# Classification using Sum of Ranking Differences of Outlier Measures

**Tony Lemos, John H. Kalivas**
Department of Chemistry
Idaho State University
921 S. 8th Avenue, STOP 8023 Pocatello, ID 83209, USA
lemoton2@isu.edu, kalijohn@isu.edu

## Abstract

A useful application in analytical chemistry is classifying unknown samples into classes. Single-class classification is a type of classification approach where only one well-defined class is of interest. Outlier detection is useful for defining class membership for unknown samples, since outlier detection removes samples that are not represented by the sample class space. When using outlier detection, there are two problems: which outlier measure to use and the tuning parameter value for the chosen outlier measure. The proposed technique for single-class classification using outlier measures eliminates these two problems. To avoid selecting any one particular outlier measure, multiple measures are evaluated by using sum of ranking differences (SRD). The method of SRD is used to evaluate multiple outlier measures to obtain a consensus in classifying a sample. In regards to tuning parameters, a parameter window is used to avoid doing more work, such as having a training set of samples to select a tuning parameter. Wavelength selection and fusing spectra from different instrument is used in conjunction with SRD to provide a robust characterization of the class of interest. Presented are results for the new classification approach on spectral food data sets.

## Objectives

- Create a simple procedure to perform one-class classification
- Utilize multiple outlier measures to obtain a consensus in classifying a sample

## Background

Definitions
- Target class – Class of interest
- Non-target samples – Samples not belonging to the class of interest

Two types of classification techniques
- Discriminant Classification
  - Samples are classified into more than one predefined class
- One-class classification
  - Samples are classified into one predefined target class

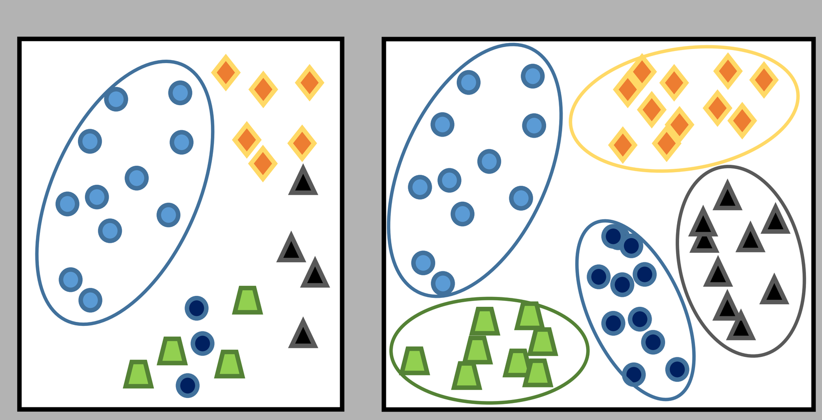


Figure 1 – Classification scenarios: One-class classification (left), discriminant classification (right)

Outlier Detection
- Outlier - An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs[1]
- Outlier detection is one-class classification have same principal idea
  - Differentiating between data that appears normal (belonging to a class) abnormal
- Difference: Application
  - Outlier detection – Which samples are not conforming to the normal behavior of similar samples?
  - One-class classification – Is this sample behavior similar enough to the other samples to belong to their class?

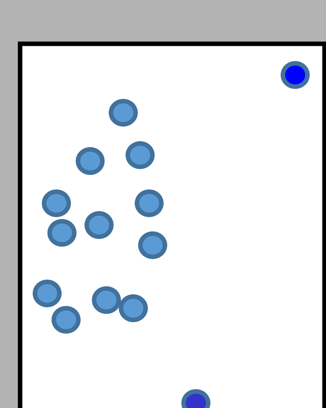


Figure 2 – Sample observations

[1] Barnett, V. and Lewis, T.: 1994,Outliers in Statistical Data. John Wiley & Sons., 3rd edition

## Approach

- 17 outlier measures

| |
|---|
| **Comparing sample to class** |
| Mahalanobis Distance (MD) |
| Q-residual (Q) |
| Sinβ |
| Divergence Criterion (DC) |
| **Comparing sample to mean class** |
| Determinant |
| Euclidean Distance |
| Inner product correlation |
| cosθ |
| Constrained Procrustes Analysis |
| Unconstrained Procrustes Analysis |
| Extended Inverted Signal Correction Difference |

Table 1 – List of outlier measures used for one-class classification

Require a tuning parameter (up to rank r number of eigenvectors)

1, 1–2, 1–3,…, 1–r eigenvectors
- Involves training each measure

Introducing a tuning parameter window
- Diversifying the collection of outlier measures
- Simplifies classification

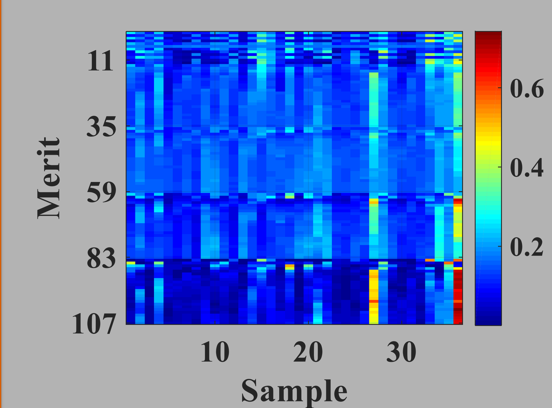
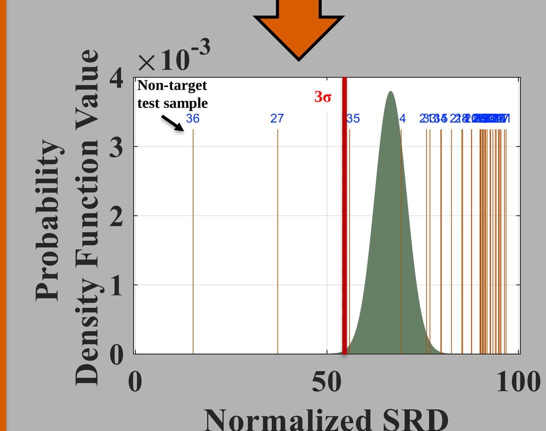



Figure 2 – Outlier measures scaled to unit length (across rows) classification. Merits: 1–11 vector to mean, 12–35 MD, 36–59 Sin(β), 60–83 Q, 84–107 DC

Sum of ranking differences (SRD)
- Comparison of columns (samples) across rows (merit)
- Determines a rank for each sample




Comparison of ranks by random numbers (CRRN)
- Determine the probability that the SRD sample rankings is not a random ranking

Figure 3 – The SRD normalized rankings of each sample with the random ranking distribution and the 3σ threshold

Classification Quality Measures

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TN}{TN+FP}$$

$$\text{Specificity} = \frac{TP}{TP+FN}$$

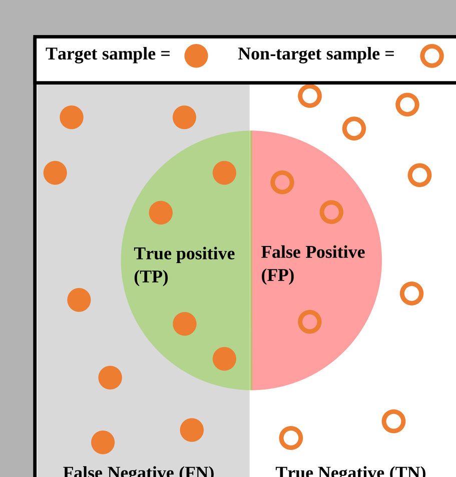



Figure 4: Illustration of true positive, true negative, false positive and false negative

## Data sets

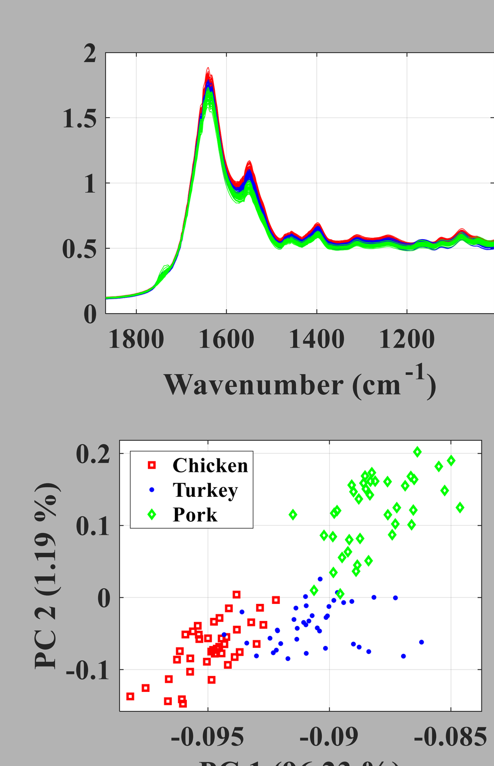



Figure 5: Spectra (top) and the principle component (PC) plot (bottom) for each meat

Meat Mid-infrared (MIR)
- 40 samples for each class

Process:
- 5 samples from each class for validation
- Maximum tuning parameter window: 24
- 10 splits

Strawberry puree MIR data
- 351 strawberry samples
- 632 non-strawberry (strawberry adulterated with other fruits) samples
- Process
  - 30 validation samples from each class
  - Outlier clean the target class
  - 10 splits
  - Stack wavelength regions

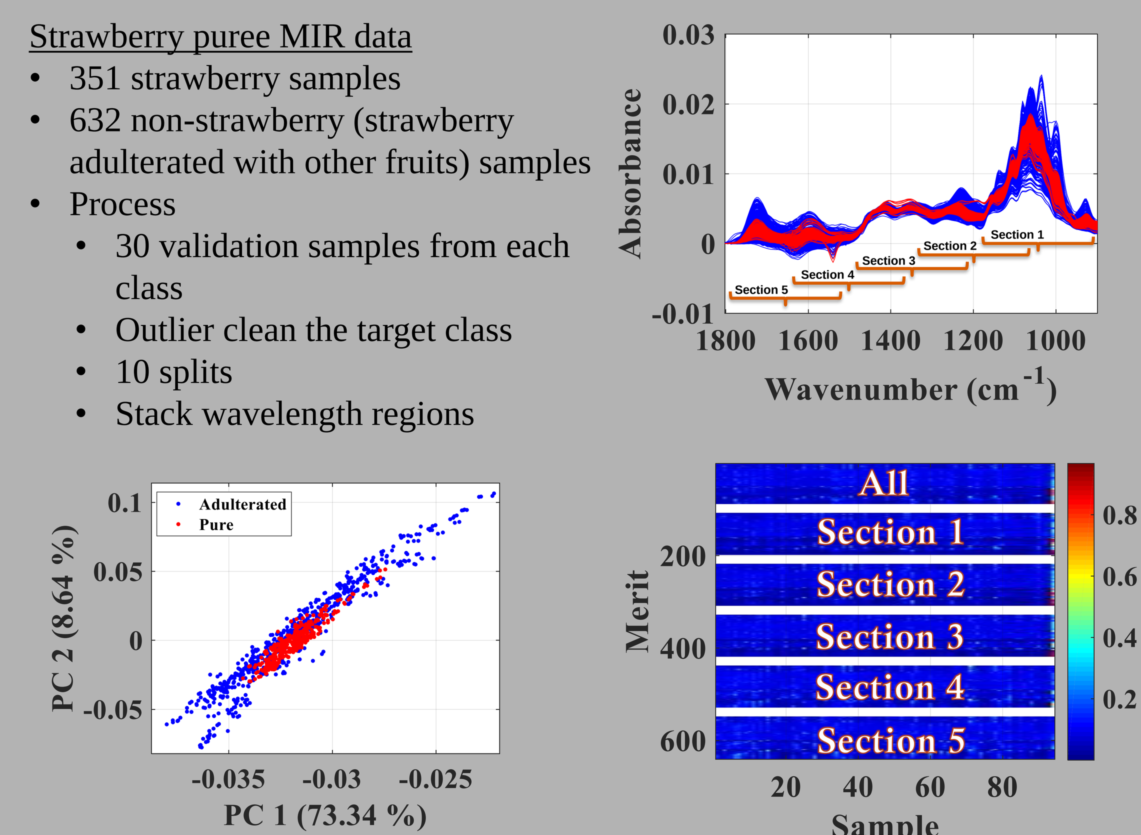



Figure 6: Spectra with wave selected regions (top), the principle component (PC) plot (bottom left) for the pure and adulterated strawberry samples and the SRD input (bottom right)

Italian Beer
- Classes
  - 19 Birra del Borgo – ReAle (target)
  - 41 other craft beers – 'non-ReAle'
    - 12 Birra del Borgo
    - 29 different location
- Measured on 5 instruments

Process:
- Stack the instruments
- 3 validation samples from each class
- Maximum tuning parameter window: 12
- 20 splits

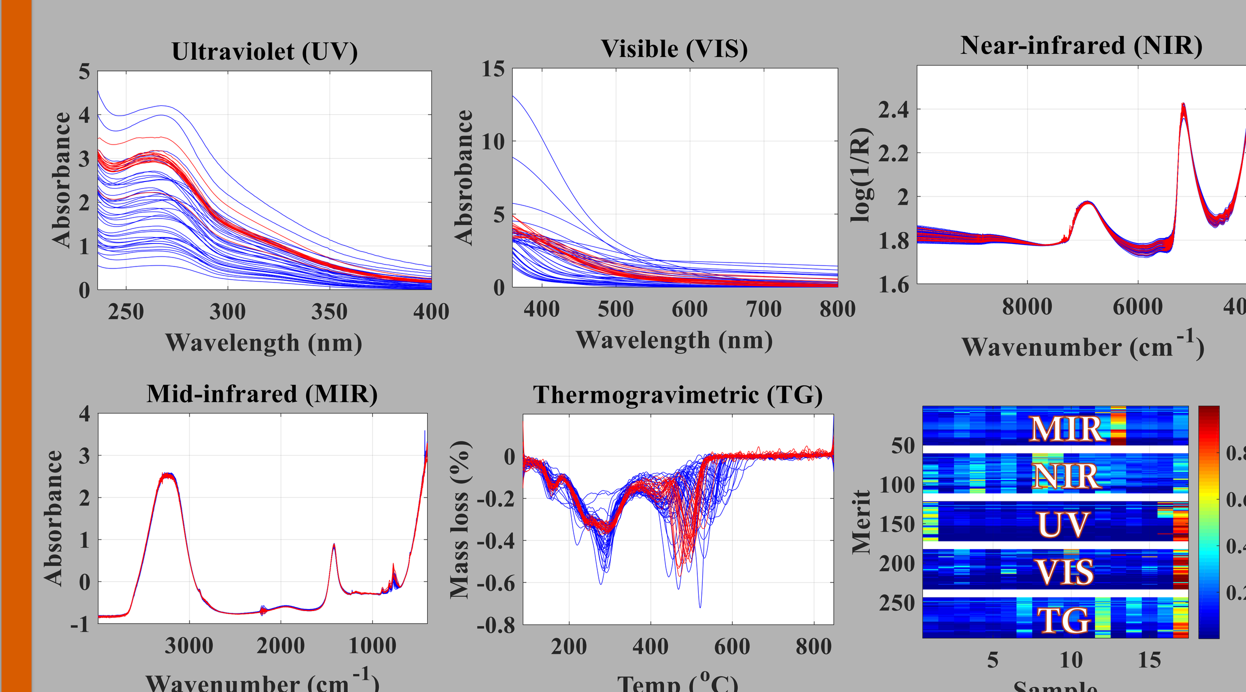



Figure 7: Spectra for ReAle and non-ReAle bear on each instrument. The stacking of the instruments is illustrated on the bottom right figure.

## Results

Meat Results

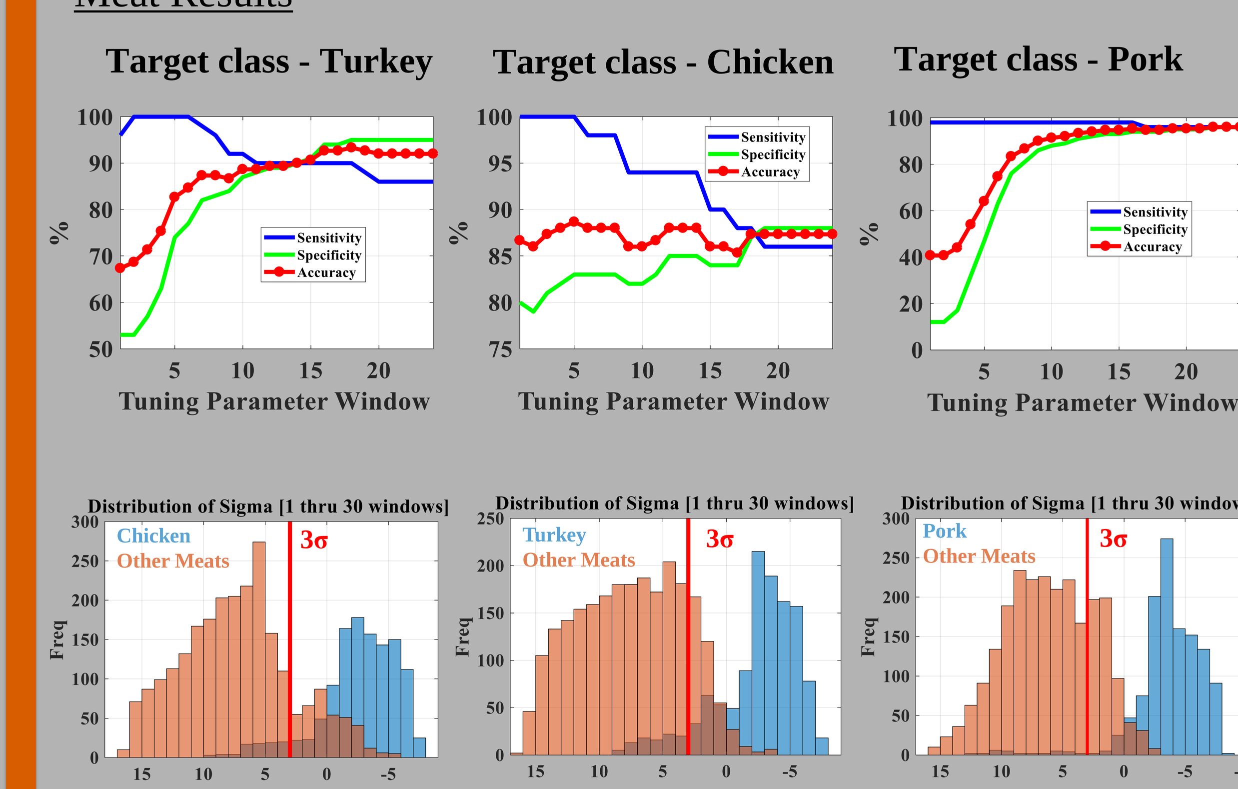



Figure 8: The overall meat results (top row) and the distribution of the sigma for each validation sample across each tuning parameter window (bottom row)

Strawberry Puree Results

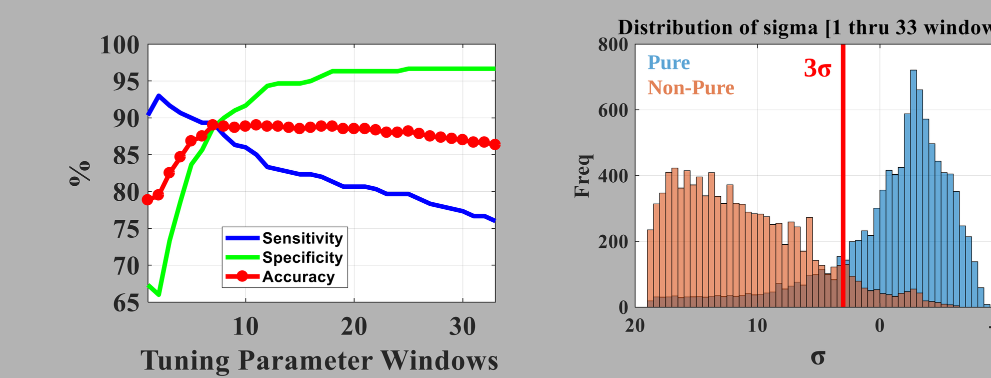



Figure 9: The overall strawberry results (left) and the distribution of the sigma for each validation sample across each tuning parameter window (right)

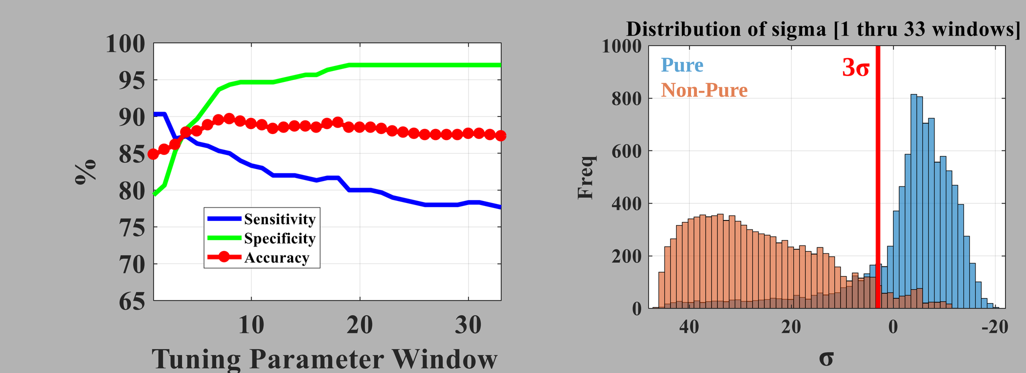



Figure 10: The overall strawberry results (left) and the distribution of the sigma for each validation sample across each tuning parameter window (right)

Italian Beer Results

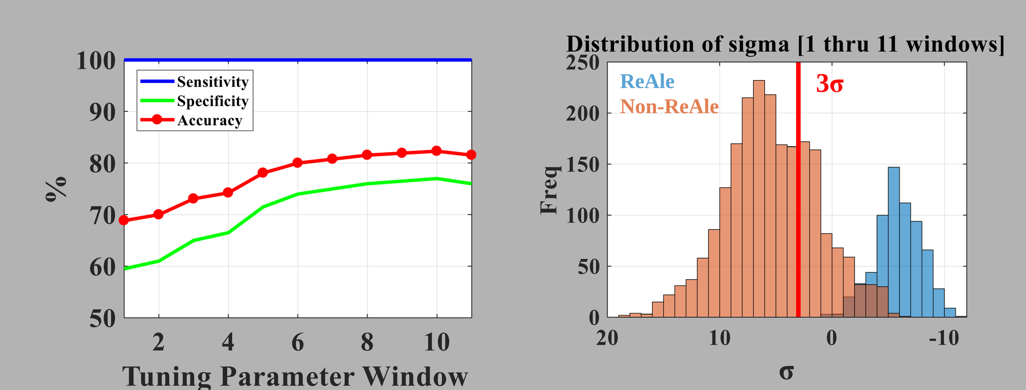



Figure 11: The overall beer results (left) and the distribution of the sigma for each validation sample across each tuning parameter window (right)

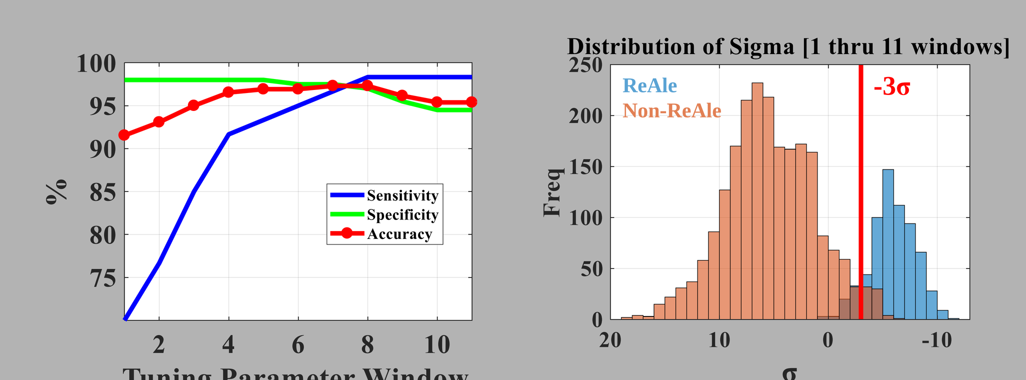



Figure 12: The overall beer results (left) and the distribution of the sigma for each validation sample across each tuning parameter window (right) with -3σ threshold

## Conclusion

- SRD is an effective one-class classification technique
  - Generally increases in accuracy at higher windows
  - Flexibility of SRD
    - Outlier measures
    - Instruments
    - Preprocessing methods
  - Tuning parameters
    - Tuning parameter window
    - Adjust sigma threshold

**Acknowledgements:** Work supported by the National Science Foundation under grant No. CHE-1506417 (co-funded by CDS and E Programs) and is gratefully acknowledged by the authors.