Applying Statistical Distance Metrics for Dimensionality Reduction

Anna Bell

THE UNIVERSITY OF UTAH

Introduction

A dimensionality reduction problem in data chemistry is addressed using statistical distance measurements. Pairwise combinations of chemical features are scored for their clustering of points labelled for their favorability in asymmetric catalysis. The meaningfulness of each score is validated using a permutation testing method.

Motivations in Data Chemistry

Finding Hotspots in Complex Catalytic Reaction Spaces

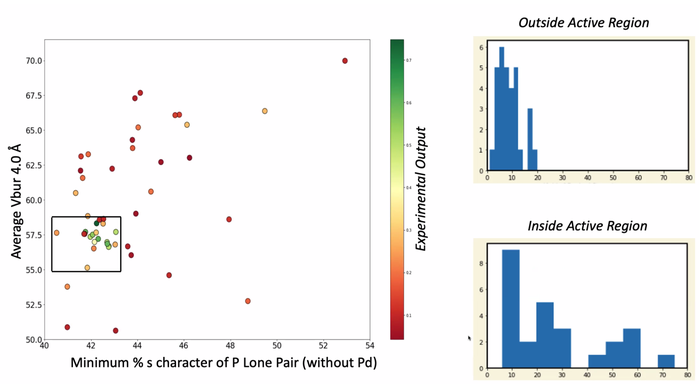
At the University of Utah's Department of Chemistry, the Sigman Lab is searching for "hotspots."

A hotspot may be described as densely concentrated target reactions resulting from experimental asymmetric catalysis.

When catalytic reactions are not linearly separable or easily modeled, hotspots in a lower-dimensional space identify features important to selectivity in organometallic reactions.

Hotspots also inform new catalyst and ligand design as well as predict reactivity for certain catalytic reactions.

Example of a 2-Dimensional Feature Space
Containing a Hotspot, Hand-Selected by Dr. Lucy Van Dijk

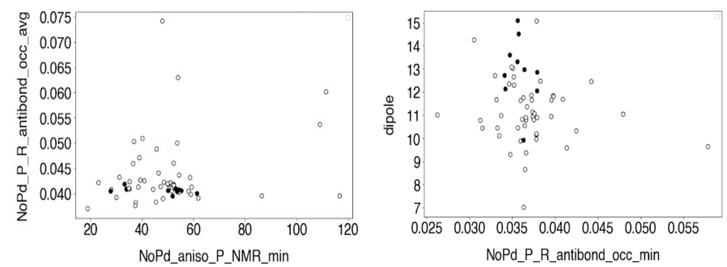


The Hotspot Search Algorithm

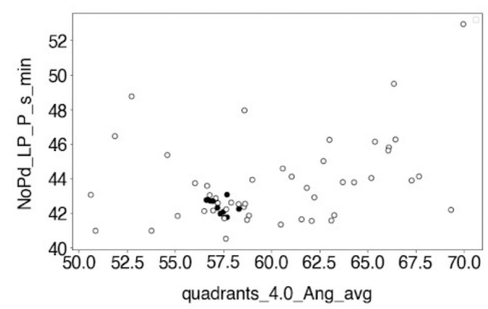
A hotspot is scored by evaluating its:

- **density** of desirable outcomes,
- **variance** of desirable outcomes from all others,
- **significance** of the distribution of desirable outcomes compared to the distribution of all outcomes.

Examples of Low-Scoring Feature Combinations
(which got high scores with other scoring methods)



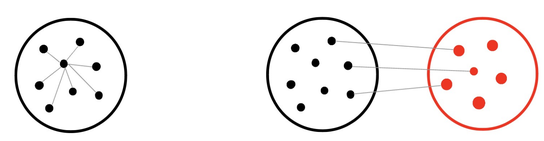
Example of a High-Scoring Feature Combination
(using the S_Dbw scoring method)



(target reactions are back, the rest are white)

Scoring Method

A hotspot should contain desirable outcomes that are compact and well-separated from the other data points. The S_Dbw* score measures the cohesion and separation to score a potential hotspot.



Cohesion Separation

* M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set."

Algorithm Steps

For each pairwise combination of features:

1. Score the hotspot for cohesion and separation:

   S_Dbw = IV + ID

   IV: "**Intra-cluster Variance**"
   ID: "**Inter-cluster Density**"

2. Determine whether the S_Dbw score is meaningful using a permutation method that generates a p-value for the **significance of a hotspot's distribution.**

Null Hypothesis: The S_Dbw score is the result of chance. Future scope expansion are unlikely to produce similar reaction outcomes.

Alternative Hypothesis: The hotspot's distribution is independent. Catalytic experiments with compounds identified by the hotspot yield similar results, providing new insights to the mechanism of asymmetric catalysis.

Thank You

Dr. Jeff Phillips
University of Utah
School of Computing

Dr. Matthew Sigman
University of Utah
Department of Chemistry

Dr. Meysam Alishahi
University of Utah
School of Computing

Dr. Lucy Van Dijk
University of Utah
Department of Chemistry