Tool

* + Dash Interface--Diego
  + Plotly Plots--Kaylyn

## Unit Tests--Kaylyn

* Example Notebook--Kaylyn

Monday:

* Presentation made by time to meet with Josh/Wes--all
* Give to Josh/Wes--all
* Finalize pitch--Julia
* Turn in stuff to Kelly--Kaylyn

Wednesday: Final Presentation from 12:30-2:30 PM.

## **5/24/21 Weekly Meeting Notes**

* Need to change number of impurities
  + Analyze this
* Use Dash to visualize factor impact
  + Everyone should be thinking about what information is meaningful to see visually by the user
* Fix the elevator pitch
  + Julia - ask friends if they understand what I’m saying and where to improve. What’s not clear?
    - Also Julia make sure the version in the Google Drive is most current
  + Everyone can think about how it can improve
* Organize both repo’s.
* Final presentation is June 9th 12:30-2:20pm.
* Update Gantt Chart - Julia

## **5/21/21 Meeting with Just Notes** Discussion:

* Josh seems okay with random uniform distribution. Maybe try LHS if we have time but I’m thinking we likely won’t.
* Surprised by the isotherm comparison data (why are these transferable between each other?)
* Might need better clarity on yield/purity shape differences and color differences in the plots.
* Resin type makes sense. This is an area where the model is not necessarily transferable. Can we effectively predict across resins using a training set with many different resin types? -- is Josh going to have to split things based on resin if he wants to predict with a specific resin type?
* Do we want to focus on the impurity route? Fix the NN architecture and then vary input size. Adding more impurities is going to be more complicated physics; more interactions between more molecules is going to be more complex.
* **HIGHLIGHT: What areas of input parameter space does the model struggle with the most?**
* **JOSH’S DELIVERABLE: What are the problems we’re having in terms of generating training examples? How does using the MM work? Best practices and future improvements we have for the MM.**
* **FEEDBACK ON OUR DELIVERABLES: Longer form PPT presentation for Josh as well (similar to our final presentation). Ask Josh what kind of analysis is interesting to the scientists. Big fan of the integrative dash app idea.**

Problems with the Mechanistic Model:

* Benedicte explains the iterations on producing data with the mechanistic model.
* Josh wasn’t using those named proteins at all in his initial mol res scan file.
* In the lab, when you’re changing resin type, you’re also changing the protein resin interaction parameters (this might be why the models aren’t necessarily super transferable).
* The capacity of different resins right now is what is currently breaking the model’s transferability between resin type datasets. (we haven’t really changed the protein resin interaction parameters because we’re sampling randomly) Should be able to predict across resins reliably if we use capacity as an input to the model.
* **Worth trying two impurities to see how interactions vary.**

## **5/17/21 Weekly Meeting Notes**

Discussion of Tasks:

* Data Generation as soon as possible
  + Generation of the necessary datasets finished?
  + Uniform distribution makes more sense if we’re just pulling from a random range (no reason to weight things differently)
  + 200 combos is pretty good for n\_combos when using MM
* Project Presentation for Josh BY FRIDAY
  + Jupyter Notebook with comparisons between data sets over the three models that we’ve built (varying the things that Josh is interested in?)
  + Final project proposal (what are we giving to Josh at the end of the quarter?)
* Visualization and Final Project Construction

STANDUP ON FRIDAY: Benedicte

BY NEXT MONDAY, MAY 24TH: Fill in after meeting with Josh

## **5/10/21 Weekly Meeting Notes**

Discussion of Tasks:

* Pitching our product (video due Friday)
  + Julia
* Fixing models, standardizing pipelines, unit testing
  + Diego + Kaylyn (more Kaylyn than Diego)
* Data generation
  + Compare Langmuir and GIEX isotherm. Try different impurities and different combinations of impurities. Vary sampling strategies.
  + What other differences in data input are going to be important? Julia is working on resin type.
  + Julia + Benedicte + Steven
* Should we start the visualization aspect?

STANDUP ON FRIDAY: Steven

BY NEXT MONDAY, MAY 17TH: Some conclusions on the comparison of datasets to discuss with Josh.

## **5/3/21 Weekly Meeting Notes**

Discussion of Tasks:

* Data Processing pipeline w/ unit tests
  + Need to split experiments so they are in testing OR training, not both.
  + Diego + Kaylyn
* Accuracy metric pipeline w/ unit tests
  + Need to use the same accuracy metrics for all three models.
  + Make decent visuals to compare models.
  + Diego + Kaylyn
* Data Generation: Standardize naming system and provide explanations for changing parameters between datasets.
  + Compare Langmuir and GIEX isotherm.
  + Try different impurities and different combinations of impurities.
  + Vary sampling strategies.
  + What other differences in data input are going to be important?
  + Julia + Benedicte + Steven

STANDUP ON FRIDAY: Kaylyn

BY NEXT MONDAY, MAY 10TH: Finished two pipelines with unit tests, multiple data sets generated so that we can compare model performance across sampling strategies/isotherms/impurities. Schedule meeting with Josh if we need more discussion on impurities.

## **4/29/21 Meeting with Just Notes**

Input Parameter Space: Some of the inputs seem pretty arbitrary.

* Look at the paper that Josh and Julia were talking about; should have ranges of values for generating training sets for NN builds.
* K\_ov is equivalent to K\_eff, might be a couple of orders of magnitude too high
* Axial diffusion parameter (D\_L): related to hydrodynamic radius of the protein, but Josh picked this number from a paper. Pick one that coordinates with monoclonal antibodies since those are the most interesting.
* From Julia’s notebook; k\_eq seems to be the most important factor for both yield and purity.
* Plot true yield and purity with predictions and then plot Julia’s linear regression line.
* Do we change column length or inlet concentrations? (Josh initially set the column parameters taken from the AICHE paper. Arbitrarily chosen antibody concentration bc it makes sense based on experience.)--Might not be as

important but we are open to explore this. Keep column parameters the same, maybe change inlet concentration.

* Multiple different impurities, different types of impurities.
* <https://onlinelibrary.wiley.com/doi/epdf/10.1002/elsc.201400247>
* Adding impurities^^^

Other Notes:

* Look at a subset of cut values. Maybe overestimating the accuracy of the models because the cut values are dependent on each other (they’re not completely unique).
* All points associated with one chromatography experiment should be either in training or in testing, but not within both.
* Yield and purity are calculated at specific cut points.
* Identify the ideal experimental setup for predicting high yield and purity for cut points that correspond to low experiment time.

Interactive visualization tool is the next step after models and sampling strategies.

## **4/26/21 Weekly Meeting Notes**

Reminder: Meeting with Just on Thursday from 10:30-11:30 AM!

Discussion of Tasks:

* Code for simple linear regression model, looking into optimization.
  + Julia started this this morning
  + Kaylyn will help if you run into trouble!
* Code for basic NN and starting the optimization process.
  + Kaylyn
* Code for probabilistic bayesian NN, looking into optimization.
  + Diego
* Literature Search/Modification of MM for building training sets (ongoing throughout project)
  + Benedicte is working on this
  + Also Julia

STANDUP ON FRIDAY: Kaylyn

BY THURSDAY, APRIL 29TH: Push all three separate model notebooks to Github, make progress on literature search/MM modification so we can ask any further questions to Just, somewhere within the optimization process for NN architecture. Able to compare and determine the best type of NN?

BY NEXT MONDAY, MAY 3RD: Update on FRIDAY after class.

## **4/19/21 Weekly Meeting Notes**

Reminder: Attend Wesley’s NN tutorial on Thursday from 1-2 PM.

Discussion of Tasks:

* Literature Search (input parameter ranges)
  + Julia
  + Benedicte
* Skeleton Code for NN (developing initial notebook where we can vary architecture to optimize)
  + Kaylyn
  + Diego (week 1)
* Building training sets w/ different sampling strategies--try uploading to the private github, if that doesn’t work talk to Kaylyn
  + Julia
  + Steven
* Visualization tool? (maybe talk about in the next few weeks)
  + Benedicte
  + Diego (week 2)

STANDUP ON FRIDAY: Benedicte

BY NEXT MONDAY, APRIL 26TH: Everyone has working skeleton code so that we can start varying architecture to optimize. We have training sets that we can train the NN on.

BY FRIDAY, APRIL 30TH: Optimization for NN architecture completed. Starting optimization for mechanistic model and building training sets. (Sensitivity analysis?)

Kaylyn will send an update email to Just on April 23rd with our progress and then schedule a meeting with Just for April 30th.

## **4/16/21 Meeting with Just Notes**

Questions to ask Just:

* Clarify specifically what we can put on the public repo (no datasets but NN notebooks?)
  + Anything that we develop is fine in our public repo.
* Diego has some questions about the bitbucket code.
* Input Parameters: do we need to develop new input parameters/types of chromatography/types of proteins?
* Input Parameter Ranges: is there a good range of parameter values for us to vary within? Where do we get this information?
* Sampling: how should we be picking from the range of the input parameters?
* Neural Networks: are there any suggestions for neural network design?

Target Protein + Impurity (need GEX v. Langmuir isotherm parameters for both)

Inputs: 2 salt concentrations (gradient elutions), linear flow rate, combinations of binding isotherm parameters (generate random combinations of molecular resin interaction parameters/salt concentrations)

Keep fixed column size

Sensitivity analysis w/ respect to different areas of the molecular resin interaction input space

Start building skeleton code to predict outputs from inputs (how does the size of the neural network or type of NN affect the error in prediction? How does the number of training examples impact the error? NN architecture, training set size, before influence of parameters)

Look at newer paper for possible input parameter range? 4 different proteins as a starting point. Check citations between two papers to get a handful of examples for proteins and binding isotherm parameters. PERFORM LITERATURE SEARCH

Predicting purity and yield is easier (predicting the chromatogram is not necessary).

Dynamically updating model? Vanilla deep neural network. Training data is likely going to have a bigger impact.

Adaptive Sampling: could we find the areas of the NN where the data isn’t as good and increase the amount of data to predict in that space?

Try fixed network architecture and update input space.

Visualization tools? Interactive interface that builds response surfaces like those in the Kramer paper.

\*\*Add slider bar ;)

Idea: All build one standard neural network and then each one of us iterates on a certain aspect of the design to figure out the trends and then optimize the network as well as the creation of datasets.

## **4/12/21 Weekly Meeting Notes**

Figure out bitbucket repo so we can get data (discuss with Just on Friday)

Read about/select a sampling strategy (discuss with Just on Friday)

What exactly can we put on the public facing repo? (discuss with Just on Friday)

Ask Wesley about building ANNs (schedule tutorial for Thursday)

-Discuss the best type of ANNs for this project

-Select # of hidden layers/optimizers/neurons/activation functions/etc.?