Electron Density-Based Machine Learning for Accelerating Quantum Calculations

Joshua Lansford and D. G. Vlachos

2019 Blue Waters Symposium, Sunriver OR
June 5, 2019
**Materials Gap in Catalysis: Theory and Experiments**

Physics + Data science\(^{[4]}\) is needed to understand both *dynamic changes*\(^{[4]}\) and static properties of complex materials.

---

Infrared (IR) spectroscopy of dispersed Pt atoms and nanoparticles for CO oxidation

- Vibrational spectroscopy is a precise (<1% uncertainties) surface technique that is rapidly advancing.
- Spectra are relatively insensitive to temperature and can be used in-situ or operando

The Argument for CO as a Probe Molecule

<table>
<thead>
<tr>
<th>Site-type</th>
<th>Freq. [cm(^{-1})]</th>
</tr>
</thead>
<tbody>
<tr>
<td>atop</td>
<td>2070</td>
</tr>
<tr>
<td>bridge</td>
<td>1830</td>
</tr>
<tr>
<td>fcc</td>
<td>1760</td>
</tr>
</tbody>
</table>

- C-O frequency depends on both site-type and site coordination
- C-O has well defined peaks that can be visually identified by the human eye and brain
- There are no quantitative methods to determine surface structure from vibrational spectra

Exp. CO Frequency on Pt(111)\([1]\)

Low coverage adsorption, at various temperatures\([2]\)

\[ \omega = 1,997 + 10.0 \text{CN} \text{ cm}^{-1} \]

Goal
• Determine local microstructure of Pt nanoparticles from experimental vibrational spectra using CO as a probe molecule

Plan
• Assess accuracy of DFT in recreating IR spectra
• Provide an overview of surrogate modeling
• Combine data science techniques with expert knowledge to better understand data and improve sampling, highlighting data visualization
• Illustrate key details of the surrogate models for generating synthetic IR spectra and learning the corresponding local structure
• Show model results and provide an application to experimental vibrational spectra
C-O stretch frequencies for CO at an atop site

Generalized Coordination Number (GCN) is a coordination number weighted by second nearest neighbors \([1]\)

\[
\overline{CN}(i) = \sum_{j=1}^{n_i} \frac{cn(j)n_i}{cn_{\max}}
\]

[Calle-Vallejo et al., Angew. Chem. Int. Ed. 53, 8316 (2014).]

C-O frequency is a descriptor of local structure but in experiments we must untangle spectra generated from many CO molecules on many different GCNs – We need intensities!
Methods: Generating Spectra from First Principles

Compute Intensities using the derivative in dipole moment $\mu$ (dynamic dipole moment) with respect to the normal mode displacement ($Q$).\footnote{Porezag and Pederson. Phys. Rev. B. 54, 11 (1996)}

$$I_{i}^{IR} = \frac{\mathcal{N} \pi}{3c} \left| \frac{d\mu}{dQ_i} \right|^2$$

$$\frac{d\mu}{dQ_i} = \sum_{k=1}^{3N} \frac{\partial \mu}{\partial R_k} X_{ki}$$

- Normal mode (hessian of the forces) for identifying peak locations (frequencies)
- VASP\footnote{G. Kresse and J. Furthmüller, Phys. Rev. B. 54, 11169 (1996).} for computing electron densities
- CHARGEMOL\footnote{T. A. Manz and N. G. Limas, RSC Advances 6, 47771 (2016).} for integrating over the electron densities to get the dipoles
- Matrix product of the dipole Jacobian and the normal mode vectors to compute intensities


1) Existing literature supports accuracy of measuring and computing frequencies on surfaces\(^1,2\)

2) There is not always a one-to-one correspondence between intensity and concentration

3) There are more frequencies than just the C-O stretch

---

IR Spectra of CO on Pt(111) with a c(4x2) Overlayer

DFT generated spectra reproduces experimental spectra (frequencies and intensities)


Surrogate Model Overview: Iterative Design

DFT Data → Simulated Spectra → Surrogate Spectral Model → Multinomial Regression → Local Structure

Synthesizing Spectra
- Outlier removal
- Harmonic approx.
- Lateral interactions
- Spectral mixing
- Convolution

Surrogate Model Overview: Iterative Design

DFT Data

Simulated Spectra

Local Structure

Surrogate Structure Model
Outliers inhibit learning both because they result in large gradients during training and because there are not enough samples with similar feature values to predict them.
Removing samples that are **not local minima** on the potential energy surface applies expert knowledge to remove unphysical outliers.
Surrogate Model Details: The Activation Function

\[ f_i = \frac{e^{a^T w_i}}{\sum_{k=1}^{K=4} e^{a^T w_k}} \]

Input Layer (501 intensities)

Hidden Layer(s)

Output Layer (site-type/GCN range)

- \( f_1(X) \) (\% Atop)
- \( f_2(X) \) (\% Bridge)
- \( f_3(X) \) (\% 3-fold)
- \( f_4(X) \) (\% 4-fold)
Wasserstein loss with $[0,0,0,1]$ for three probability sets compared to the single-valued kl-divergence.

$$W^2 = \sum_{n=1}^{C} \left[ \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} t_i \right]^2$$

Kl-divergence compares probabilities between two distributions at each index ($p_i$ and $t_i$), while Wasserstein compares the cumulative probability at each index ($CDF(P)_i$ and $CDF(T)_i$) and takes into account inter-class relationships [1].

Model Results: Site-type Histogram

Example of Synthetic Spectra

Actual and Predicted Histograms

- 3-fold and 4-fold
- Atop
- Bridge

Relative Intensity vs. Frequency [cm\(^{-1}\)]

CO coverage distribution

- Atop
- Bridge
- 3-fold
- 4-fold

Simulated Distribution
Model Prediction
Model Results: Generalized Coordination Histograms

Simulated GCN Histogram

- Green bars: Simulated GCN Distribution
- Purple bars: Model Prediction

GCN Groups Determined by Clustering

<table>
<thead>
<tr>
<th>GCN group</th>
<th>GCN values</th>
<th>GCN group</th>
<th>GCN values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-1.8</td>
<td>7</td>
<td>5.5-6.1</td>
</tr>
<tr>
<td>2</td>
<td>1.8-2.8</td>
<td>8</td>
<td>6.1-6.6</td>
</tr>
<tr>
<td>3</td>
<td>2.8-3.7</td>
<td>9</td>
<td>6.6-7.2</td>
</tr>
<tr>
<td>4</td>
<td>3.7-4.3</td>
<td>10</td>
<td>7.2-7.9</td>
</tr>
<tr>
<td>5</td>
<td>4.3-4.9</td>
<td>11</td>
<td>7.9-8.5</td>
</tr>
<tr>
<td>6</td>
<td>4.9-5.5</td>
<td>12</td>
<td>High Coverage Low-index planes</td>
</tr>
</tbody>
</table>

CO coverage distribution vs. Generalized Coordination Group
Experimental Application: Spectra from Literature

Experimental Spectroscopy

- Pt(111) (c4x2)
- Pt(111) 0.17 ML CO
- Pt(110)
- 55 nm Au@0.7 nm Pt/Pt

CO saturated 0.5 M H₂SO₄ at -0.1 V*

STM of 55 nm Au @0.7 nm Pt/Pt

Pt(111) c(4x2) 0.5 ML

Pt(110) 1 ML

* A voltage of -0.1 V will only shift the C-O frequency by 2.9 cm⁻¹.
A combination of LEED and TPD measurements tell us that at 0.5 ML this c(4x2) overlayer results in 50% atop and 50% ridge sites. At high pressures this spectra could correspond 62% atop and 38% bridge.

Trends in LEED studies suggest at low coverages almost all CO is adsorbed at atop sites on Pt(111)

Pt(110) can undergo reconstruction, however, at the maximum coverage of 1 ML it is observed to deconstruct with all CO in the atop position.

Because the nanoparticle system is in liquid, coverages are low. This would preclude ordered high spatial overlayers of the low-index planes. The uniformity of the nanoparticles would suggest that most occupied sites are at a low-index plane of the same site.

The supposed high-coverage Pt(110) surface has significant 4-fold contribution. This is unexpected.
The parts of the spectra resulting in predicted adsorption at 4-fold sites for Pt(110) (yellow line) is likely due to the extended tail below 400 cm\(^{-1}\) and the slight bump at 1700 cm\(^{-1}\).
Conclusions

• We are able to synthesize spectra with a surrogate model efficiently
• We successfully implemented a multinomial neural network to predict the proportion of occupied site-types and GCN histograms of synthetic spectra
• We demonstrated the applicability of this model to experimental data
• We iteratively used data science tools and philosophies with expert knowledge to identify areas of our combined \{target, feature\} space that needed more data and to generalize our model to high coverage systems with varying convoluting functions
Future Work on Blue Waters

Transition State Theory$^{[3,4]}$

\[
r_i = k_i \prod_j C_j
\]

\[
r = \kappa \frac{k_B T}{h} \exp\left(-\frac{\Delta G_i^\ddagger}{k_B T}\right) \prod_j C_j
\]

Transition State Theory and the Potential Energy Surface

- Reactants
- Saddle Point
- Transition State
- Products
- ΔH‡: Enthalpy
- local minima
Transition State Theory and the Potential Energy Surface

\[ G = H - TS \] \[ S = \text{function(}vibrations\text{)} \]

H: Enthalpy
G: Gibbs Energy
S: Entropy

Transition State Theory Computational Complexity

\[ O(N^3) \]

\[ O(N^4) \]

Vibrational Calculations
Outline of Future Work

Issues with the current technique for addressing the materials gap

1. Need more data
2. Frequency calculations are very slow!

- The electronic density distribution completely specifies the energy of a chemical system’s state and can be calculated using density functional theory (DFT) based on the Kohn Sham equation\cite{1}
- Frequencies at equilibrium can be computed directly from equilibrium (ground state) electron density\cite{2}

Combining geometric and electronic density information we should be able to generate a chemical representation that facilitates extrapolation.

1. Need more data – automatic structure generation for generative adversarial networks
2. Frequency calculations are very slow! – deep neural networks trained on electron density

Funding from DARPA and RAPID and funding from the Blue Waters Graduate Fellowship for the next phase of my work

Professor Dionisios G. Vlachos for advisement

The Vlachos group