ChBE 6746/4746 Project Report
Title: Performing Process Optimization on an Organic Electronics Dataset

Group 1: Ian Clark, Aaron Liu, Lily Luan, Young Hee Yoon

Introduction:

Developing engineering solutions with polymer thin-film materials is often challenging due to the need to consider a vast design space of parameters. One such example exists within the field of polymer-based organic electronics; one of the most studied materials for polymer electronics research is poly(3-hexylthiophene) (P3HT), a polymer semiconductor that is well-reported in literature. In a typical process, the P3HT is dissolved in a solvent (analogous to an "ink") and cast onto a device substrate, where the performance is characterized using the **mobility** (μ *[=] cm2/V•s*) as a figure of merit. This mobility value is very sensitive to a variety of processing characteristics.

For example, recently, work within the Reichmanis/Grover groups at Georgia Tech compiled a small P3HT device dataset derived from about 20 papers from experimental literature. Though all data points are samples with the same polymer (P3HT), the range of reported mobility values spans 6 orders of magnitude (Figure 1).

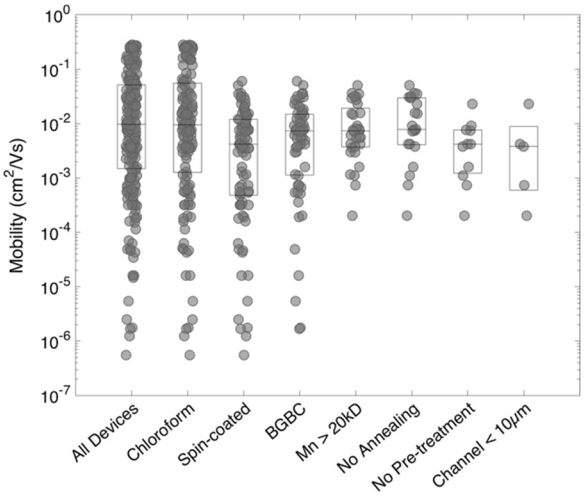

Figure 1. Reported mobility of devices in P3HT dataset satisfying progressively tighter constraints of commonly reported processing conditions. Response values span a range of six orders of magnitude [1].

Based on a preliminary data analysis that narrowed down important processing variables based on physicochemical intuition and domain expertise, it was found that the highest mobility values could be narrowed down to a design range [2]. However, neither a dedicated machine learning nor model optimization study has been explored at a deep level due to dataset challenges (missing values, sparsity, etc.)

Dataset Description and Cleaning:

The overall goal of this project was to explore some of the challenges associated with this dataset by applying concepts derived from the CHBE 6746/4746 course. Some important characteristics (and challenges) of this dataset include:

1. The size of the dataset is small, with 216 observations and approximately 30 variables.
2. The variables include a mix of discrete, continuous, and categorical variables.
3. Not all the data has been cleaned. For example, in one of the features, authors either reported a numerical value (i.e. 96%) or a string descriptor (i.e. "high" or ">90%") for the regioregularity.
4. The physical or mathematical model is not known for this problem, and behavior is expected to be nonlinear.

Given these dataset challenges, this project has focused heavily on data processing and cleaning in the early stages. Once the data was cleaned to an adequate level, we fit a handful of preliminary surrogate models attempting to model the complex behavior in the dataset.

The following table lists the original 32 features:

Author	Year	DOI	Mn (kDa)	Mw (kDa)	PDI	RR	Initial Concentration (mg/ml)
Solv1	Solv2	VFSolv1	Boiling Point (C)	Hansen Radius	Age Time (Hours)	Age Temp (C)	Substrate Treatment
Process Environment	Spin Rate (RPM)	Spin Time (s)	Dip Rate (mm/min)	Dip Time (min)	Film Thickness (nm)	Anneal Temp. (C)	Ann Time (Hours)
Mobility Environment	OFET Regieme	OFET Configuration	Channel Length (um)	Channel Width (mm)	Vds (V)	Electrode Material	Deposition Method

Initially, feature removal through data cleaning was performed as follows:

1. First, the features that are believed to be irrelevant to the scope of this project, or contain insufficient data were removed:

    a. Journal information {'Author', 'Year', 'DOI'} – may be important, but may not provide physicochemical insights
    b. Regioregularity {'RR'} – Might be important, but the information is not well represented in dataset. For example, while many reported values show up as the desired, exact numerical values (96%), a majority are less exact descriptors ('high', '>90%'). Since the representation for most of these observations is highly uncertain, we chose to remove the column.
    c. Ternary blend information, or use of a second solvent {'Solv2', 'VFSolv1'} – 192 of 218 do not report this variable. Given the lack of data, and the speculation that this behavior may be complex to model or impute with the available data, we chose to remove the feature.
    d. Redundant primary solvent information {'Solv1'} – Including the string name of a solvent would require excessive one-hot encoding and creating more features. Instead, since Boiling Point and Hansen Radius are reported, the desired solvent behavior is already contained in these features
    e. Features removed in favor of transforming information into simpler features (see 3rd point):

        i. {'Age Temp (C)', 'Age Time (Hours)'} → 'Aged' (Binary)
        ii. {'Anneal Temp. (C)', 'Ann Time (Hours)'} → 'Annealed' (Binary)
        iii. {'Spin Rate (RPM)', 'Spin Time (s)'} → 'Deposition_Method_SPUN' (Binary)
        iv. {'Dip Rate (mm/min)', 'Dip Time (min)'} → 'Deposition_Method_DIPPED' (Binary)

f. Some remaining features were hypothesized to not be **significantly** sensitive (by orders of magnitude) due to previous physicochemical knowledge (that would be nice to include if data was less noisy), or the noise in the data was similarly too difficult to account for: {'Process Environment', 'Mobility Environment', 'OFET Regime', 'Vds (V)', 'Electrode Material', 'Film Thickness', 'Channel Width (mm)', 'Channel Length (um)'}[2],[3],[4],[5].

2. Molecular weight information (M_w and M_n) and polydispersity (PDI) have a known mathematical relationship, and those with 2 out of 3 values reported can be filled in using the following:

$$PDI = \frac{M_w}{M_n}$$

Number average molecular weight (M_n) is reported for all values, and it may be desirable to remove one of the other two features due to their dependence.

Additionally, a correlation analysis found that the two molecular weight features (M_w and M_n) were highly correlated (Figure 2). Therefore, one of these two (M_n) was removed.

3. As mentioned above, some features were transformed from multiple string descriptors within one feature or multiple features in favor of a single binary variable indicating whether a particular process was employed. Some further details follow:

   a. 'Substrate Treatment': Many authors use a surface modifier to alter the interface between the polymer and the substrate. Though the mobility is ultimately somewhat sensitive to the surface modifier's chemistry, we simplify this feature to provide "substrate treated vs. untreated" information to avoid one-hot encoding multiple descriptors
   b. Annealing information was also transformed into a single binary variable by the following criteria: If the annealing temperature is ≤ 25°C or the annealing time is reported as '0', then the binary variable was set to zero; otherwise, the sample was annealed, and the value was set to '1'. A similar exercise was performed for Aging information.
   c. Since we will only consider whether the processing method was used or not, the categorical 'Deposition method' feature was one-hot encoded to generate binary variables with each category, 'DIPPED,' 'DROPPED', and 'SPUN'. The removed processing variables (see point 1) are important for some processes, but the behavior is unlikely to be captured here. Based on the diversity of values available, we hypothesize that the behavior related to these is not captured anyway.
   d. The other remaining categorical variable, 'OFET Configuration,' was also one-hot encoded into binary variables for each type, 'BGBC' and 'BGTC'.

4. Finally, all remaining rows that contained missing values in any feature were removed. 64 values were missing from 'Mw (kDa)' and 'PDI' and 16 from 'Initial Concentration (mg/ml)'. The number of data was reduced from 218 to 146, removing the 72 rows in total with missing values. In general, we believe the data is **missing at random**. There is a reason for the missing data (some authors chose to only report one molecular weight but not PDI or did not believe some variables were important to report – these rows are missing for all rows belonging to the same publication). However, the reasons for missing data may not be significantly related to the outcome.

After these data were cleaned and preprocessed, a dataset of 146 observations and 13 features remained. This approximately satisfies a heuristic criterion of 10 times the number of observations as features.

Exploratory Analysis and Basic Data Visualization

To build basic intuition with the remaining features in the cleaned dataset, each feature was plotted against the target output ('Mobility') to see if general trends or variable sensitivity could be observed. Two examples that generally match physicochemical expectations are shown in Figure 2 below. It has been reported in literature that larger degrees of polymerization (i.e., molecular weight) can correlate with improved mobility up to a certain threshold. This is observed in Figure 2(a), where the lowest mobilities are attributed to the lowest available molecular weights, but the mobility values appear to plateau at around $M_w \sim 25$ kDa. This one-dimensional analysis does not appear to explain the variations across 2-3 orders of magnitude past that threshold. In Figure 2(b), we see that all annealed samples generally have a higher frequency of larger mobility values than non-annealed samples. This would make sense, as annealing ensures proper removal of any residual solvent after processing, which may degrade device performance.

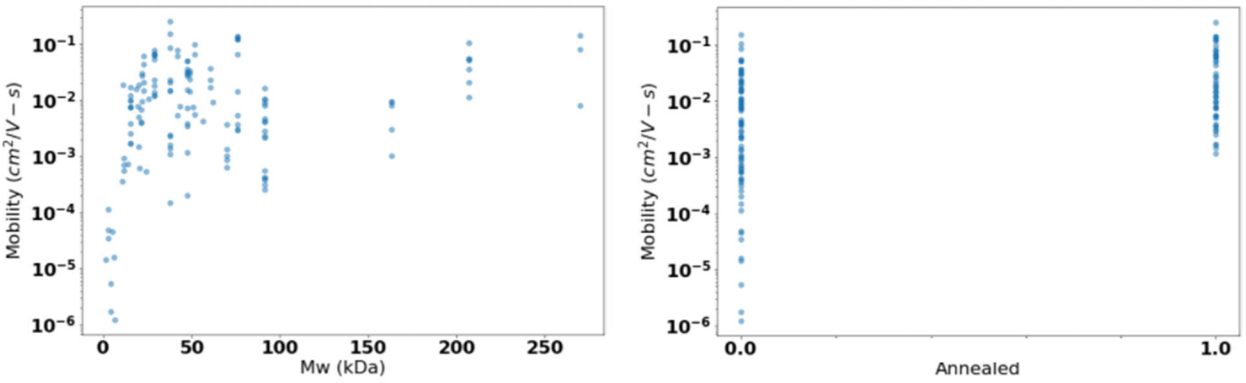

Figure 2. (a) Plot of molecular weight (M_w) vs. target output, mobility. (b) Plot of binary feature "Annealed" vs. target output, mobility. Plots for other features are available in Jupyter Notebook.

Since this dataset is a compilation from multiple literature sources (not simulated data), the available design space is fixed; however, Figure 3 shows various pairs of plots within the remaining continuous variables. Some features are sampled at diverse values, while other plots contain empty spaces that have not been sampled. For instance, experimenters frequently used chloroform (which has a boiling point around 334 K); this mode is clearly visible in the third row of plots in Figure 3 and indicates that possibly some regions are under-sampled. It is important to note that while we performed transformations and one-hot encoding to "simplify" the feature set, the presence of new binary variables creates an additional challenge in constructing accurate surrogate models. The handling of these binary variables was a key consideration in model fitting and optimization.

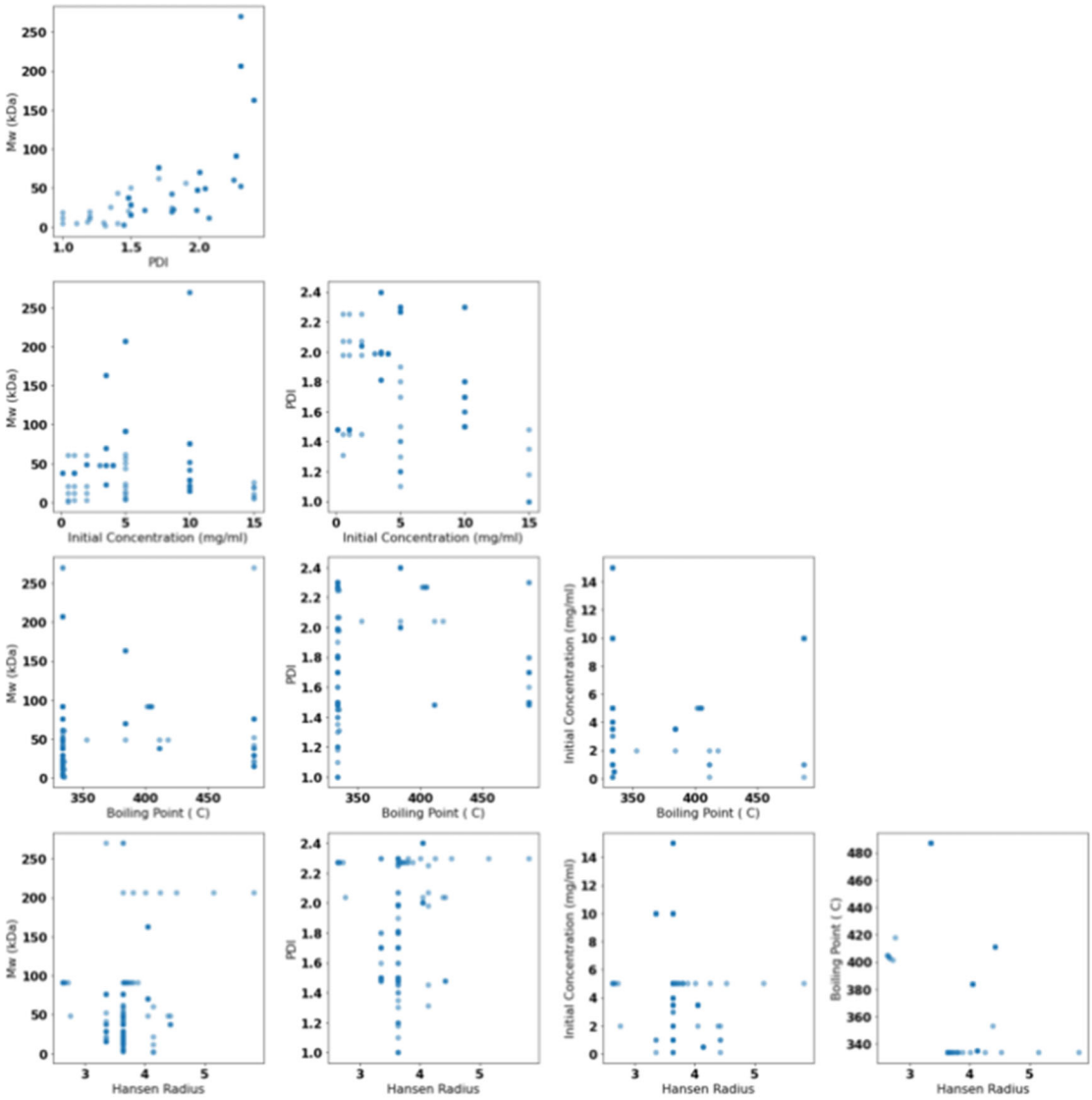

Figure 3. Visualization of 2-D design spaces for remaining continuous features in cleaned dataset.

Optimization Formulation and Method: After the number of features was trimmed, we attempted to fit various regression or data-driven surrogate models to the data. The objective function for optimization was formulated via the coefficients of the models, and the objective was to maximize the mobility. Various regression models were fit to the 13 independent variables and the response variable (mobility). The 13 features in the dataset will be used as variables, as follows:

- **Non-negative real variables**: Mw (kDa), PDI, Initial Concentration (mg/ml), Boiling Point (°C), Hansen Radius
- **Binary variables**: Substrate Treatment, Annealed, Aged, OFET configuration_BGBC, OFET configuration_BGTC, Deposition Method_DIPPED, Deposition Method_DROPPED, Deposition Method, SPUN

Coefficients for the relationship equation between the mobility and the variables were determined from the data-driven optimization modeling equation. For optimizing the resulting surrogate models, constraints were informed by hypothetical future experiment. These bounds and constraints were formulated from domain expertise for each variable, and knowledge that some binaries could only be turned on or off:

Table 2. Optimization bounds for the surrogate optimization

Variable	Lower Bound	Upper Bound
Mw (kDa)	1	120
PDI	1.0	2.5
Concentration	0	10
Boiling Point	334	487
Hansen Radius	2	6
All binary variables	0	1

Constraints:

```
Con1 = Constraint(expr = configuration_BGBC + configuration_BGTC == 1)
Con2 = Constraint(expr = deposition_SPUN + deposition_DIPPED + deposition_DROPPED == 1)
```

The behavior captured by the P3HT dataset is expected to be nonlinear, and likely nonconvex. Depending on preliminary analysis, the optimization problem to maximize mobility is likely to be a MINLP. Therefore, Baron or IPOPT solvers were used as appropriate.

Surrogate Models and Results

Prior to surrogate model fitting, we performed scaling on the dataset. For the input matrix, z-score scaling was performed by mean centering the data and dividing by the standard deviation for each continuous feature. This was favored over min-max scaling for the purpose of handling outliers. Binary variables were left unscaled because the scaled continuous variables were on a similar order of magnitude, and to simplify constraint handling for optimization later. As for the output matrix, the values were scaled by applying a logarithm to all the values, given the spread in orders of magnitude. Multiple surrogate models were fitted with the preprocessed dataset and were tuned and compared to determine how accurate a model could be fitted.

After scaling the data, four models (ordinary linear regression, linear terms with LASSO, quadratic terms with LASSO, linear SVR) were fit to the coefficients using Pyomo. Further, a Gaussian Process Model and Neural Network were fit using *scikit-learn*. **Figure 4** and **Table 2** show the parity plots for the models used and their respective R^2 values:

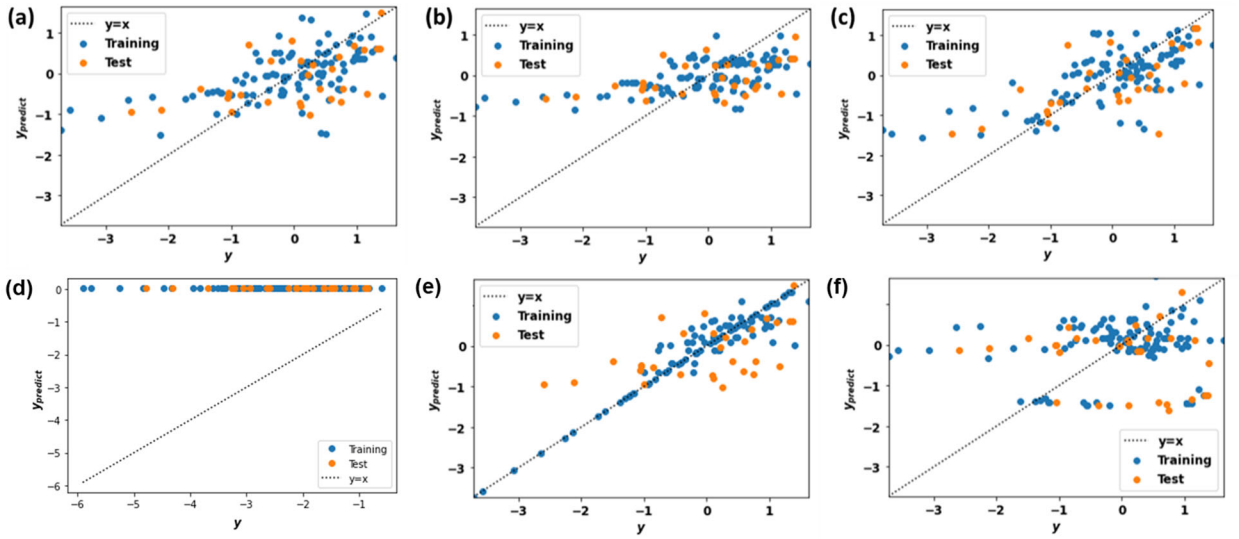

Figure 4. Updated parity plots for preliminary surrogate models. (a) Linear Regression; (b) LASSO (linear); (c) LASSO (quadratic terms); (d) Linear Support Vector Regression; (e) Gaussian Process Modeling; (f) Neural Network

Table 3. Preliminary result of fitted surrogate models and their accuracies on training set and testing sets

Model	MSE (Testing)
Linear Regression	0.753
LASSO (linear)	0.747
LASSO (quadratic)	0.622
SVR	6.105
Gaussian Process	0.975
Neural Network	1.411

As shown in **Figure 4** and **Table 3**, none of the surrogate models represent a great cross-validated accuracy with the preprocessed dataset. The Gaussian Process Model performs well with the training data, which is not surprising since the model should closely interpolate the input. However, the prediction of the testing data is not improved over the others. Given that the current iteration of the Neural Network does not perform well with the training data, perhaps better hyperparameter optimization would be needed.

For further evaluation, SVR and the neural network were removed from further consideration due to poor performance. For each remaining surrogate model, the mobility was maximized subject to the given constraints listed above in Table 2. For this report, we present the results from the four surrogate models (linear regression, linear-LASSO, quadratic-LASSO, and Gaussian Process Model) (Table 4):

Table 4. Optimization results for surrogate derived for selected models

Variable	Linear	LASSO (linear)	LASSO (quad)	GPM
Objective: Mobility [cm^2/V-s]	0.029	0.017	0.017	0.94
Mw (kDa)	120	120	120	120
PDI	2.5	2.5	2.5	2.29
Concentration (mg/mL)	10	10	0	5
Boiling Point (K)	487	487	487	334

Hansen Radius	6	6	4.54	4.01
Substrate Treatment	0	0	0	0
Annealed	1	1	1	1
Aged	1	1	0	1
OFET_Configuration_BGBC	1	1	1	1
OFET_Configuration_BGTC	0	0	0	0
Deposition Method_DIPPED	1	1	1	1
Deposition Method_DROPPED	0	0	0	0
Deposition Method_SPUN	0	0	0	0

The results show that the surrogate models mostly capture some behavior that might have been expected. For example, based on the basic visualization plots (Figure 2) and previous physical knowledge, we may expect the optimum to have a higher molecular weight (Mw) and greater polydispersity (PDI). Additionally, the models appear to have a consensus that certain processing steps such as annealing and aging are important to perform, while substrate treatment does not seem to be favored. Dip coating is also unanimously preferred for all samples, which is interesting because out of the three processing methods, dip coating is most likely to induce structural alignment of nanofibrils leading to improved mobility (in other words, this result might be physically explainable).

Generally, concentration is purported to lead to higher mobilities, but the optimal concentration is not always at the extremes of the specified bounds [0, 10], which shows that there may be an optimal value within that range. This might indicate an interesting variable to perform future design of experiments in the laboratory, since the models do not have a consensus. Other variables that do not have a strong consensus are Hansen Radius and Boiling Point. However, while this could also be interesting to explore further in a more controlled DOE, solvent selection is tricky since an experimenter cannot necessarily choose a single solvent that satisfies both a desired Boiling Point and Hansen Radius (solubility). Caution must also be taken in evaluating conclusions from this data. For example, BGBC seems to be favored for all devices, but this could be due to over-representation in that value; that is, most experiments in the dataset use BGBC, so it is possible that there were not any high-performing experiments that used BGTC. Further evaluation of these results might dig deeper into the data diversity within the retained rows.

Finally, the information lost from data cleaning should be considered. We believe the largest potential effect could be from omitting data based on lack of representation. For example, mobility is known to be sensitive to regioregularity (RR), but we were forced to omit this variable due to the heterogeneity of the data types and lack of specific descriptors (see earlier section). Additionally, the binary variables are a distilled version of the actual behavior reported in the papers. Instead of '1' vs. '0' for "Annealing", how long does a device need to be annealed to lead to the desired mobility? Though the exercise was driven by some domain knowledge, cleaning the original number of features from 32 to 14 certainly removed important information that could be useful to include in a richer version of this dataset.

Conclusions and Future Work

Overall, the preliminary model presented in this work looks to be able to model simple trends that may make physical sense. These results also show that the data cleaning that was comprehensively performed for this dataset did enable basic models, albeit improvement is needed by performing future work. One targeted area of improvement could be more careful handling of the binary variables; some suggestions are below.

Moving forward, further work on surrogate model fitting might be implemented to attempt to improve the prediction of the actual data. Given that the dataset may be noisy and difficult to fit, we are also interested in breaking the dataset into smaller, local problems based on what information is available. Some ideas include:

1. Determine the variables that are more meaningful than others and assign different weights on the variables accordingly.
2. Surrogate model fitting using subsets of data can be performed:
    - Fitting on continuous variables only
    - Turning "on" and "off" binary variables one at a time; for example, fit a surrogate to the data while holding 'Aged' = 0 or 'Aged' = 1, and make a comparison to the two cases based on the resulting surrogates
3. Conduct efficient global optimization (EGO) to determine where to sample next for future research experiments. This could also involve extracting and analyzing the error function from the GPM.

Additionally, a deeper analysis of what information was lost during data cleaning is also important for evaluating results. This can be done through a set of hypotheses, such as "how biased are the observations toward using chloroform as a solvent?" and making individual evaluations on a case-by-case basis.

Individual Contributions

- **Project Proposal**
  - Writing: AL, IC, YY
  - Editing: LL
  - Project Ideas
    - P3HT: AL
    - Energy of Buildings dataset: IC
    - Various porous polymer membrane datasets: YY
- **Dataset Preparation**
  - Dataset cleaning: AL, IC
  - Discussion of cleaning steps: LL, YY, AL, IC
  - Writing: YY, AL, IC
- **Optimization Formulation and Method**
  - Writing: IC, AL, YY
  - Constraint generation based on domain knowledge: AL
  - Constraint writing for python: IC
  - Application of constraints to code: YY, LL
- **Surrogate Model Construction and Optimization**
  - Data visualization: AL, IC
  - Linear, Ridge, LASSO in scikit-learn for baseline comparisons to Pyomo: IC
  - Linear pyomo model + code template for other models: YY
  - LASSO + linear pyomo model, optimization, rescaling: LL
  - LASSO + quadratic pyomo model attempt: AL
  - SVR model attempt (sklearn): IC
  - Writing: AL, YY, LL
  - GPM and NN surrogate: AL, YY
- **Final Report**
  - Writing: AL, IC
  - Editing: YY, LL
  - Figures: AL, IC, YY, LL

References

1. Persson, N.; McBride, M.; Grover, M.; Reichmanis, E., Silicon Valley meets the ivory tower: Searchable data repositories for experimental nanomaterials research. *Curr. Op. in Solid State and Mater. Sci.* **2016**, 20 (6), 338-343.
2. McBride, M.; Persson, N.; Reichmanis, E.; Grover, M., Solving Materials' Small Data Problem with Dynamic Experimental Databases. *Processes* **2018**, 6 (7).
3. Choi, D.; Chu, P.-H.; McBride, M.; Reichmanis, E., Best Practices for Reporting Organic Field Effect Transistor Device Performance. *Chemistry of Materials* **2015,** *27* (12), 4167-4168.
4. Chang, M.; Lee, J.; Kleinhenz, N.; Fu, B.; Reichmanis, E., Photoinduced Anisotropic Supramolecular Assembly and Enhanced Charge Transport of Poly(3-hexylthiophene) Thin Films. *Advanced Functional Materials* **2014,** *24* (28), 4457-4465.
5. Sirringhaus, H., Device Physics of Solution-Processed Organic Field-Effect Transistors. *Advanced Materials* **2005,** *17* (20), 2411-2425.
6. Wu, D.; Kaplan, M.; Ro, H. W.; Engmann, S.; Fischer, D. A.; DeLongchamp, D. M.; Richter, L. J.; Gann, E.; Thomsen, L.; McNeill, C. R.; Zhang, X., Blade Coating Aligned, High-Performance, Semiconducting-Polymer Transistors. *Chemistry of Materials* **2018,** *30* (6), 1924-1936.