Expanding the addressable chemical space with libraries of computed spectra

Calvin Stevenson, Ty Abshear, Graeme Whitley, Michelle D'Souza, Ph.D.* | **John Wiley & Sons, Inc.**

Results

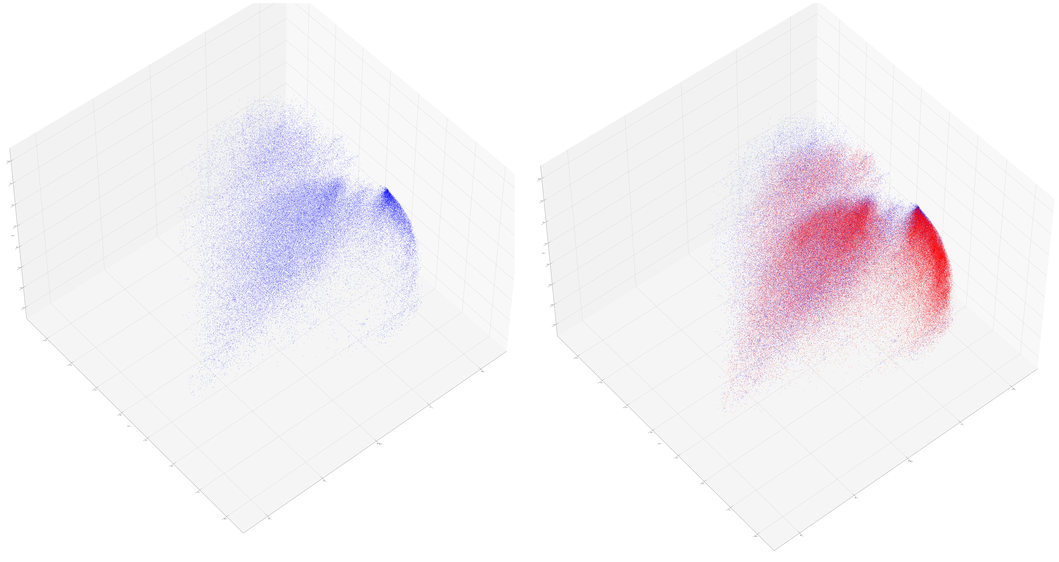

Figure 1: The first plot shows the empirical spectral datasets alone (blue), while the second plot shows the empirical datasets along with the SmartSpectra (red) included to demonstrate how it augments the chemical space.

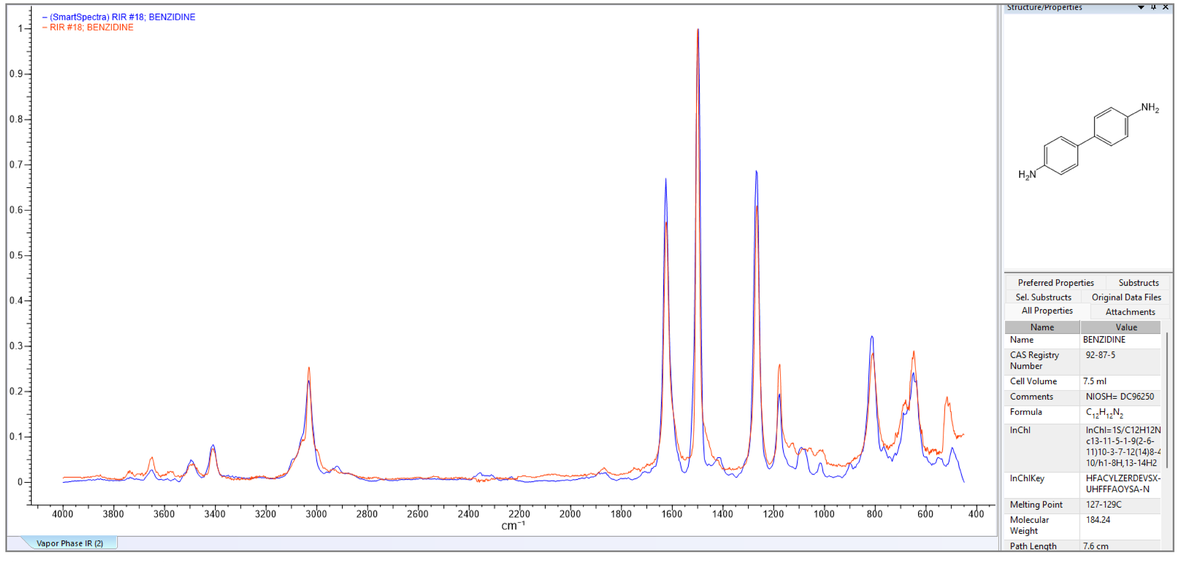

Figure 2: The resulting overlap of SmartSpectra Vapor Phase (blue) vs. empirical Vapor Phase (orange). Here you can compare the individual result of a SmartSpectra record to an empirical Wiley example.

Abstract

Using an AI-powered spectrum prediction engine derived from its high-quality, comprehensive databases of measured spectra is a current strategy to expand chemical compound coverage by generating computed spectral data. Augmenting coverage of empirical databases within the bounds of a model (the chemical space of the underlying training set) is a strategy to help improve overall available compound coverage for unknown identification, especially for rarer compounds and materials.

Our validation studies on each of the SmartSpectra computed datasets demonstrate that these computed libraries, constructed from extensive and high-quality empirical reference datasets, demonstrate performance levels closely approaching that of empirical datasets.

Method

The data used for these libraries is owned by Wiley. The software used was KnowItAll 24.0.59.0. The structure data predicted with the SmartSpectra algorithm had to be contained in the chemical structure fingerprint of the model, where the chemical structure can be created entirely by parts of the model's chemical structure fingerprint fragments.

The prediction models were validated rigorously using external validation studies. The hit lists were used as an accurate field test for how users would experience using the library. Two different tests were run: replicate hit list analysis and test set hit list analysis. Essentially one test checks how well predictions and replicates perform against empirical data, while the other test checks to see how the predictions perform on external data sets. For these tests, spectrum searching was used. The search algorithm used the KnowItAll correlation algorithm.

To see the full validation papers, visit https://sciencesolutions.wiley.com/.

Table 1. Validation statistics

Data sets	Average Hit List Position	Top 10 Hit Percentage
FT-IR, Replicate Analysis		
Validation Test 1: Replicates Searched on Empirical Test Set (5%)	3.97	94%
Validation Test 1: Replicates Searched on SmartSpectra Test Set (5%)	2.94	97%
Validation Test 2: Replicates Searched on Empirical Test Set (10%)	1.11	100%
Validation Test 2: Replicates Searched on SmartSpectra Test Set (10%)	9.45	85%
Raman, Hit list Analysis		
Model Test Set	6.1	91%
JASCO Test Set10	13.1	82%
JASCO Test Set with Sadtler Empirical Data11	14.6	81%
JASCO Test Set with Sadtler Empirical Data (Outliers Removed)	5.7	89%
Vapor Phase IR, Hit list Analysis		
Validation Test 1: Replicates Searched on Empirical Test Set	1.05	100%
Validation Test 1: Replicates Searched on SmartSpectra Data	1.05	100%
Validation Test 2: Wiley's Vapor Phase IR Model Test Set	4.01	93.5%
Validation Test 3: Sigma-Aldrich test set (included in the KnowItAll IR Spectral Database Collection)12 Vapor Phase IR Library Test Set	7.60	90.3%
Validation Test 4: Sigma Aldrich Vapor Phase IR Library Test Set with additional Wiley Data included in the searched databases	8.94	88%

Summary

Wiley Science Solutions has developed and used a set of validation tools and predictive algorithms to generate multiple libraries of computed spectra within the bounds of our current libraries' chemical space. These SmartSpectra collections currently contain 250,000 FT-IR records, 215,427 Vapor Phase IR records, and 33,163 Raman records.

Based on our validation studies, we have determined that the computed libraries demonstrate a high level of performance approaching that of empirical databases. These libraries have shown the ability to characterize and classify unknowns by enhancing the coverage within the bounds of Wiley's empirical data chemical space.