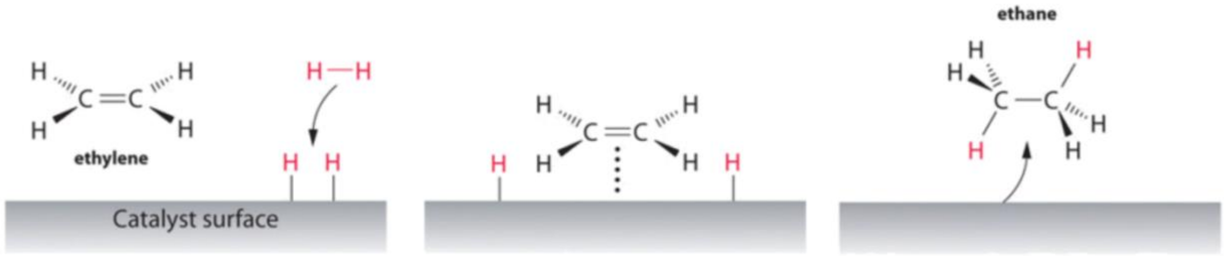
The use of Machine Learning to Predict Properties of Heterogeneous Catalysts

From fertilizers, plastics, fuels, and pharmaceutical drugs, modern society heavily relies on chemicals to sustain our lives. In particular, the mass production of commodity small molecules like ammonia, ethylene, propylene, methanol are extremely important and are mass produced industrially. However, the production of these commodity chemicals has led to large amounts of green-house gas emission. For instance, the synthesis of ammonia (NH_3), a precursor for fertilizers, is currently industrially synthesized through the Harbor-Bosch (H-B) process from atmospheric nitrogen (N_2). The development of the H-B process has enabled mass production of fertilizers, supporting the rise of modern, intensive agriculture and thus the rise in global populations. However, the H-B process, while being extremely efficient, requires the chemical reaction to take place at extremely high pressures, leading to huge emissions of CO_2 into the atmosphere. In this reaction, atmospheric nitrogen (N_2) and hydrogen (H_2) is inserted into a reactor with an iron-based catalyst (Fe_2O_3) (Equation 1). In theory, the iron-catalyst helps speed up the reaction of nitrogen and hydrogen and helps the reaction proceed under milder reaction conditions. However, despite the use of the iron-catalyst, the reaction can only take place at harsh conditions above 400 °C and 150 bar to go over the energy barrier necessary to break and form new bonds.

$$N_2 + 3H_2 \rightarrow 2NH_3 \tag{1}$$

To maintain the temperature and pressure at these extreme conditions, a lot of energy input in required, thus leading to a large release of carbon-dioxide into the atmosphere. Therefore, an analogous chemical reaction that can have comparable efficiencies at lower pressures and a chemical route that doesn't lead to emission of CO_2 is therefore of high interest. One way to achieve this is by using catalysts that can lower the energy barrier of the reaction, allowing the desired to be produced, even with a low energy input. While an iron catalyst is used in the H-B process, an even more efficient catalyst may allow for this reaction to proceed at even lower temperature and pressures. Therefore, the development of cheap, durable, and efficient catalysts is an important area of study in order to allow for a sustainable future.

The development of catalysts for industrial scale usages have currently been focused on heterogeneous catalysts, often mainly composed of solid metals. When a chemical of interest such as H_2 and N_2 gas comes close to these metal surfaces, through a process called adsorption, the bonds of the molecules are weekend and will be bound to the metal surface, speeding up the chemical reaction (Scheme 1). The development of such catalysts has traditionally relied on chemical intuition and experimental data. In chemistry, mechanistic insight on how the reaction takes place, is a huge part of developing chemical technologies. For instance, previous reports have shown that alkali metals are particularly effective at activating atmospheric nitrogen and rigorous experiments have been conducted to exhaust possible combinations of chemical to optimize the metal substrates and reaction conditions.[1] Similar efforts are happening through the synthesis of various solid catalysts using a combination of metals and organic substrates.[2] Additionally, not all catalysts are activated with the addition of heat, known as thermal-catalysis. Light-driven catalysis (photocatalysis) and electricity-driven catalysis (electrocatalysis) have gained popularity in recent years, diversifying strategies using heterogeneous catalysis.[3,4] Recently, many interests have been raised on the use of data-driven strategies for catalyst development, revolutionizing the material discovery paradigm.[5] However, the use of artificial intelligence (AI) and machine learning (ML) has been a novel approach in chemistry, and further improvements are possible in applying various areas of chemistry. Moreover, these approaches are often avoided by chemists that have limited background in computer-science, slowing down the application of AI/ML into chemistry and chemical catalysis.



In this study, the *Catalyst Property Database* (CPD) will be used. CPD is a database that contains adsorption energies of various small molecules calculated from quantum-mechanical calculations, mainly density functional theory (DFT) calculations . The adsorption energy is a good indicator for the performance of the catalyst as a lower adsorption energy allows for the reaction to proceed smoothly, meaning that the catalyst is more efficient. The database contains several attributes such as the metal used as the catalyst and various parameters that describe the structural characteristics of the catalyst (Table 1).

Table 1. The potential labels and corresponding attributes that can be used in the dataset.

Potential labels	Potential Attributes
Adsorption energy, adsorption site	Cell Symmetry, facet, first layer, second layer, fixed substrate, formula, nanoparticle size, potential, primary class, secondary class, point-group.

A great improvement in the world of heterogeneous catalysis will be our ability to accurately predict the adsorption energies of molecules and back-track properties that lead to the desired results. This will allow us to use this data and sample potential untested catalysts that may be of interest to test next, accelerating the material discovery process. Therefore, creating a regression-based ML model that can predict the adsorption energies will allow us to screen for untested materials to then be tested experimentally. Moreover, a categorical prediction ML model that can predict the adsorption sites will allow us to predict certain mechanisms.

This will be accomplished by using a combination of basic statistical techniques and multiple machine learning algorithms. The steps are outlined below.

1. Gain clear understanding of the chemical meaning behind each attribute.
2. Create a smaller test data to use for implementation of the ML algorithms.
3. Use basic statistical methods such as linear regression to gain a basic understanding of the correlation between the label and attributes as well as correlations within attributes.
4. For instances with missing attributes, use appropriate imputation methods. Then, clean up several attribute values so that the program can interpret them smoothly.
5. For categorical labels (adsorption site), use very basic ML algorithms such as k-nearest neighbors and decision trees to see how well the model performs. If program doesn't perform as intended, try with subsets of data that focus on a particular type of attribute (ex. Try with data that only uses Fe metal). If not successful, try replacing/re-mapping the attribute data (ex. Fe can be remapped to as atomic number, atomic mass, atomic radii, electron configuration, etc.).
6. Take similar strategies as step 4 to predict numerical labels using regression algorithms such as support vector machines (SVM) with varying kernels, and various types of neural networks (NN).

Currently, I have gained a clear understanding of the chemical meaning behind each attribute and have mapped out potential ways in which I can express each attribute as some are categorical data that can be turned into numerical data. With the final deadline for this project being May 16th, I will be looking for a completion of the project by May 10th, to allow for revisions to be made to the program. A rough schedule in which I will complete the tasks are outline below (Table 2).

Table 2. Schedule to complete the project.

Task	Date
Create a small data set for implementation	April 12th
Download and clean up data	April 19th
Apply basic statistical methods	April 26th
Apply ML algorithms (numerical)	May 3rd
Apply ML algorithms (categorical)	May 10th
Prepare final submission	May 16th

References

1. Chang, F. *et al.* Alkali and Alkaline Earth Hydrides-Driven N_2 Activation and Transformation over Mn Nitride Catalyst. *J Am Chem Soc* **140**, 14799–14806 (2018).

2. Arroyo-Caire, J., Diaz-Perez, M. A., Lara-Angulo, M. A. & Serrano-Ruiz, J. C. A Conceptual Approach for the Design of New Catalysts for Ammonia Synthesis: A Metal—Support Interactions Review. *Nanomaterials* **13**, 2914 (2023).

3. Wang, J., Marchetti, B., Zhou, X.-D. & Wei, S. Heterogeneous Electrocatalysts for Metal–CO_2 Batteries and CO_2 Electrolysis. *ACS Energy Lett* **8**, 1818–1838 (2023).

4. Lopat'eva, E. R. *et al.* Heterogeneous Photocatalysis as a Potent Tool for Organic Synthesis: Cross-Dehydrogenative C–C Coupling of N-Heterocycles with Ethers Employing TiO2/N-Hydroxyphthalimide System under Visible Light. *Molecules* **28**, 934 (2023).

5. Foppa, L. *et al.* Data-Centric Heterogeneous Catalysis: Identifying Rules and Materials Genes of Alkane Selective Oxidation. *J Am Chem Soc* **145**, 3427–3442 (2023).