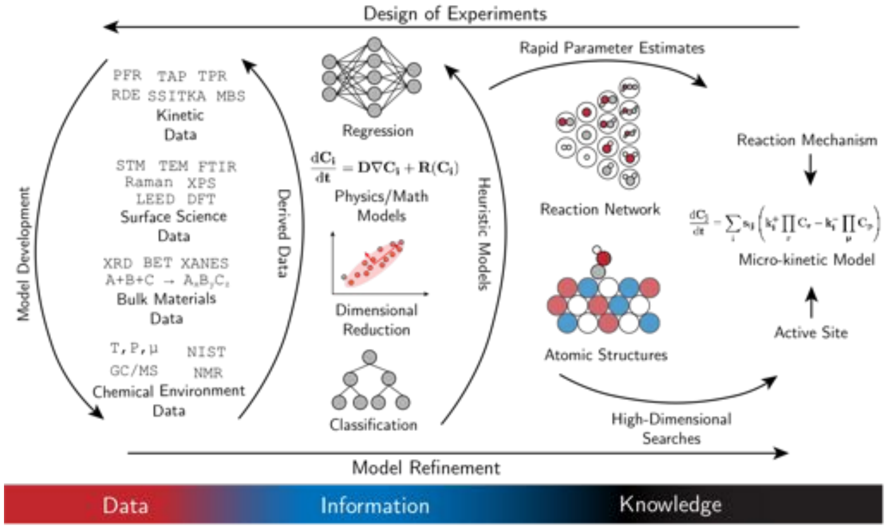
Catalysis Informatics: Utilizing machine learning and data science to extract knowledge from catalytic data



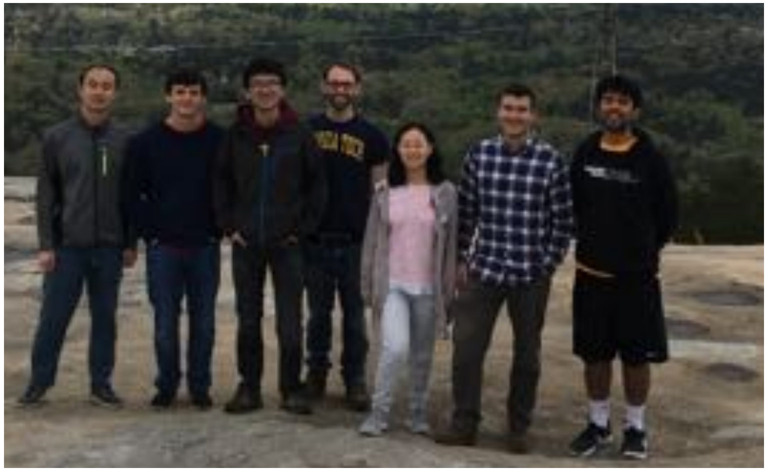
A.J. Medford
Assistant Professor
Dept. of Chemical & Biomolecular Engineering
Georgia Institute of Technology

06.07.18

Machine Learning in Science and Engineering Conference
Carnegie-Mellon University

Acknowledgements

- Georgia Tech
  - Adam Yonge
  - Sean Najmi
  - Ben Comer
  - Chaoyi Chang
  - Fuzhu Liu
- Idaho National Labs
  - Ross Kunz
  - Sarah Ewing
  - Rebecca Fushimi
  - Tammie Borders
- Funding
  - USDOE EERE Advanced Manufacturing Office Next Generation R&D Projects Contract no. DE-AC07- 05ID14517



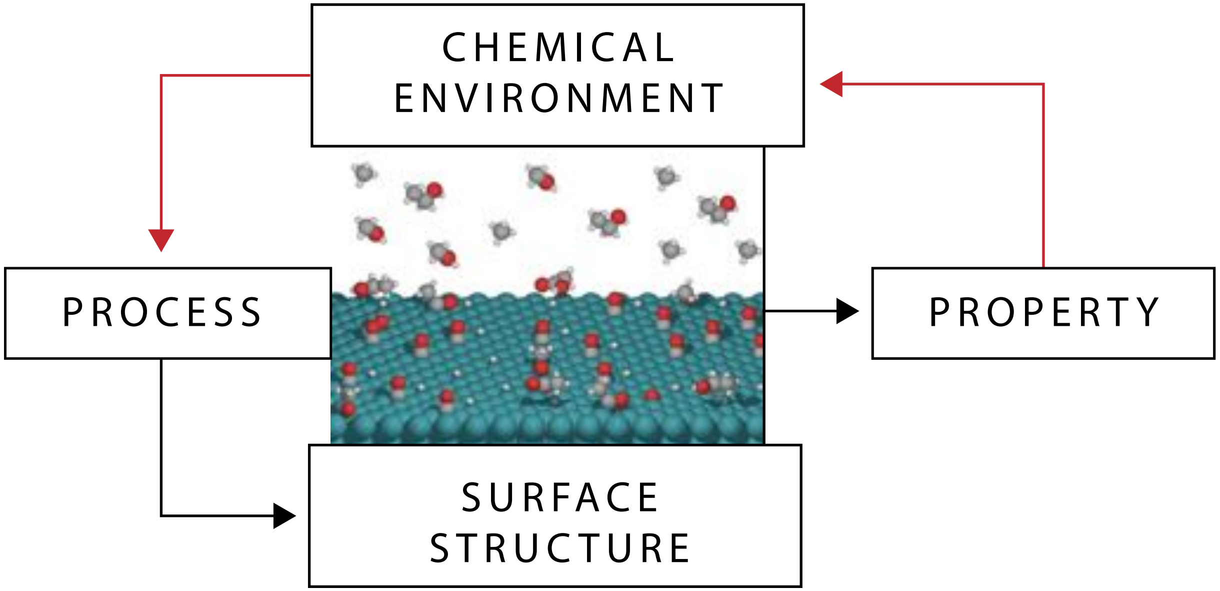
A.J. Medford, R. Kunz, S. Ewing, T. Borders, R. Fushimi. ACS Catalysis - accepted

(Heterogeneous) catalysis is a unique problem where machine-learning methods have significant potential

Outline

- Catalysis Data to Knowledge
    - Process-Structure/Environment-Property
    - Data-Information-Knowledge
    - Catalysis data types
- Catalysis Informatics
    - Macro-scale data
    - Micro-scale data
    - Bridging the gap
- Catalysis Knowledge
    - Reaction mechanism determination
    - Active site searches
    - Descriptor-based screening
- Future Directions

Catalysis informatics is a necessary sub-discipline

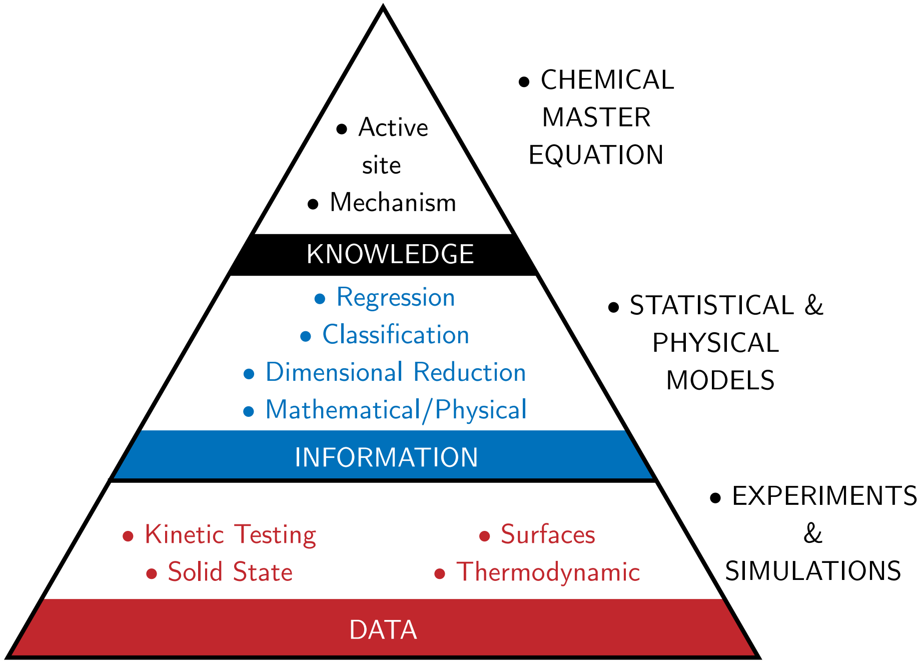


- Heterogeneous catalysis involves interaction of molecules + materials
  - Intersection of cheminformatics and materials informatics
- Process-Structure-Property paradigm fails for catalysis
  - Same material responds differently depending on environment
- Catalysis is a dynamic phenomenon
  - Catalysts alter their environment, which can induce structure changes

A.J. Medford, R. Kunz, S. Ewing, T. Borders, R. Fushimi. ACS Catalysis - accepted

The goal of catalysis informatics is to convert data to “knowledge”

$$\frac{dP(S_i)}{dt} = \sum_j \left(A_{ji} P(S_j) - A_{ij} P(S_i) \right)$$



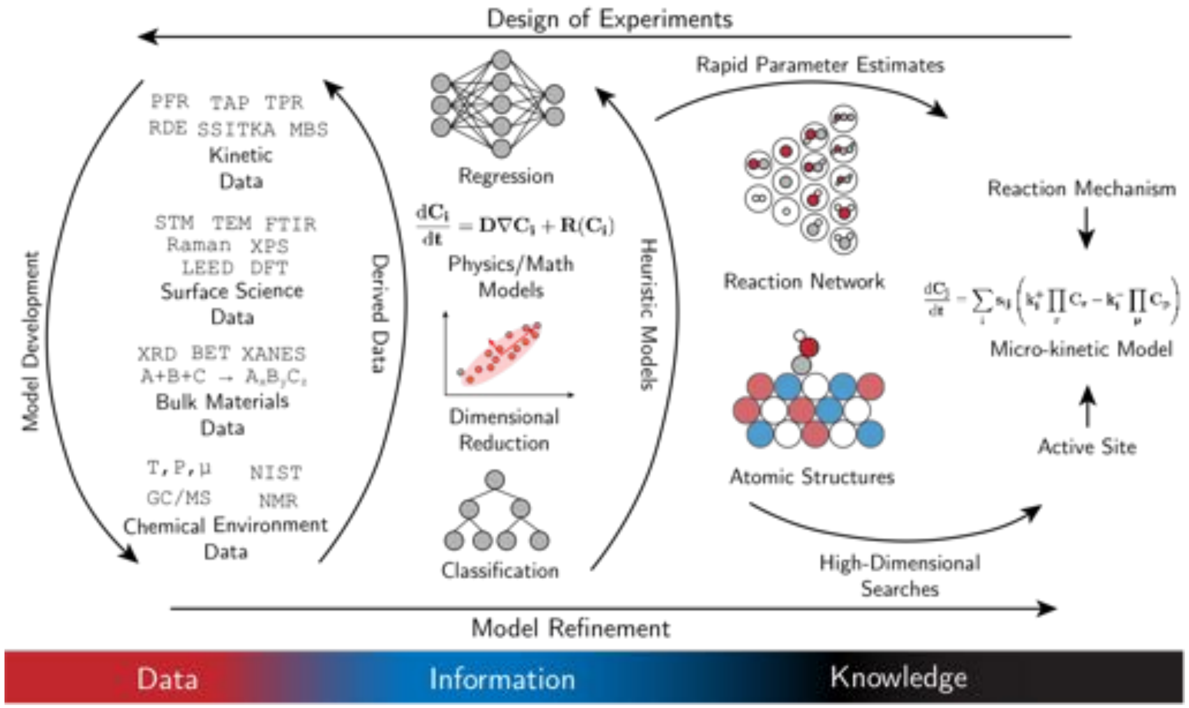
- Chemical master equation = knowledge

$$\frac{dC_j}{dt} = \underbrace{\sum_i s_{ij}}_{\text{mechanism}} \left(\overbrace{k_i^+}^{\text{active site}} \underbrace{\prod_r C_r}_{\text{mechanism}} - k_i^- \prod_p C_p \right)$$

  - Active site(s)
  - Reaction mechanism(s)
- Observable properties (rate, selectivity, reaction order, etc.) can be obtained from master eqn

A.J. Medford, R. Kunz, S. Ewing, T. Borders, R. Fushimi. ACS Catalysis - accepted

Data/information/knowledge is a dynamic continuum

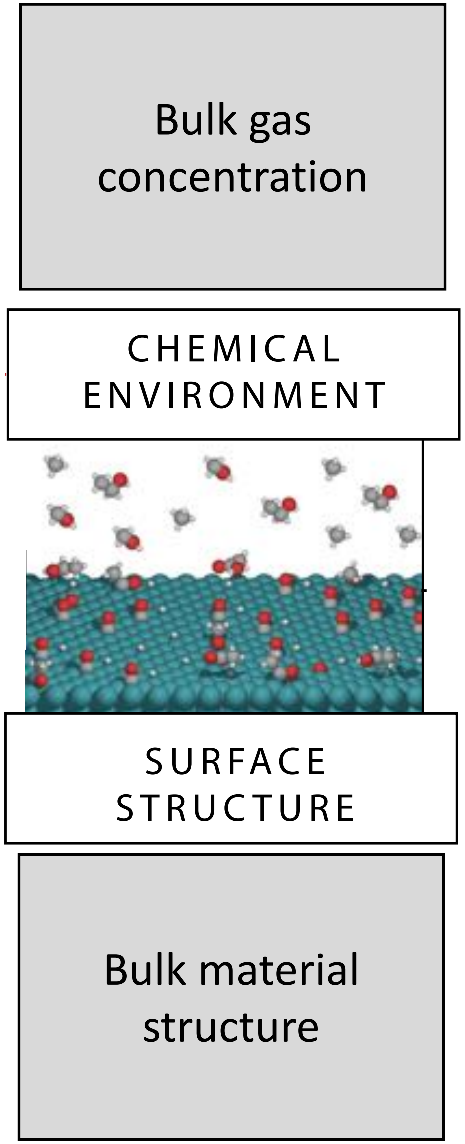


- Boundaries between data/info/knowledge are fuzzy
  - Derived data, TPR, etc.
- Knowledge extraction is a dynamic process
  - Design of experiments/calculations
  - Model refinement

A.J. Medford, R. Kunz, S. Ewing, T. Borders, R. Fushimi. ACS Catalysis - accepted

Catalysis includes many diverse sources of data

- Chemical environment data (cheminformatics)
  - Chemical potentials, molecular structures, etc.
  - Essentially cheminformatics
- Reaction kinetics data
  - Reaction rates, selectivity, stability, etc.
  - Involves dynamic concentrations over time
- Surface science data
  - Oxidation state, adsorption energies, etc.
  - Must be surface-sensitive
- Bulk materials data (materials informatics)
  - Material stability, composition, process, etc.
  - Essentially materials informatics



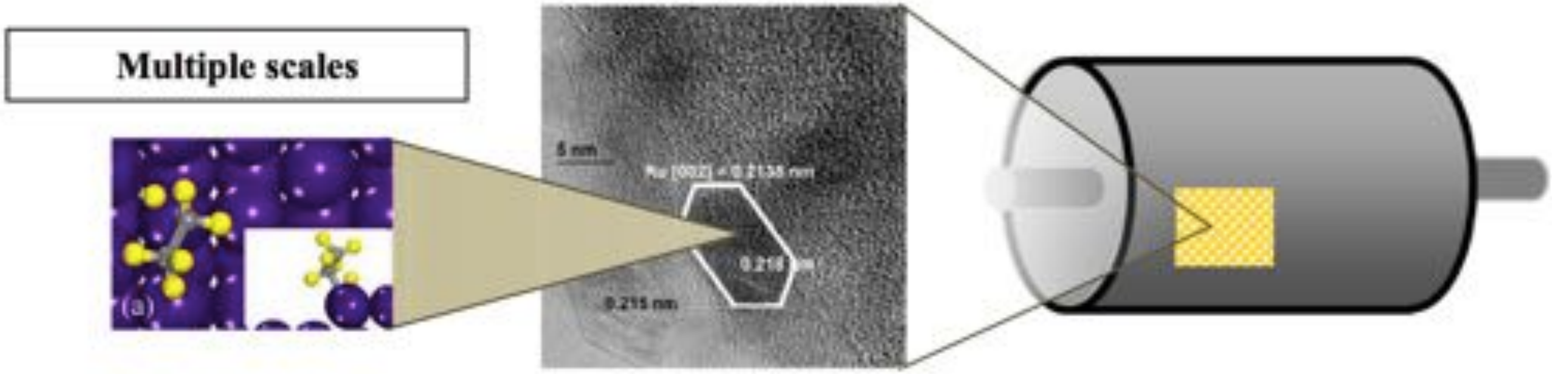
Bulk gas concentration

CHEMICAL ENVIRONMENT

SURFACE STRUCTURE

Bulk material structure

Catalysis data spans many scales and direct measurements are challenging



Micro-scale (surface science):

- DFT calculations
- XPS, UPS
- TPD, TPR
- Molecular beam
- SEM, TEM

Bridging the gap:

- In-situ/operando spectroscopy
- Transient kinetics
- Modulation-excitation spectroscopy

Macro-scale (reaction kinetics):

- Combinatorial testing
- Plug-flow/batch reactors
- 3-electrode cells
- Fluidized bed reactor

Catalysis informatics = multiscale modeling ++
+ diverse experimental data
+ statistical/ML models and heuristics

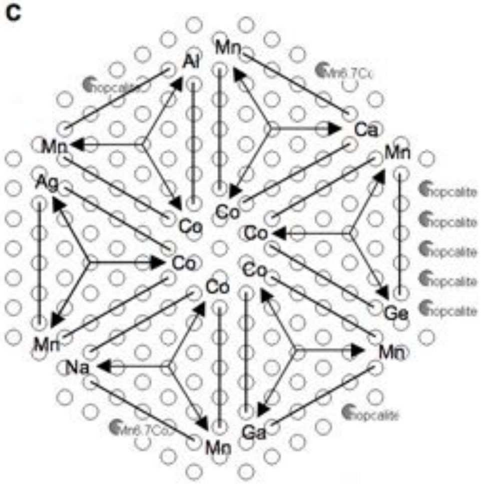
M. Salciccioli, M. Stamatakis, S. Caratzoulas, and D. G. Vlachos, Chemical Engineering Science, vol. 66, no. 19, pp. 4319–4355, Oct. 2011.

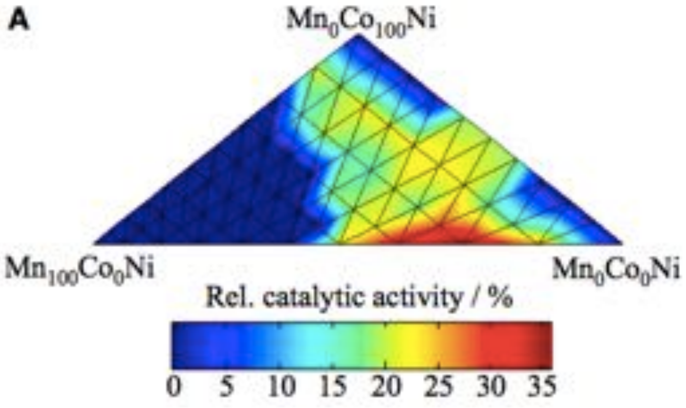
Outline

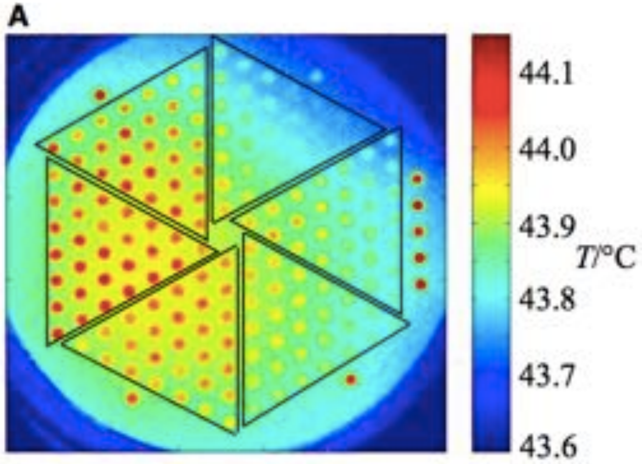
- Catalysis Data to Knowledge
  - Process-Structure/Environment-Property
  - Data-Information-Knowledge
  - Catalysis data types
- Catalysis Information
  - Macro-scale information
  - Micro-scale information
  - Bridging the gap
- Catalysis Knowledge
  - Reaction mechanism determination
  - Active site searches
  - Descriptor-based screening
- Future Directions

Combinatorial testing provides a rich source of catalysis testing data

- Test 100 – 1000 catalyst compositions at a time
- Rapid route to testing effects of composition



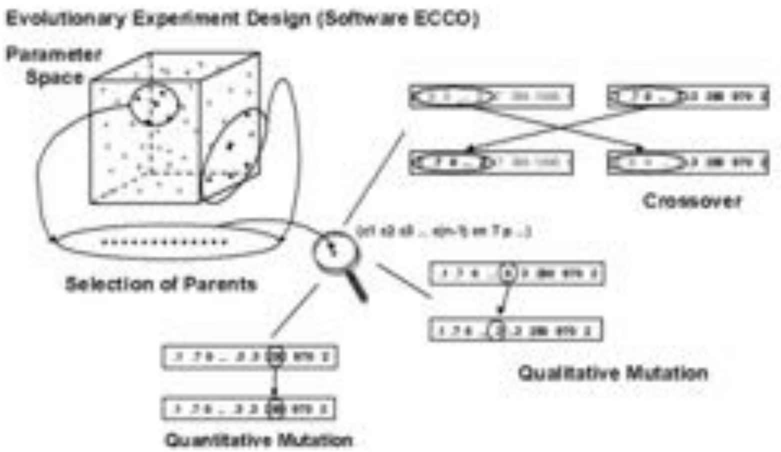


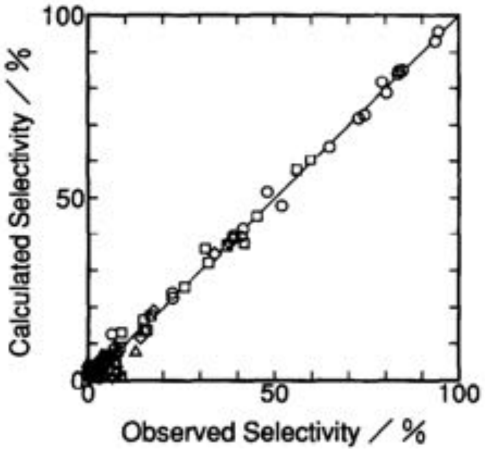


J. W. Saalfrank and W. F. Maier, Comptes Rendus Chimie, vol. 7, no. 5, pp. 483–494, May 2004.

Early examples of machine-learning in catalysis from combinatorial data

- Genetic algorithms accelerate searches
  - 5-component alloy → 150 million possibilities
  - 10,000 per day → 40 years of testing
- Supervised regression can seek trends in results
  - Neural networks commonly employed
  - Accuracy is reasonably good
  - Used as early as 1990’s

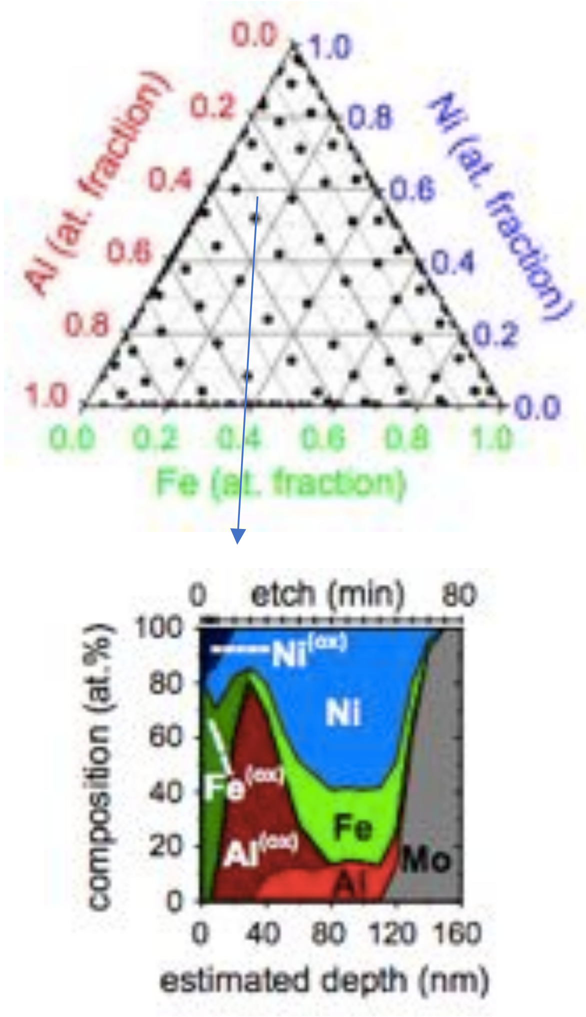




D. G. Duff, A. Ohrenberg, S. Voelkening, and M. Boll, Macromolecular Rapid Communications, vol. 25, no. 1, pp. 169–177, Jan. 2004.
S. Kite, T. Hattori, and Y. Murakami, Applied Catalysis A: General, vol. 114, no. 2, pp. L173–L178, Jul. 1994.

… but combinatorial testing is rarely used in catalyst design

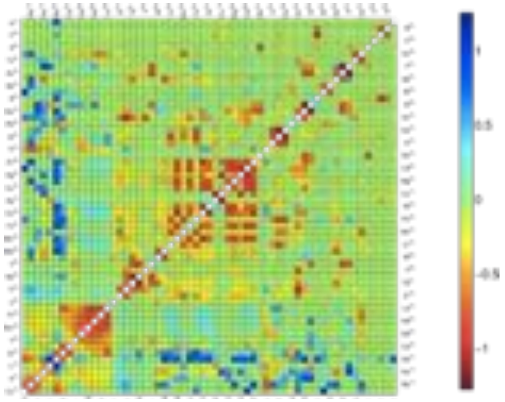
- False negatives are common
  - Surface concentration is not equal to bulk concentration
  - Surface defect structures and impurities can be important
  - Support effects may not be captured
- Transferability of models
  - Neural-net models may be biased by material class or experimental setup

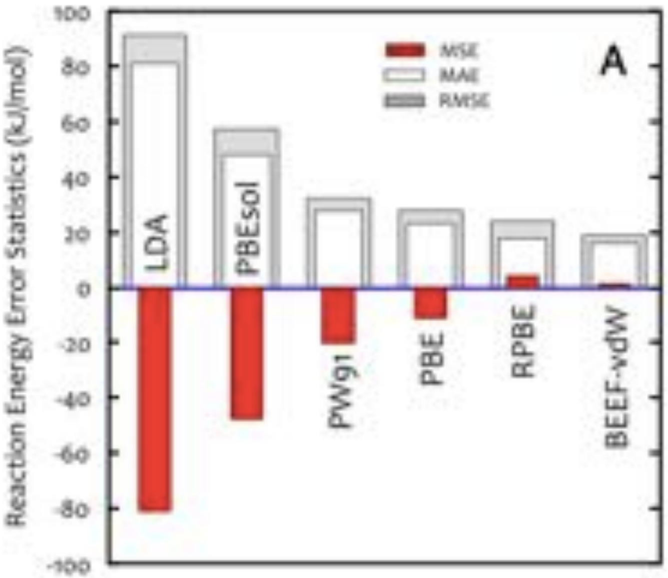


M. A. Payne, J. B. Miller, and A. J. Gellman, Corrosion Science, vol. 91, pp. 46–57, Feb. 2015.

Atomic-scale high-throughput data is primarily provided by DFT

- DFT data advantages:
  - Fast vs. experiments
  - Systematic
  - Well-defined
  - Reasonably accurate
- DFT data disadvantages:
  - Slow on absolute scale (hr-day)
  - Many possible options
  - Need atomic structure
  - Not perfectly accurate



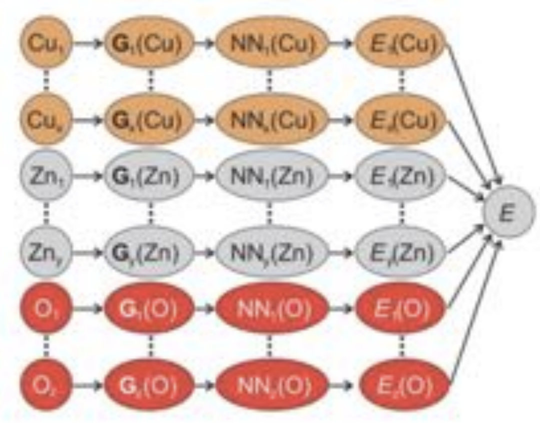


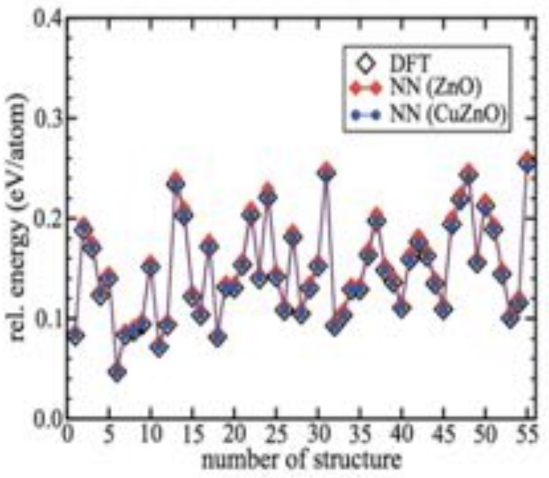
https://materialsproject.org/docs/structurepredictor

J. Wellendorff, T. L. Silbaugh, et. al. Surface Science, vol. 640, pp. 36–44, Oct. 2015.

Machine learning has been widely adopted by the DFT catalysis community

- Neural networks (and other ML models) can "learn" DFT results
  - Accuracy within <0.05 eV
  - Flexible for new and arbitrary systems (reacting interfaces!)
  - Orders of magnitude faster than DFT
- Some disadvantages:
  - Requires lots of training data (~10k)
  - Fails outside of training region
  - Not good for many-element (>~3) systems

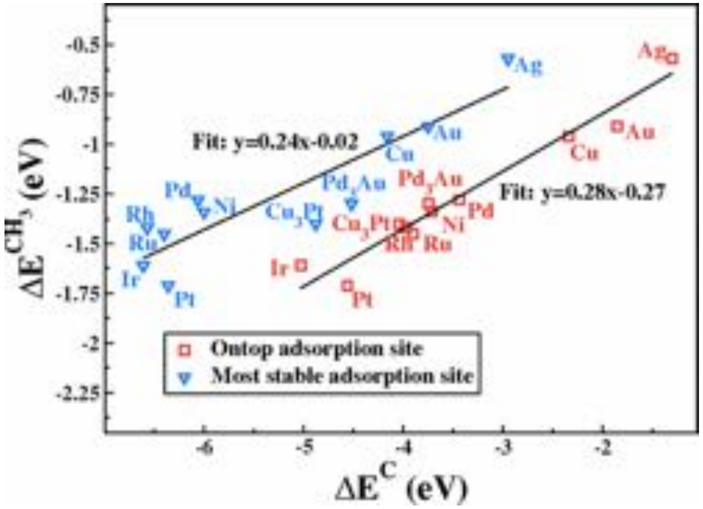


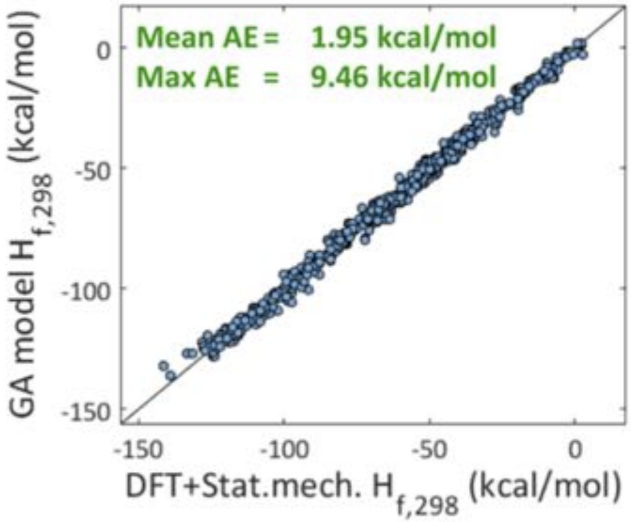


N. Artrith, B. Hiller, and J. Behler, physica status solidi, vol. 250, no. 6, pp. 1191–1203, Nov. 2012.
A. Khorshidi and A. A. Peterson, Computer Physics Communications, vol. 207, pp. 310–324, Oct. 2016.

Simpler correlation models are also widely used to rapidly predict energies

- Many model types:
  - Linear scaling relations
    - Predict binding across materials
  - Coordination number scaling
    - Predict binding across surface structures
  - Group additivity
    - Predict binding across adsorbate types
- Advantages:
  - Much faster
  - Less training data
  - More physical insight
- Disadvantages:
  - Less flexible/transferrable
  - Less accurate
  - More knowledge needed

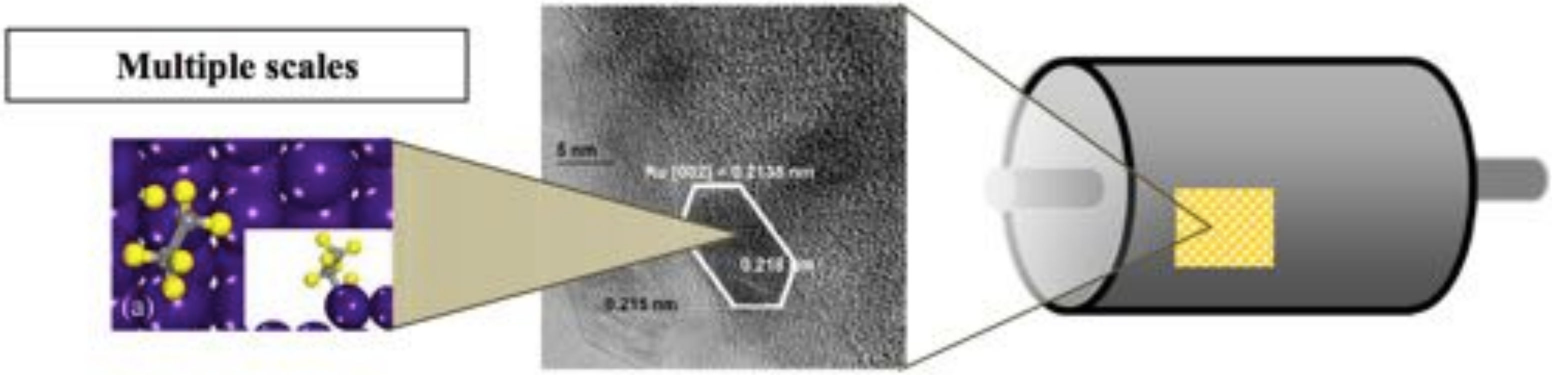




F. Abild-Pedersen et. al. Physical Review Letters, vol. 99, no. 1, Jul. 2007.

G. H. Gu and D. G. Vlachos, The Journal of Physical Chemistry C, vol. 120, no. 34, pp. 19234–19241, Aug. 2016.

There are many opportunities for machine learning in bridging the scale gap



Density functional theory

- Machine-learning forcefields
- Energy estimations

Catalyst testing

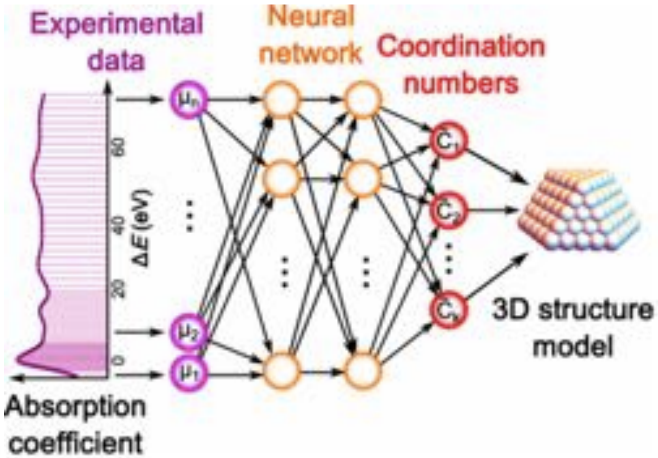
- Combinatorial approaches
- Genetic algorithms

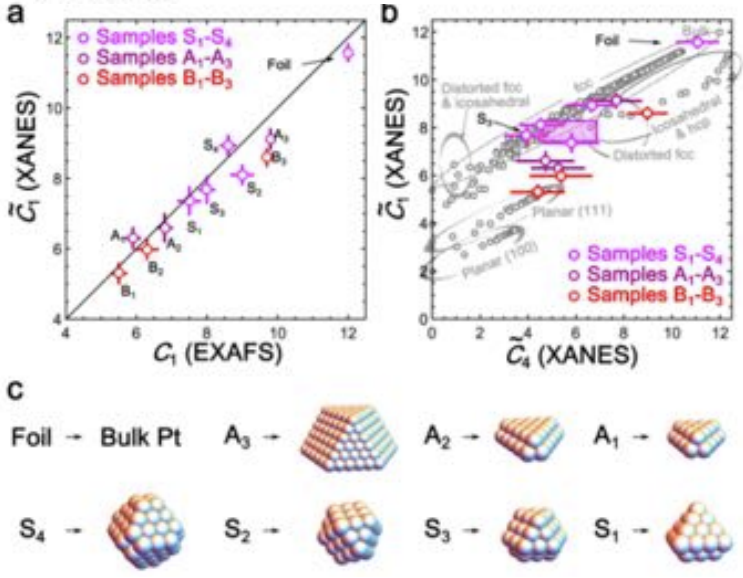
- In-situ/operando spectroscopy
- Transient kinetics
- Modulation-excitation spectroscopy

} Opportunities for ML

M. Salciccioli, M. Stamatakis, S. Caratzoulas, and D. G. Vlachos, Chemical Engineering Science, vol. 66, no. 19, pp. 4319–4355, Oct. 2011.

Neural network analysis of in-situ XAFS data can determine particle shape

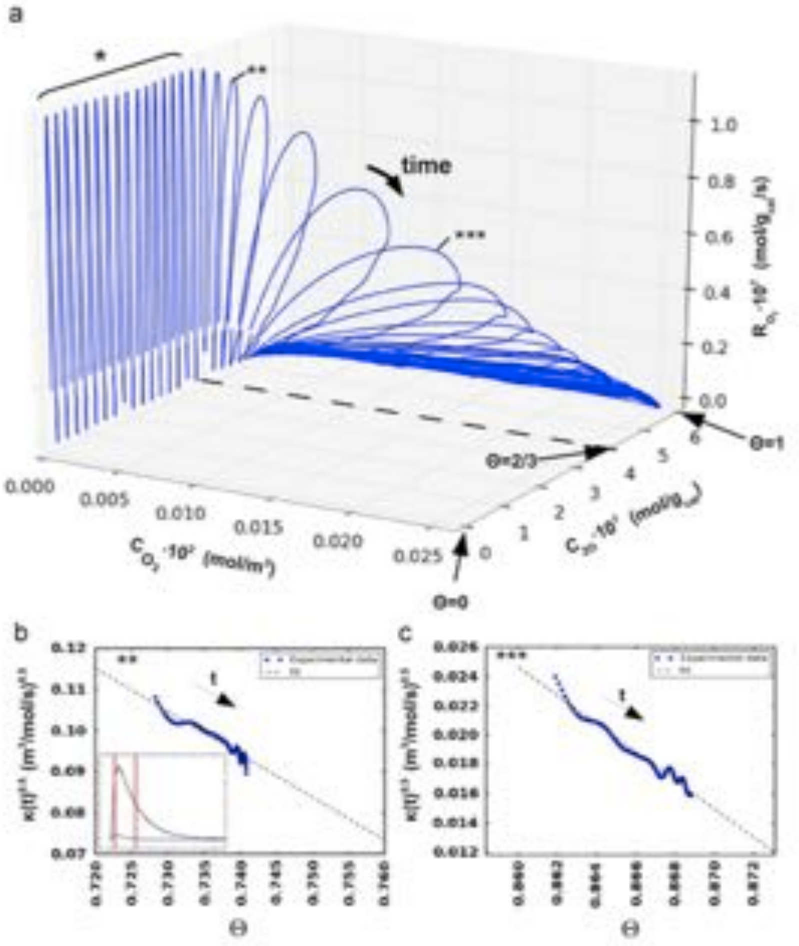




- Neural net trained to identify coordination number based on in-situ XAFS spectra
- Coordination number of different particle types predicted from atomic models
- Comparison gives insight into particle shape

J. Timoshenko, **D. Lu**, Y. Lin, and A. I. Frenkel, The Journal of Physical Chemistry Letters, vol. 8, no. 20, pp. 5091–5098, Oct. 2017.

TAP pulses are a high-throughput measurement of rate/concentration

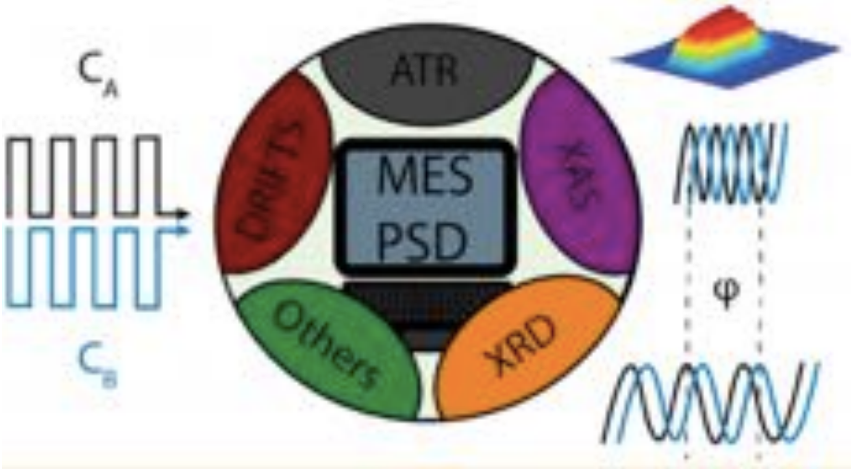


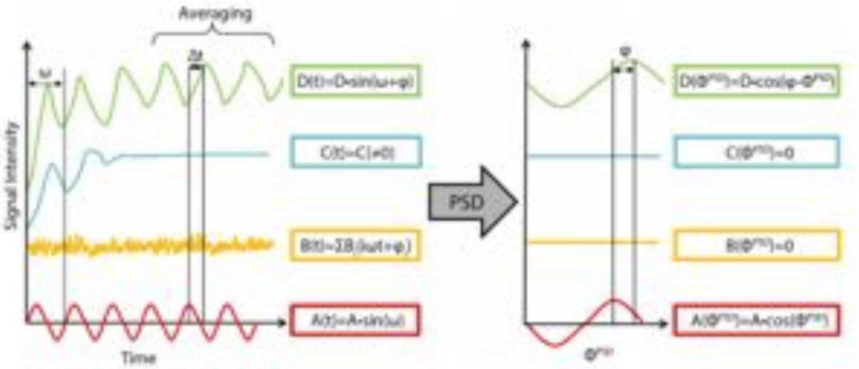
- Temporal analysis of product (TAP) is a surface science technique that operates on real catalysts
  - Similar to TPR
- Repeated TAP pulses sample different regions of rate/concentration space
- Complex data analysis needed to extract knowledge about mechanism

E. A. Redekop, G. S. Yablonsky, D. Constales, P. A. Ramachandran, C. Pherigo, and J. T. Gleaves, Chemical Engineering Science, 66, 24, 6441–6452, 2011.

Modulation-excitation spectroscopy can provide rich insight after analysis

- Modulation-excitation combines in-situ experiments and transients
- Works with many spectroscopic techniques
- Provides insight into spectator/active surface species
- Complex data analysis required



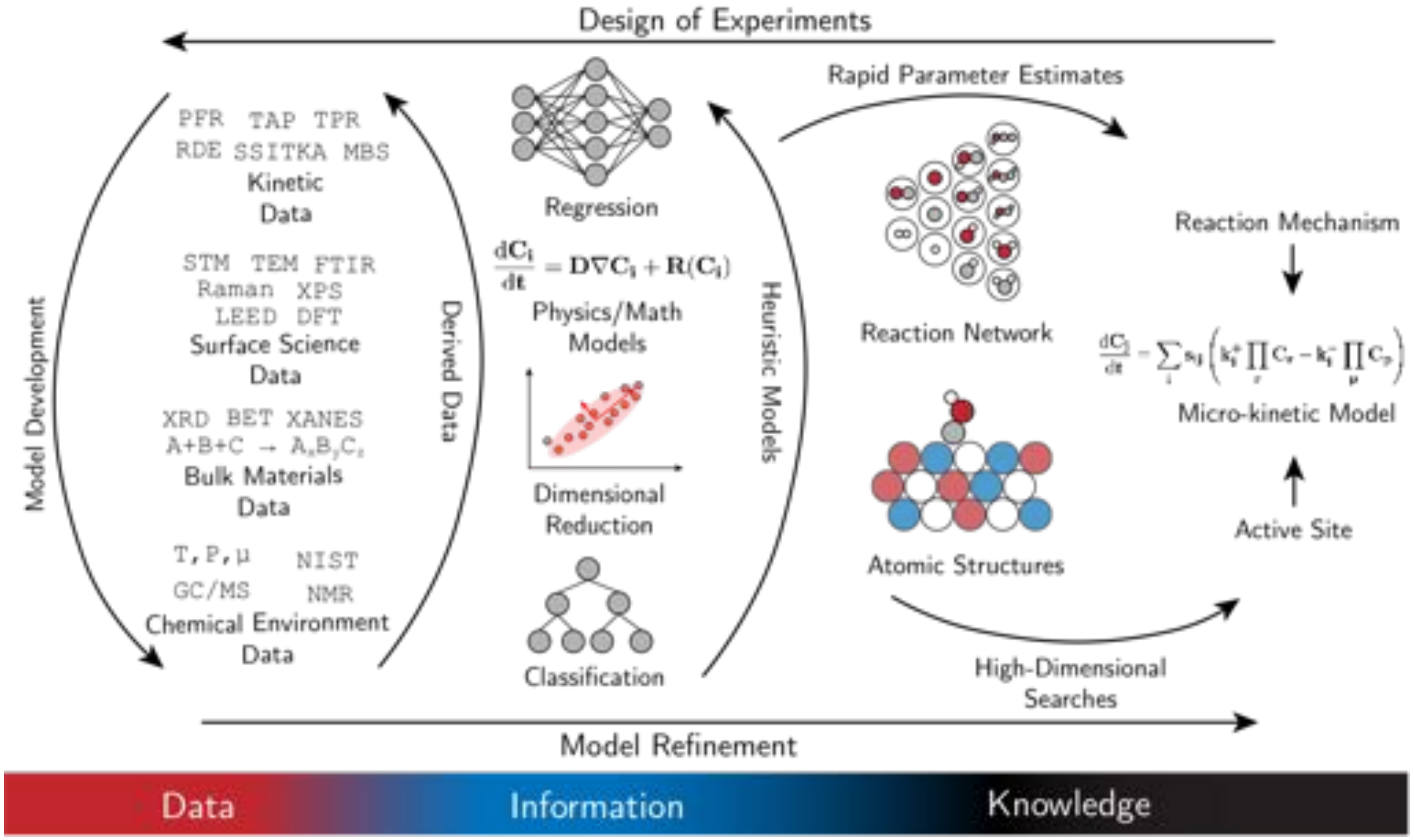


P. Müller and I. Hermans, Industrial & Engineering Chemistry Research, vol. 56, no. 5, pp. 1123–1136, Jan. 2017.

Outline

- Catalysis Data to Knowledge
  - Process-Structure/Environment-Property
  - Data-Information-Knowledge
  - Catalysis data types
- Catalysis Informatics
  - Macro-scale data
  - Micro-scale data
  - Bridging the gap
- Catalysis Knowledge
  - Reaction mechanism determination
  - Active site searches
  - Descriptor-based screening
- Future Directions

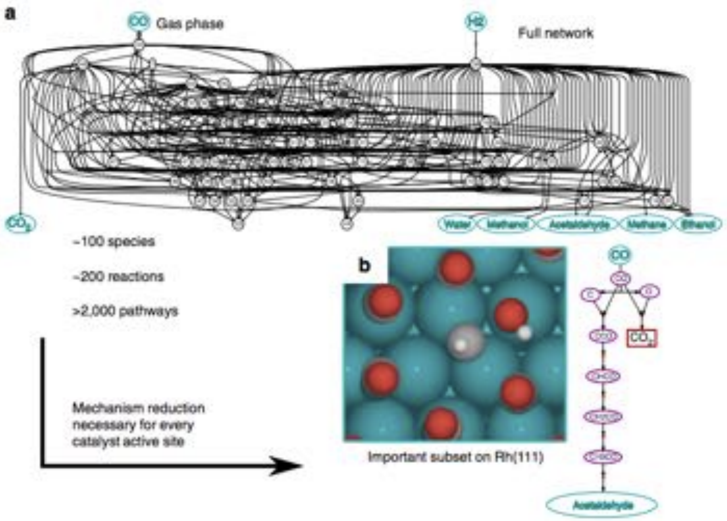
Constructing the master equation requires the active site(s) and reaction mechanism(s)

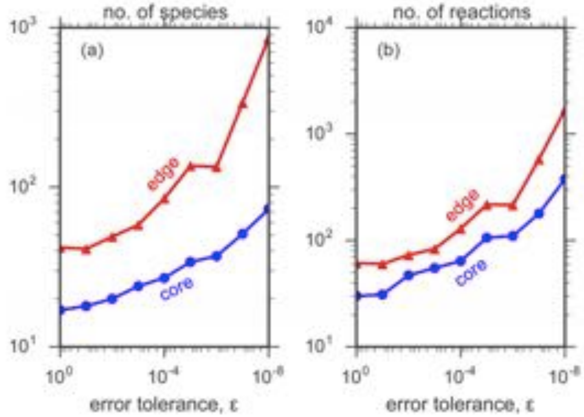


- Finding the reaction mechanism(s) and active site(s) are high-dimensional searches
  - Machine-learning can accelerate these searches!

Reaction mechanism determination needs rapid parameter estimates

- Reaction network size increases combinatorically with molecule size
- Two strategies:
  - Global – enumerate everything and reduce
  - Local – start from reactants and include most important
- Both accelerated by “ML”
  - Global – Iterative search + UQ
  - Local – Group additivity

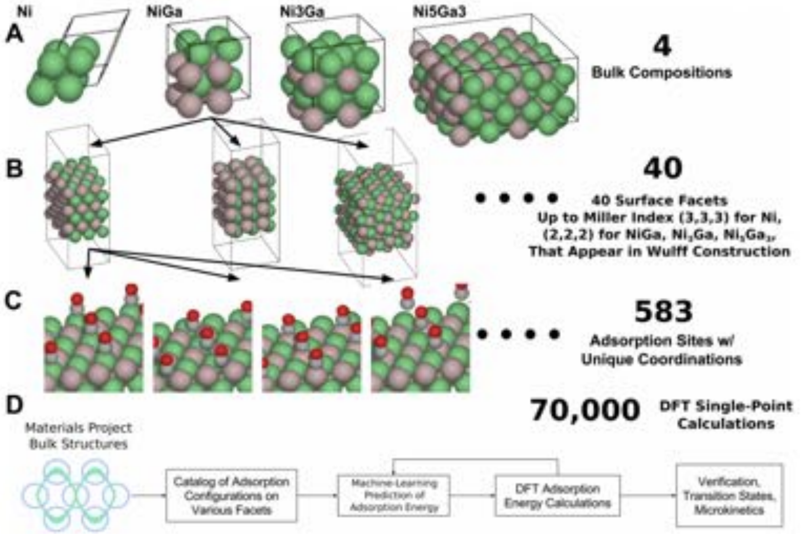


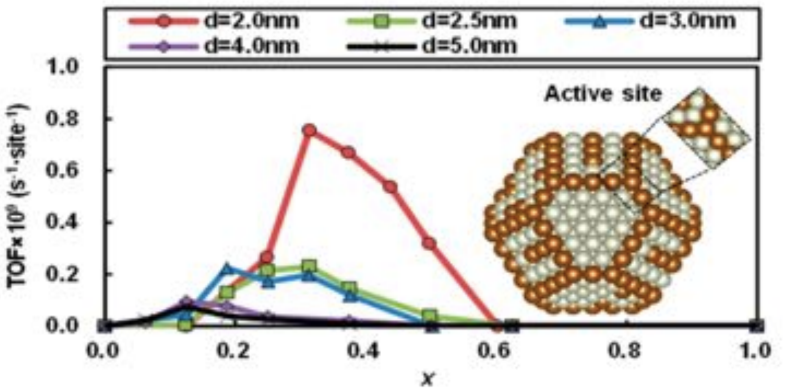


Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, Nature Communications, vol. 8, p. 14621, Mar. 2017.
C. F. Goldsmith and R. H. West. The Journal of Physical Chemistry C, vol. 121, no. 18, pp. 9970–9981, May 2017.

Machine-learning models enable exhaustive active-site searches

- Real catalysts have diverse range of active sites
  - Alloys -> varied compositions
  - Nanoparticles -> varied surface structures
  - Oxides/compounds -> vacancies/defects
- Coupling rapid energy estimates with nanoparticle models enables more robust active-site identification for realistic models



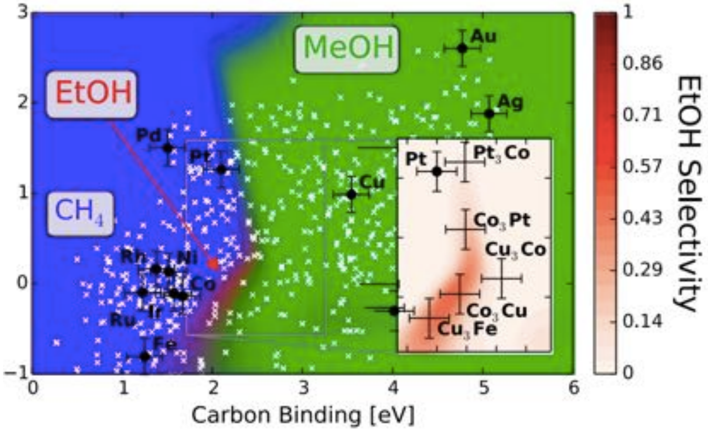


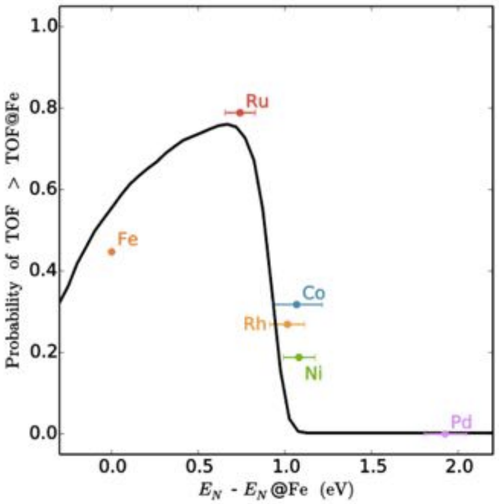
Z. W. Ulissi, M. T. Tang, et. al. ACS Catalysis, vol. 7, no. 10, pp. 6600–6608, Aug. 2017.

R. Jinnouchi and R. Asahi, The Journal of Physical Chemistry Letters, vol. 8, no. 17, pp. 4279–4283, Aug. 2017.

Volcano plots can integrate data across materials to generate predictions

- Knowledge of active site and mechanism:
  - **explains** catalytic activity/selectivity for a given material
  - enables **prediction** of activity/selectivity for new materials
- Volcano plots are informatics models
  - Integrate lots of data with regression model and kinetic models
  - Dimensional reduction for catalyst design
- Other opportunities for data science/ML:
  - Uncertainty quantification
  - Improved regression/descriptor selection





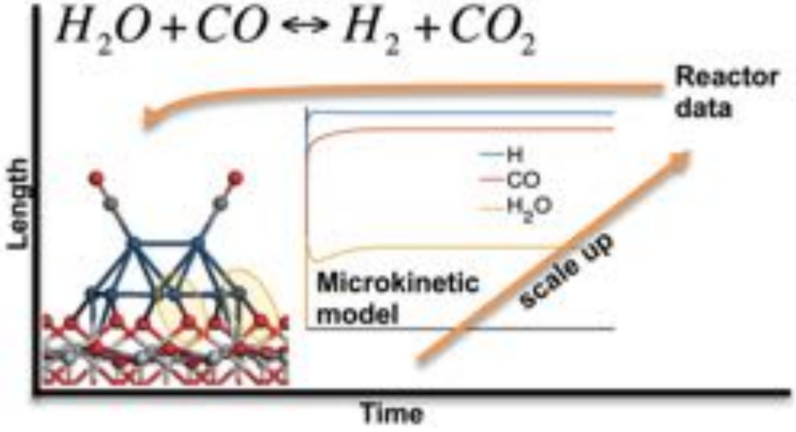
A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, et. al. Journal of Catalysis, vol. 328, pp. 36–42, Aug. 2015.
A. J. Medford, J. Wellendorff, A. Vojvodic, F. Studt, F. Abild-Pedersen, et. al. Science, vol. 345, no. 6193, pp. 197–200, Jul. 2014.

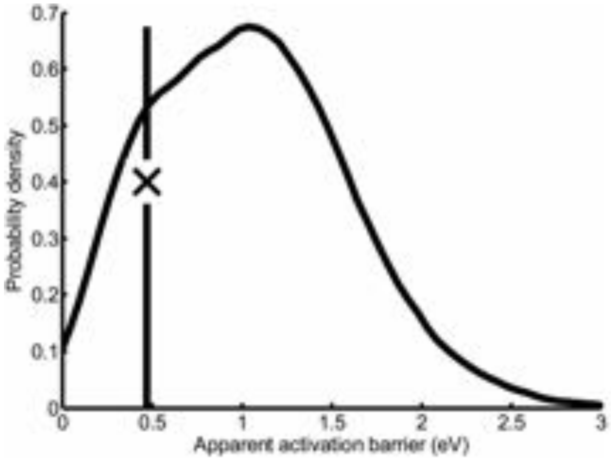
Outline

- Catalysis Data to Knowledge
  - Process-Structure/Environment-Property
  - Data-Information-Knowledge
  - Catalysis data types
- Catalysis Informatics
  - Macro-scale data
  - Micro-scale data
  - Bridging the gap
- Catalysis Knowledge
  - Reaction mechanism determination
  - Active site searches
  - Descriptor-based screening
- Future Directions

Machine-learning has primarily generated knowledge from computational data

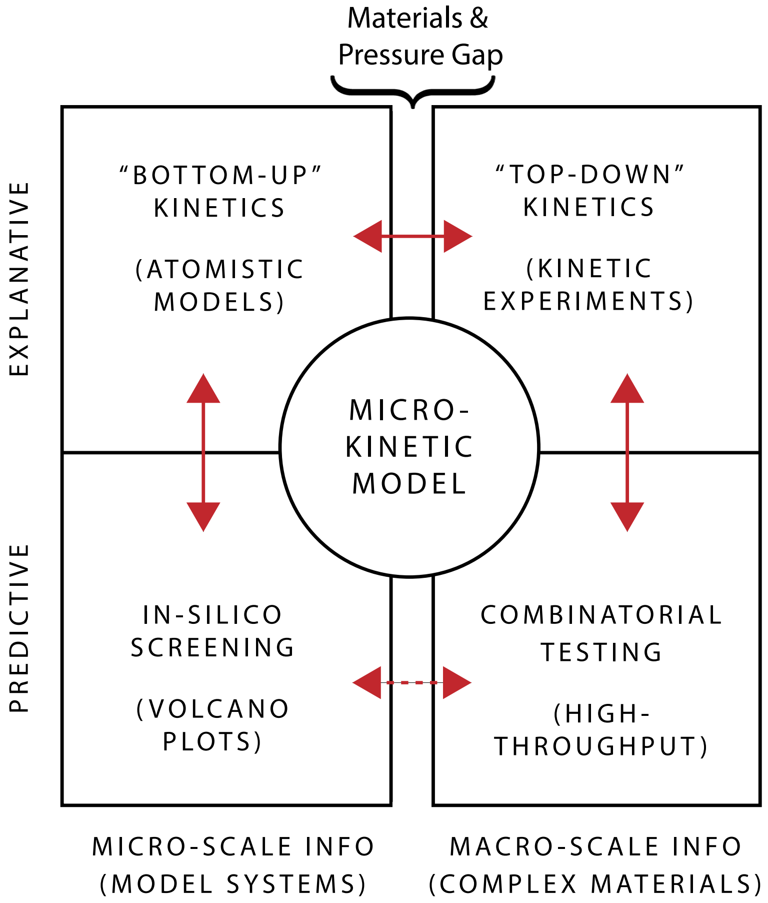
- Natural starting point:
  - Systematic, widely available, similar skillsets needed
  - Illustration how ML can accelerate knowledge generation
- Next step: Integrate experiments!
  - Experimental data complements theory
  - Fusing both can lead to robust new knowledge





E. Walker, S. C. Ammal, G. A. Terejanu, and A. Heyden, The Journal of Physical Chemistry C, vol. 120, no. 19, pp. 10328–10339, May 2016.
E. A. Walker, D. Mitchell, G. A. Terejanu, and A. Heyden, ACS Catalysis, vol. 8, no. 5, pp. 3990–3998, Mar. 2018.

Catalysis "knowledge engines" seek to automate and integrate knowledge creation

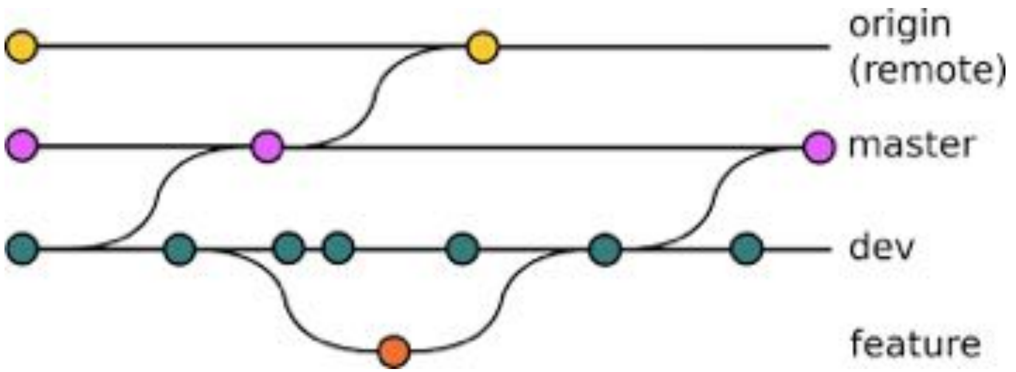


- Catalysis "knowledge engines" were first proposed by Caruthers et. al. in 2004
  - Combine high-throughput experimentation with model fitting
- Recent advances improve feasibility
  - computational catalysis
  - machine learning
  - open-source development
  - data infrastructure
- Integrate information from disparate sources to obtain generalizable knowledge
  - Need more quantitative connections and uncertainty quantification!

J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, et. al., Journal of Catalysis, vol. 216, no. 1–2, pp. 98–109, May 2003.
A.J. Medford, R. Kunz, S. Ewing, T. Borders, R. Fushimi. ACS Catalysis - submitted

Software implementations can enable development of “knowledge engines”

- Previously published approaches are not openly available
- Open-source tools and community effort can accelerate the field
  - Standard interfaces
  - Improved reproducibility
- Modern tools make this easy!



J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, et. al., Journal of Catalysis, vol. 216, no. 1–2, pp. 98–109, May 2003.

Catalysis is a unique problem where machine-learning methods have significant potential