MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS

Sergio Pablo García Carillo

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

UNIVERSITAT ROVIRA i VIRGILI

More is Different: Modern Computational Modeling for Heterogeneous Catalysis

Sergio Pablo García Carillo

DOCTORAL THESIS
2022

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Sergio Pablo-García Carrillo

More is Different: Modern Computational Modeling for Heterogeneous Catalysis

DOCTORAL THESIS

Supervised by
Prof. Núria López Alonso

Institute of Chemical Research of Catalonia (ICIQ)
and Rovira i Virgili University (URV)

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2022

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

hola
Institut Català d'Investigació Química
Av. Països Catalans, 16
43007 Tarragona. Spain

Prof. Núria López Alonso, group leader at the Institute of Chemical Research of Catalonia.

I STATE that the present study, entitled "**Modern Computational Modeling for Heterogeneous Catalysis**", presented by Sergio Pablo-García for the award of the degree of Doctor in Chemical Science and Technology, has been carried out under our joint supervision at the Institute of Chemical Research of Catalonia and that it fulfills all the requirements to be eligible for the International Doctorate Award.

Tarragona, May 5rd, 2022

Prof. Núria López Alonso

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Sponsors

The work presented in this Ph.D. thesis has been funded by the Institute of Chemical Research of Catalonia (ICIQ), member of the Barcelona Institute of Technology (BIST). The generous computer resources provided by the Barcelona Supercomputing Centre (MareNostrum) and the Spanish Supercomputing Network are also acknowledged.

UNIVERSITAT ROVIRA I VIRGILI

Barcelona Institute of
Science and Technology

Acknowledgements

I will start this section by saying that I am not a people's person, I enjoy quietness and solitude the most. However, and often to my regret, human beings are gregarious animals and need mutual support to achieve complex goals. Therefore, I think it is fair to say that I am merely the catalyst for this work, as it has been my environment and my connections with other humans that have made it possible. In the following pages I will proceed to list those who I believe have had the greatest influence on this project, starting with my non-academic circle:

- **My parents** for raising me as best as they could.
- **My brother** to whom I owe my passion for computers and science.
- **My sisters** for the love and support that they have given me over the years.
- **My nieces** for pizza Fridays and bass sessions.
- **J. M. Prieto** for all the support he has given me the last 20 years.
- **O. Sánchez** for the gaming sessions and long talks.
- **A. Gil** for Quake, Nicholas Cage, Whisky, Statistics and philosophy.
- **F. Silva** For the amazing cover.
- **C. Cabezas** for having been by my side since I told her that someday she would call me Doctor.

The academic circle:

- **Prof. J. Ribas** for encouraging me to start this journey.
- **Dr. Q. Li** for his teachings on Density Functional Theory and Automation.

- **Dr. R. García-Muelas** for his generous guidance and teachings, without forgetting the Dvorak distribution.
- **A. Sabadell-Rendón** for the mathematical discussions and all the projects that we developed together.
- **Dr. R. Vargas** for sharing with me his knowledge of machine learning and introducing me to the school of automatic differentiation.
- **Dr. J. Kjell** for their generous guidance and support during my stay in Toronto.
- **Prof. A. Aspuru-Guzik** for giving me the opportunity to visit his group and inspiring my work.
- **Prof. N. López** for inspiring me with her passion for science and helping me to become the scientist I am today.
- **All my collaborators** for making all our projects possible.

Special Thanks:

- ***Patogrupo*** for the *patojuego* evenings. Specifically:
  - **D. Garay** for Emacs.
  - **R. Pérez** for Python.
  - **A. Villar** for Miau.
- **Gnus** for the coffee breaks and the uncomfortable talks.

This list condenses in a little number of words the help and support you have given me. However, *little* do not mean meaningless, and I believe that you all know the weight these words carry for me.

Before closing this chapter, I would like to show my gratitude to the entire **GNU/Linux** community. They not only provided the complete set of tools that I needed to complete this work but also an excellent documentation.

Contents

Abbreviations

C_n	forall $n \in \mathbb{N}$ organic molecules containing n carbon atoms
E_a	Activation Energy
ΔE	Reaction Energy
ANN	Artificial Neural Networks
ASE	Atomic Simulation Environment
bcc	Body Centered Cubic
BEP	Brønsted-Evans-Polanyi
BMS	Bayesian Machine Scientist
CHE	Computational Hydrogen Electrode
CI-NEB	Climbing-Image Nudged Elastic Band
CML	Chemical Markup Language
CNN	Convolutional Neural Network
DFT	Density Functional Theory
$\mathrm{eCO_2R}$	Electroreduction of Carbon Dioxide
FAIR	Findable Accessible Interoperable Reusable
fcc	Face Centered Cubic
FE	Faradaic Efficiency
GGA	Generalized Gradient Approximation
GNN	Convolutional Graph Neural Network
GR	Gaussian Regressor
hcp	Hexagonal Close Packed
KMC	Kinetic Monte Carlo

LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Local-Density Approximation
LSR	Linear Scaling Relationships
MAE	Mean Absolute Error
MCMC	Markov-Chain Monte-Carlo
MK	Microkinetic
NEB	Nudged-Elastic Band
OD-Cu	Oxide Derived Copper
ODE	Ordinary Differential Equation
PAW	Projected-Augmented Wave
PBC	Periodic Boundary Conditions
PBE	Perdew-Burke-Enrzerhof
PCA	Principal Components Analysis
PES	Potential Energy Surface
PPR	Projection Pursuit Regression
PW91	Perdew-Wang 91
ReLU	Rectified Linear Unit
RF	Random Forest
SELFIES	Self-Referencing Embedded Strings
SL	Statistical Learning
SMILES	Simplified Molecular-Input Line-Entry System
SSE	Sum of Squared Errors
t-SNE	t-Distributed Stochastic Neighbor Embedding
TS	Transition State
US-PP	Vanderbilt Ultrasoft Pseudo-Potentials
VASP	Vienna Ab Initio Simulation Package
WF	Work Function

List of Publications

1. **Turning chemistry into information for heterogeneous catalysis.** *S. Pablo-García, M. Álvarez-Moreno, & N. López,* Int. J. Quantum Chem. **2020**, 121, 1, 26382.

2. **Performance of metal-catalyzed hydrodebromination of dibromomethane analyzed by descriptors derived from statistical learning.** *A. J. Saadun, S. Pablo-García, V. Paunović, Q. Li, A. Sabadell-Rendón, K. Kleemann, F. Krumeich, N. López, & J. Pérez-Ramírez,* ACS Catal. **2020**, 10, 11, 6129–6143.

3. **Dimensionality reduction of complex reaction networks in heterogeneous catalysis: from linear-scaling relationships to statistical learning techniques.** *S. Pablo-García, R. García-Muelas, A. Sabadell-Rendón, & N. López,* Wiley Comput. Mol. Sci. **2021**, 11, 6, 1540.

4. **Nuclearity and host effects of carbon-supported platinum catalysts for dibromomethane hydrodebromination.** *A. J. Saadun, S. K. Kaiser, A. Ruiz-Ferrando, S. Pablo-García, S. Büchele, E. Fako, N. López, & J. Pérez-Ramírez,* Small **2021**, 17, 16, 2005234.

5. **Electrochemical reduction of carbon dioxide to 1-butanol on oxide-derived copper.** *L. R. L. Ting, R. García-Muelas, A. J. Martín, F. L. P. Veenstra, S. T. Chen, Y. Peng, E. Y. X. Per, S. Pablo-García, N. López, J. Pérez-Rámirez, & B. S. Yeo,* Angew. Chem. Int. Ed. **2020**, 59, 47, 21072–21079.

6. **Generalizing performance equations in heterogeneous catalysis from hybrid data and statistical learning.** *S. Pablo-García, A. Sabadell-Rendón, A. J. Saadun, S. Morandi, J. Pérez-Ramírez, & N. López,* ACS Catal. **2022**, 12, 2, 1581-1594

7. **Mechanistic routes toward C_3 products in copper-talaysed CO_2 electroreduction.** *S. Pablo-García F. L. P. Veenstra, L. R. L.*

Ting, R. García-Muelas, F. Dattila, A. J. Martín, B. S. Yeo, J. Pérez-Ramírez, & N. López, Catal. Sci. Technol. **2022**, 12, 2, 409-417.

The contributions of Sergio Pablo-García to the previous articles are the following:

1. First draft, development of automation software, Density Functional Theory calculations and design and drawing of interactive figures.

2. Theoretical part of the first draft, Density Functional Theory calculations and statistical analysis of the chemical data. Shared Authorship with A. J. Saadun from ETH Zürich, who performed the wet lab experiments and provided the experimental data.

3. Automation section of the first draft, design and drawing of the figures.

4. Support with the automation of Density Functional Theory calculations to detect Transition States.

5. Visualization and exploration of different reaction paths included in the studied reaction network.

6. First draft, Density Functional Theory calculations and statistical analysis. Shared Authorship with A. Sabadell-Rendón, from ICIQ, who was in charge of the Bayesian Machine Scientist executions, data collection and filtering.

7. Theoretical part of the first draft, development of software for automatic network generation and analysis, density functional theory calculations and data analysis. Shared Authorship with F. L. P. Veenstra from ETH Zürich and L. R. L. Ting from NUS, who performed the wet lab experiments and provided the experimental data.

Abstract

The combination of Experimental observations and Density Functional Theory studies is one of the pillars of modern chemical research. As they enable the collection of additional physical information of a chemical system, hardly accessible *via* the experimental setting, Density Functional Theory studies are widely employed to model and predict the behavior of a diverse variety of chemical compounds under unique environments. Particularly, in heterogeneous catalysis, Density Functional Theory models are commonly employed to evaluate the interaction between molecular compounds and catalysts, lately linking these interpretations with experimental results. However, high complexity found in both, catalytic settings and reactivity, implies the need of sophisticated methodologies involving automation, storage and analysis to correctly study these systems. Here, I present the development and combination of multiple methodologies, aiming at correctly asses complexity. Also, this work shows how the provided techniques have been actively used to study novel catalytic settings of academic and industrial interest.

This works aims to display the development and use of a complete automation framework, able to automate and analyze catalytic problems of high complexity. In **Chapter 1** I present the current status and challenges regarding complexity in heterogeneous catalysis, how computational chemists traditionally dealt with it and novel methodologies to asses it. **Part I** presents the computational tools used to ensamble an automation framework. The part is divided in 4 chapters, **Chapter 2** that presents the DFT methodologies commonly used for surface chemistry, **Chapter 3** that demonstrates the significance of automation and geometry manipulation while presents a set tool to take benefit from both, **Chapter 4** that depicts the current need of statistical learning techniques to analyze scientific data and **Chapter 5** that portrays the service of applied graph theory in chemistry. **Part II** explores the combination and use of these methods to solve specific catalytic problems. Specifically, **Chapter 6** presents the building and use of two different workflows to automate Density Functional Theory ionic relaxations and transition state searches and as well as their storage, **Chapter 7** shows the combination of unsupervised and supervised

machine learning techniques to extract chemical descriptors from a hydrodebromination reaction and how to use them to predict its activity output, this procedure is then generalized to an extended hydrodehalogenation family, **Chapter 8** shows how to improve the techniques presented in the previous Chapter with applied graph theory to automate the complete generation of a reaction network and their use to unravel the paths to Propylene and 1-Propanol in the eCO_2R, **Chapter 9** takes advantage of the graph representation of molecules to propose the architecture of a Graph Neural Network to predict binding energies of large hydrocarbons training the network with small molecules with unique functional groups. Finally, **Chapter 10** includes the conclusions of this work.

Thus, this thesis presents a set of tools to ease the computational study of catalytic systems, as well as their combination to solve high complexity problems. I successfully tested these tools on different production environments to evaluate their usability and accuracy, being able to automate calculations, identify chemical descriptors, make activity predictions and explore complex catalytic networks. This works paves the way in the use of computational tools to deal with the high complexity growth found in heterogeneous catalysis.

Chapter 1

Introduction

A catalyst is a chemical compound that accelerates a chemical reaction being not consumed during the process. Thus, the catalysts offers an alternative path for the reaction, which is more complex but energetically favored. The ability to reduce the energetic requirement of chemical reactions made them the workhorses of chemical transformations in industry, with approximately 85-90% of the products of chemical industry produced using catalytic processes. They are also important in biological processes, as living matter relies in enzymes, a particular kind of catalysts.[1]

Catalysts can have multiple forms, such as atoms, molecules, enzymes or solid surfaces. Specifically, heterogeneous catalysis studies the reactions composed by reactants in gas phase or solution and a solid catalysts. These kind of catalysts have a high economical impact, as they are widely used in industry to produce base chemicals, polymers and pharmaceutical drugs among others. Moreover, their use is not limited to chemical production; they are found in a wide range of applications, for example: abating pollution or green chemistry.[2–5]

1.1 Catalysis Complexity

Chemicals of industrial interest range from small, simple molecules to composite molecules as polymers. Catalytic architectures used to synthesize these compounds are distinguished by high complexity, including an active material, carrier, molecular modifiers, dopants, and binders.[6–10] Computational methods fail at entirely reproduce these levels of complexity as their scope barely go beyond pure crystals and simple orientations.[11] Explicitly, in heterogeneous catalysis, complexity mainly arises from three factors: (i) possible adsorption sites (catalyst structure), (ii) reaction network size (number of species inside the network) and (iii) environmental effects caused

by external forces. Thus, the overall complexity of the system limits our understanding of the problem.

In an ideal case, the full environment of the catalytic architecture should be considered, including all the molecular moieties and their different states being evaluated using expensive ab-initio methods. In this case, the extrapolation to experimental observables (e. g. product activity) is straightforward as the obtained computational description is almost complete. Nonetheless, human and computational powers are limited; relying on alternative methods and strategies able to provide a partial description of the catalytic setting is needed. There are two main approaches to bypass complexity: (i) reduce the noise of the system by identifying its most essential parts (dimensionality reduction) and (ii) scaling ab-initio calculations from the atomic level to the macroscopic level (multi-scale modeling). The use of these methodologies is not restrictive, and they are usually combined to connect selected ab-initio values with macroscopic observables.

1.2 Finding Chemical Descriptors

Chemical descriptors were at first identified by Sabatier,[12] who found that the rate of the oxide formation can be link with a single parameter. These so-called volcano plots were introduced by Balandin.[13] Volcano-shaped functions link the catalytic activity of a certain setting with a key parameter of the catalyst (**Figure 1.1**). As they are able to reduce the number of variables needed to obtain the activity values for a certain catalytic setting, chemical descriptors can be considered complexity reducers. However, they are rather elusive and their identification is usually done after extensive ab-initio studies, requiring a considerable upfront investment to be detected.[14–16]

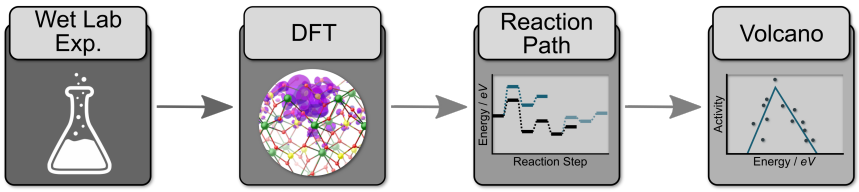


Figure 1.1: Traditional process to link experimental observables and DFT calculations. Adapted from Ref.[17].

Lately, these generalizations to families of reactants and catalysts were rooted by the phenomenological observations by Hammet [18] and Hammond postulates,[19] and Brønsted-Evans-Polanyi (BEP) [20, 21] equations, their nature lately being unravel by Density Functional Theory (DFT) through

Linear Scaling Relationships (LSR), linking thermodynamics to thermodynamics and kinetics. LSR wisely use chemical descriptors to infer thermodynamical and kinetic features. Within LSR, two different families can be discerned: (i) structural or topological features define the thermochemistry (e. g., describing the adsorption energy of intermediates) and (ii) thermochemistry defines the kinetics.[5, 22–30] The former has been successfully applied to metallic surfaces, with some degree of success in their application to oxides.[31–35] Kinetic parameters can also be approximated from thermochemical values using the BEP relations.[23, 36–39] For these relations, the Activation Energy (E_a), is approximated as a linear function of the Reaction Energy (ΔE). The factor multiplying the ΔE depends on the distribution of the Transition State (TS) among the reactives/products, being a value between 0.0 and 1.0, and 0.5 for symmetric reactions.[5, 40–42] LSR are commonly coupled with multi-scale models to predict experimental observables starting from a minimal set of DFT values.

1.3 Multi-scale Modeling

Multi-scale models are useful to connect computational data with experimental observables, for this purpose, ab-initio values are commonly introduced as input to estimate critical activity observables.[16, 43–46] Among the multi-scale methods, Microkinetic (MK) and Kinetic Monte Carlo (KMC) are commonly used in heterogeneous catalysis. MK is a mean-field method that for a given boundary conditions related with the experimental setting, solves the entire coupled Ordinary Differential Equation (ODE) system defined by each one of the different balances present in a reaction network.[46] MK and DFT can be coupled to provide an insightful analysis of the variation of composition during the reaction time.[5, 47–50] Yet, limitations in the use of DFT data for MK modeling arise with catalysts complexity, for example with highly anisotropic systems.[43] For these systems, KMC is needed.[51, 52] Instead of solving the ODE system defined in classical MK, KMC calculates the probabilities to transform the current state of a chemical system into a future state. Although KMC is able to describe more complex systems,[53–57] it is strongly demanding in terms of DFT evaluations, specially for systems that present coverage effects or anisotropy leading to a combinatorial explosion, requiring DFT evaluations for each of the included states and still not reaching chemical accuracy. In these cases, even using LSR instead of pure DFT values, use of KMC models becomes prohibitive.

Among other approaches, the use of data platforms and automation are able to mitigate this issue. Databases can be used to mine already-obtained data and feed multi-scale models, while automation relieves the method-

ological generation of novel datasets including DFT evaluations required by these models.

1.4 Databases

Scientific research rapidly grown during the last decades, this led to the increased production of scientific data. Experimental environment and results need to be strictly categorized and stored to be considered FAIR: Findable, Accesible, Interoperable and Reusable.[58] These needs led to the emergence of an increasing number of database platforms to store general or field-specific experimental data.[59–63] A single computational chemistry research project may generate around 10-15 TB of raw data, that later will be curated and uploaded to an online database. Before the genesis of these databases, scientific data was stored in physical formats kept in research centers and universities, while selected results were published into private journals or grouped into comprendia. Although this data was considered public, certainly only a selected people had (almost) instant access to these pieces of knowledge, as commonly bureaucracy was involved in the acquirement process. With the rapid deployment of internet, databases democratized data availability,[58] easing the access to scientific data to individuals not enrolled in academia. Scientific databases have become an standard, with Open Access and Data Management Plans being required by some institutions to financially support research projects.[64] Theoretical and computational fields have a long tradition of working with digital data and thus, they especially benefit from these computerized databases. However, computational data is diverse and heterogeneous, and multiple formats for the same kind of data are available, requiring further standardization to be easily accessible and interpretable.

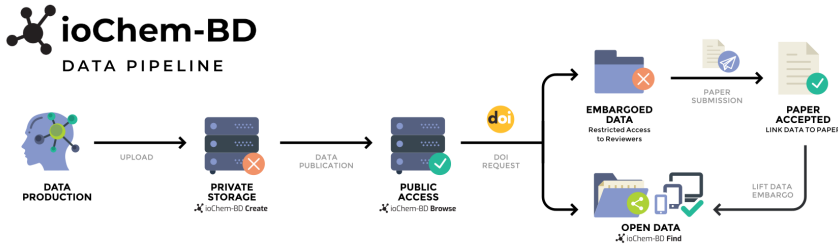


Figure 1.2: Worflow of the ioChem-BD Platform. Kindly provided by M. Álvarez.

Storing and labeling scientific computations enables its reuse for additional studies. The availability of ordered scientific data is specially useful

during the development and test of new computational models, as their results can be benchmarked against already known settings. Moreover, datasets addressing the same experimental problem can be merged, increasing the amount of information about specific processes. Still, scientific databases are a collaborative effort, and continuous contributions are required to maintain these platforms up-to-date. These databases, for example Materials Genome [65] and ioChem-BD [59] (**Figure 1.2**), are updated with novel studies that require the generation of new datasets for unique experimental conditions. Studies where both, the reaction network and the catalysts do not highly contribute to the complexity of the system, experimental and computational evaluations may manually done and later included to the database. However, for high complexity settings (dynamic catalysts, large reaction networks, multiple experimental environments, ...) automation is required.[66, 67] Laboratories have at their disposal a set of semi-automated or fully automated tools that aid performing the required experiments.[68, 69] For computational analysis, automation frameworks are also available, allowing to automate the storage of the generated data.

1.5 Deploying Automation

Since the development of the first ab-initio methods, a lot of effort has been put in their enhancement to be machine-efficient. Development of DFT methods and optimization led to the reduction of the computational time per evaluation. Together with the increasing availability of computational resources, enabled the simultaneous ab-initio modeling of multiple chemical systems. Since catalytic systems require the evaluation of large sets of singular chemical states, the bottleneck for such systems moves from calculation time to handling time. Considering that single calculations require input preparation, periodic verification and data gathering, scaling to thousand of evaluations is not straightforward.

Automation emerges to solve the scalability problem,[70, 71] being workflows its building blocks; they include the minimum set of instructions to perform a single procedure and commonly, instructions to handle problematic behaviors. Computational frameworks are able to oversee and manage complete DFT studies *via* dividing the involved steps into workflows. They have been successfully applied in heterogeneous catalysis to evaluate systems that included several chemical molecules with different geometries.[38, 66, 72],

Besides automation frameworks ease high-throughput studies, an isolated set of DFT evaluations is useless on their own, as chemical context is required to extract physical information from it. This is specially notice-

able for multi-scale methods, as they not only require thermodynamical or kinetic information, but also additional connectivity knowledge between the different components of a given chemical system. Connectivity evaluation is another step in the catalytic ladder, and as for automation, scaling is a critical issue.

1.6 Building Reaction Networks

Decomposition of molecules of industrial interest such as hydrocarbons or polymers usually undergo highly connected reaction networks composed by hundreds of intermediates.[5] Although characterization methods are in constant improvement, at the moment of writing this thesis, they are unable to track all the intermediates and reactions present during the reaction phase. Considering that limited information can be extracted from experiments, computational analysis require the complete evaluation of the molecules and paths that compose the reaction network. An additional level of entanglement is reached when catalytic surfaces are included in the process, as they present multiple adsorption options for the intermediates in the network. Accordingly, the entire evaluation of a reaction network requires automation.[73–75]

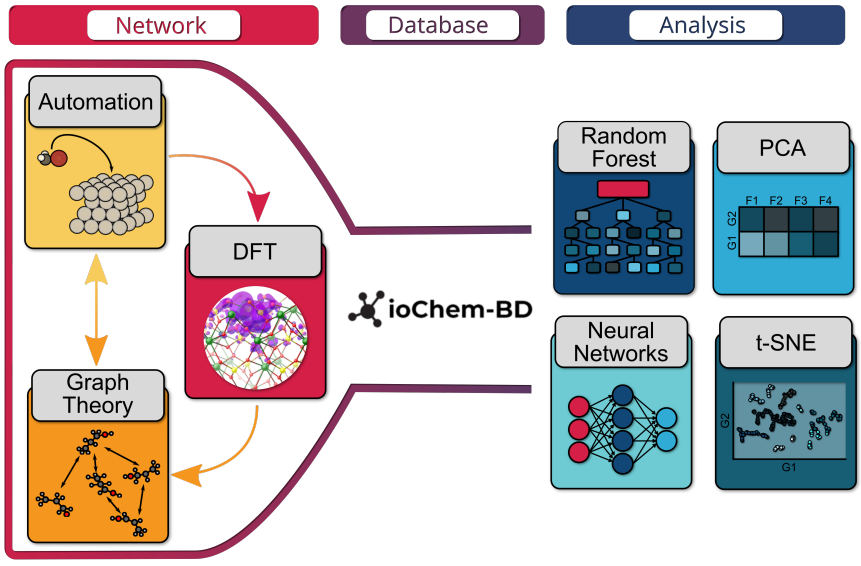


Figure 1.3: Summary of the methodology applied in modern DFT studies, including network automation, data handling and statistical learning analysis. Adapted from Ref.[17].

Mechanistic algorithms that store the connectivity of the network while

generating it can be employed to automate the process. Additionally, some of these algorithms are able to built the *ansatz* geometries of the components. In these cases, obtained structures are stored to be lately computed using DFT or other ab-initio methods.[76–80]

Self-building reaction networks can be combined with automated workflows to perform the full analysis of a reaction network while storing additional connectivity information, creating a complete description of a given catalytic setting. Multi-scale models are able to operate with these descriptions, however, they have some limitations difficult to surpass, and alternative statistics methods need to be considered. **Figure 1.3** depicts the complete process.

1.7 Statistical Learning Methods

Traditionally, data in computational chemistry was analyzed in small batches by a group of specialized scientists due to the high cost of producing a set of calculations. However, automation and data availability completely changed the paradigm, requiring tools able to analyze large quantities of data. Statistical Learning (SL) methods are able to provide this functionality. These methods are founded in the statistical analysis of data, learning from it and extracting valuable information or making predictions for unknown values. Their precision scale with the amount of data available, and thus are capable to achieve high accuracy levels with the needed amount data.

There are two large families of statistical learning methods: (i) unsupervised machine learning and (ii) supervised machine learning. The former is able to analyze the data to search for patterns while the second is noted for its prediction capabilities. Commonly, both methods are combined to obtain a deeper understanding of the studied system (**Figure 1.4**). In computational chemistry, unsupervised learning methods have been successfully applied to identify potential descriptors [82–84] of a given system and supervised machine methods have been successfully applied to make predictions on macroscopic features from DFT values,[85–94] bypassing the use of multiscale models.

1.8 Objectives

The scope of this work is to couple Big Data, Automation and Statistical Learning techniques to deploy a framework able to easily generate and link DFT and modeling with experimental results. To perform this task, this work explores the automatic generation of molecular structures, self-building

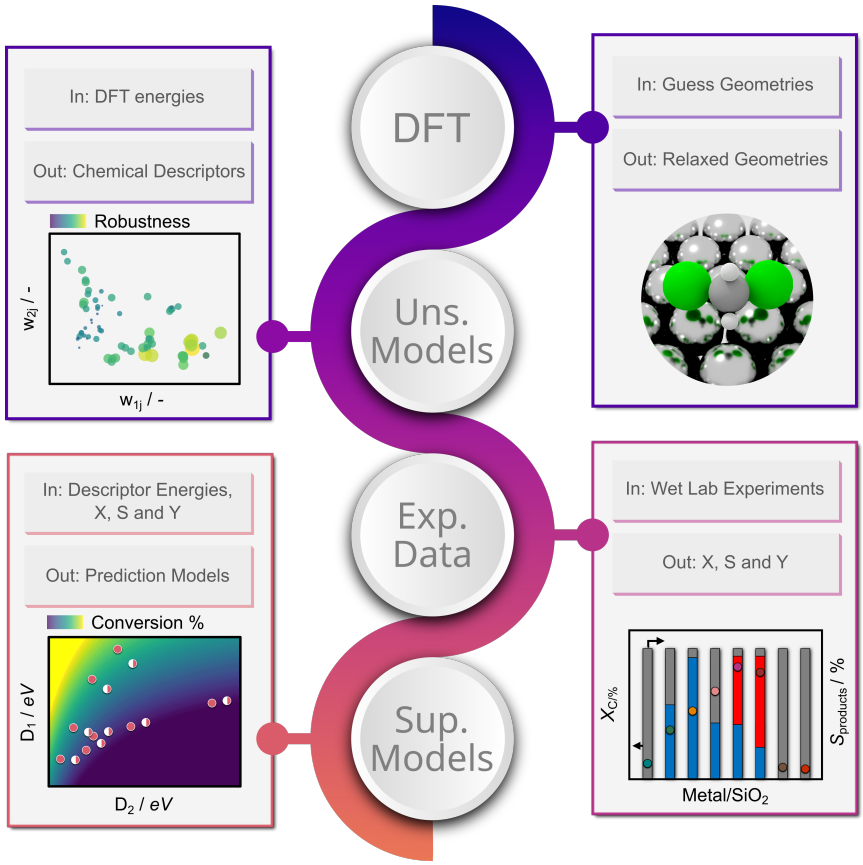


Figure 1.4: Statistical analysis procedure combining unsupervised and supervised mstatistical learning methods. Adapted from Ref.[81].

chemical networks and analysis of DFT datasets to extract chemical descriptors and produce predictions about its experimental behavior.

1.9 Summary by Chapter

- **Chapter 6** Creating a Simple Workflow

This chapter defines the development of a simple workflow to automate DFT relaxations and transition state searches on a different set of metallic surfaces. Created workflows are coupled to ioChem-BD, that performs the data curation and storage.

- **Chapter 7** Descriptors Search Using Statistical Learning

The aim of this chapter is to use the automation presented in the previous chapter to generate the data needed to study a chemical reaction of catalytic interest (CH_2Br_2 hydrodebromination reaction). Then, unsupervised machine learning techniques will be deployed to extract chemical descriptors from the DFT data. These descriptors will be employed in the prediction of experimental activities. Finally, the feasibility of generalize this procedure for the $CH_2X_2 : X \in \{F, Br, Cl, I\}$ reaction family will be explored.

- **Chapter 8** Analysis of the Carbon Dioxide Electroreduction Network

Modelling the Electroreduction of Carbon Dioxide (eCO_2R) on copper is an intrincate task due to the large number of reactions and intermediates composing the reaction network. During the eCO_2R on Oxide Derived Copper (OD-Cu) different C_3 and C_4 products such as 1-propanol and 1-butanol are obtained as products, however propylene, a compound of industrial interest for which a shortcut is predicted in certain areas, is not. The aim of this chapter is to identify the critical routes and paths that lead to the obtained experimental products *via* the exploration of the eCO_2R network and identify why propylene is not obtained. A combination of graph-theory and automation will be developed for this purpose.

- **Chapter 9** Convolutional Graph Neural Networks

The aim of this sub-chapter is to predict the DFT energy of organic molecules of industrial interest by training a Convolutional Graph Neural Network (GNN) with a selected set of key molecules with diverse functional groups.

- **Chapter 10** Conclusions and Outlook

Conclusions will be summarized and a outlook presented.

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Part I

Computational Background

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Chapter 2

Theoretical Chemistry and Density Functional Theory

To study the chemical systems presented in this work, their interatomic interactions need to be carefully described. These interactions have a crucial role at the nanoscale level and since in chemical reactions bonds are broken and formed, these phenomena is not being correctly described by classical mechanics. Then, quantum mechanics are needed in order to address this problem.

2.1 Schrödinger Equation

The state of a physical system can be described by means of wave functions through the use of the Schrödinger equation.[95] The time-independent Schrödinger equation is an eigenvalues equation where the $\hat{h}$ is the Hamiltonian operator and Ψ is the state vector of the quantum systems (**Eq. 2.1-2.2**). Ψ_n Eigenstates of the Hamiltonian are the solutions of the time-indepent Schrödinger equation, which have n associated eigenvalues such that $E_n \in \mathbb{R}$. Then, the position-space wave function of the system is defined by the expansion in terms of the positions eigenvector of the state vector (**Eq. 2.3**).

$$|\hat{H}|\Psi\rangle = |E|\Psi\rangle \tag{2.1}$$

$$\hat{H}\Psi_n = E_n\Psi_n \tag{2.2}$$

$$\Psi(\vec{r}) = \langle\vec{r}|\Psi\rangle \tag{2.3}$$

The definition of the Hamiltonian depends on the physical system. Due to the complexity at quantum level of chemical systems, some simplifications are needed to ease its calculation. Born-Oppenheimer approximation

states that as the nuclei is considerably more massive than the electrons, the response of the electrons to a movement will be faster than the response of neutrons/protons.[96] Thus, the contribution of the nuclei and the electrons can be decoupled.

$$\left[\underbrace{-\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2}_{\hat{K}}+\underbrace{\sum_{i=1}^{N}V(\vec{r}_i)}_{\hat{V}}+\underbrace{\sum_{i=1}^{N}\sum_{j<1}U(\vec{r}_i,\vec{r}_i)}_{\hat{U}}\right]\Psi=E\Psi \tag{2.4}$$

$$\langle\vec{r}|\Psi\rangle=\Psi(\vec{r}_1,\ldots,\vec{r}_N) \tag{2.5}$$

Applying this approximation, the Hamiltonian of a non-relativistic system of N electrons can be rewritten as in **Eq. 2.4**, where E is the ground state energy of the electrons, K is their kinetic energy, V is the potential generated by the nuclei and U the interaction between the electrons. **Eq. 2.5** describes the total wave function for the N electrons. it is important to note that an increment in the number of electrons also increases the difficult of the solution, being nearly impossible to analytically solve the equation manually. Numerical methods, further approximations and computers are required in the solution process.

$$\Psi(\vec{r}_1,\ldots,\vec{r}_N)=\Psi_1(\vec{r})\Psi_2(\vec{r}),\ldots,\Psi_N(\vec{r}) \tag{2.6}$$

Hartree product (**Eq. 2.6**) allows to define a second approximation to decouple the all-electron wave function as a product of single-electron wave function.[97] However, including this approximation is not enough to solve the time-independent Schrödinger equation due to its many-body problem nature. The wave function of each single-electron depends on the interaction between all the electrons U, requiring all the electrons to be determined before. Thus, the U term is defined in the Schrödinger time-independent equation but also required to obtain its solution.

$$n(\vec{r})=\langle\Psi|\hat{n}(\vec{r})|\Psi\rangle \tag{2.7}$$

$$\hat{n}(\vec{r})=\sum_{i=1}^{N}\sum_{s_i}\delta(\vec{r}-\vec{r}_i) \tag{2.8}$$

However the variable of chemical interest is the localization probability of n electrons in an arbitrary set of coordinates $(\vec{r},\ldots,\vec{r}_n)$, as it can be related with experimental observable. Therefore, the electronic density can be defined as in **Eq. 2.7-2.8** where $\hat{n}$ is the density operator, n is the number of electrons, δ the Dirac function and the second summ the spin s

of each electron i. Dirac delta is required to discard electrons occupying the same electronic states, as Pauli's exclusion principle shows that two fermions cannot be in the same state.

2.2 Hohenberg-Kohn Theorems

The basis of the DFT was introduced with the concept of an universal functional f of the electronic density $n(\vec{r})$ by Honenberg and Kohn.[98] Demonstrating the basic properties of this functional, they provided a powerful approach to find a solution for the Schrödinger equation.

These properties are defined by the two Hohenberg-Kohn theorems: (i) if n interacting electrons move in an external potential $v_{ext}(\vec{r})$, the ground-state energy is an unique functional of the density $n(\vec{r})$ and (ii) the ground-state energy can be obtained variationally: the density that minimizes the total energy is the exact ground-state density, corresponding to the many-body Schrödinger equation.

Then, instead of determining the all-electron position wave function, with a $3n - d$ form, it is possible to determine its ground-state electronic density, a function with a $3 - d$ form, to define the ground-state energy of a physical system. Henceforth, the functional is minimized by the ground-state electronic density of the system. With an ideal functional, $n(\vec{r})$ can be estimated by solving the many-body Schrödinger equation, and, by iteration, determine the ground-state electronic density/energy of the physical system.

2.3 Kohn-Sham Equations

For convenience, N-electron Schrödinger equation is decoupled into N single-electron equations to define this functional. Now, the space wave function is presented as a product of the wave functions for each electron.

$$\left[\underbrace{-\frac{\hbar^2}{2m_e}\nabla_i^2}_{\hat{K}} + \underbrace{V_N(\vec{r})}_{\hat{V}_N(\vec{r})} + \underbrace{e^2\int\frac{n(\vec{r'})}{|\vec{r}-\vec{r'}|}d^3\vec{r'}}_{\hat{V}_H(\vec{r})} + \underbrace{\frac{\delta E_{XC}}{\delta n(\vec{r})}}_{\hat{V}_{XC}(\vec{r})}\right]\psi_i(\vec{r}) = e_i\psi_i(\vec{r}) \qquad (2.9)$$

For i single-electrons, the resulting Schrödinger equations have the form of the so-called Kohn-Sham equations [99], as presented in **Eq. 2.9**. Where K is the kinetic energy of the single electron and v_n is the potential resulting from electronic interactions with the atomic nuclei.

$$V_H = e^2\int\frac{n\vec{r}}{|\vec{r}-\vec{r'}|}d^3\vec{r'} \qquad (2.10)$$

$$V_{XC} = \frac{\partial E_{XC}}{\partial n(\vec{r})} \tag{2.11}$$

The Hartree potential (V_H, **Eq. 2.10**), is the repulsion between a selected electron and the rest of electrons in the system ($n(\vec{r})$). For the Hartree potential, each electron contributes twice, as a single electron and in the electronic density term, and thus, being affected by a non-physical self-interaction effect. This effect is corrected in the last term of **Eq. 2.9**, *via* the exchange-correlation potential (**Eq. 2.11**), which is for consistency defined as a functional derivative. The E_{XC} is the only unknown value of the equation and its calculations varies depending on the method used.

$$n_1(\vec{r}) = 2\sum_i \Psi^*_{i,0}(\vec{r})\Psi_{i,0}(\vec{r}) \tag{2.12}$$

Once a convenient approximation is selected for the exchange-correlation term, the energy of the ground state for a given system is calculated via a self-consistent procedure solving the Kohn-Sham equations: (i) An initial guess for the electron density $n_0(\vec{r})$ is set, (ii) single-electron equations wave functions $\Psi_{i,0}(\vec{r})$ are defined by solving the Kohn-Sham equations, (iii) A new electronic density $n_1(\vec{r})$ is calculated applying **Eq. 2.8** to the single-electron wavefunctions and finally (iv) both $n_1(\vec{r})$ and $n_0(\vec{r})$ are compared. If they fall inside a defined tolerance interval, the former one is defined as the ground-state electronic density and the ground-state energy for the system is computed. Else, the process is repeated starting in step (iii) until convergence is achieved.

2.3.1 Exchange-Correlation functional

According to Pauli's exclusion principle, two electrons cannot be in the same specific electronic state if they also share the same spin. Thus, electrons with the same spin shall be separated, reducing the potential energy of the system. This reduction is known as the exchange energy of the system, and as Hartree-Fock approximates the all-electron wave functions into several single-electron wave functions, the energy of the single-electrons shall be corrected by a factor, known as the correlation energy. Both energies are included in the exchange-correlation energy (E_{XC}) presented in **Eq. 2.11**. Several approximations are possible with different ratios of accuracy/cost.[44].

2.3.2 Lecal-Density Approximations

The simplest approximation is the Local-Density Approximation (LDA), presented by Kohn and Sham in 1965.[99]

$$E_{XC,LDA}[n(\vec{r})] = \int n(\vec{r})\epsilon_{XC}^{homo}(\vec{r})d^3\vec{r'} \tag{2.13}$$

$$\frac{\delta E_{XC,LDA}}{\delta n(\vec{r})} = \frac{\partial \epsilon_{XC}^{homo}(\vec{r})n(\vec{r})}{\partial n(\vec{r})} \tag{2.14}$$

LDA is defined as follows: the exchange correlation-energy for an arbitrary system is derived assuming the exchange-correlation energy per electron at coordinates $\vec{r}$, $\epsilon_{XC}(\vec{r})$ is equivalent as in a homogeneous electron gas with the same local electronic density ϵ_{XC}^{homo} as presented in **Eq. 2.13-2.14**. Stochastic methods are then used to calculate a local exchange-correlation energy function for high-density electron gases, and then interpolated for intermediate and low-density electron gases.[100] For metals and other materials with constant valence electron density, LDA provides a reliable result, contrary, for systems with local density variation (atoms and molecules) LDA overestimates the bond energies [101] and thus, other approaches are required.

2.4 Generalized Gradient Approximation

Generalized Gradient Approximation (GGA) solves the density variation problem for system with slowly varying electronic densities *via* the inclusion of the spatial variation of the electronic density by taking into account the local electron density $n(\vec{r})$ and its gradient $\nabla n(\vec{r})$ (**Eq. 2.15**).

$$E_{XC,GHCA}[n(\vec{r})] = \int n(\vec{r})\epsilon_{XC}[n(\vec{r}), \nabla n(\vec{r})]d^3\vec{r'} \tag{2.15}$$

The most widely employed non-empirical GGA functionals are the Perdew-Wang 91 (PW91) [102] and Perdew-Burke-Enrzerhof (PBE). [103] However, GGA functionals do not account London dispersion forces, they may be problematic estimating the physisorption of molecules on metal surfaces.[104, 105] Thus, many corrections to include dispersion forces have been proposed: Grimme's empirical methods,[44, 106–108] vdW-DF,[109] vdW-DF2 [110], DFT-D3 [108] and the Tkatchenko-Scheffler method.[111] These methods are commonly used to improve the accuracy of the adsorption energies, specifically, this work employs the DFT-D2 correction reparametrizing the C_6 values for transition metals to avoid overbinding.[112]

2.5 Periodic Systems

In heterogeneous catalysis, modeling the adsorption of molecules over metallic surfaces is a common procedure needed in almost each catalysts evaluation. Metals are crystals, with an atomic distribution repeated over a three dimensional space.[113] The minimum atomic structure needed to model a metal is named primitive cell. When an adsorbate in also considered, the primitive cell is also expanded to include it. Periodicity is described by the Bravais lattice, usually composed by three linearly independent translation vectors $\vec{a}_1$, $\vec{a}_2$ and $\vec{a}_3$ representing three crystal axes.

$$\vec{R} = n_1\vec{a_1} + n_2\vec{a_2} + n_3\vec{a_3} \tag{2.16}$$

Primitive and translation vectors are obtained when each basis satisfy **Eq. 2.16** for $\forall i \in \{1,2,3\}, a_i \in \mathbb{Z}$. Primitive cell (or Wigner-Seitz cell) is defined by the three $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ primitive axes. This cell contains the minimum number of atoms needed to represent an arbitrary crystal structure. In general, the periodicity of crystal can fall into a set of fourteen different lattice types. For this project, we used metallic surfaces that fall into the hcp, fcc and bcc lattices.

$$\begin{aligned}
\vec{b_1} &= 2\pi\frac{\vec{a_2}\times\vec{a_3}}{\vec{a_1}\cdot\vec{a_2}\times\vec{a_3}} \\
\vec{b_1} &= 2\pi\frac{\vec{a_3}\times\vec{a_1}}{\vec{a_1}\cdot\vec{a_2}\times\vec{a_3}} \\
\vec{b_1} &= 2\pi\frac{\vec{a_1}\times\vec{a_2}}{\vec{a_1}\cdot\vec{a_2}\times\vec{a_3}}
\end{aligned} \tag{2.17}$$

$$\vec{G} = v_1\vec{b_1} + v_2\vec{b_2} + v_3\vec{b_3} \tag{2.18}$$

Reciprocal lattice of the Bravais lattice can be built from the reciprocal cell vectors $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ defined in **Eq. 2.17**. The lattice vector $\vec{G}$ is then defined as the combination of the reciprocal cell vectors and $\forall i \in \{1,2,3\}, u_i \in \mathbb{Z}$. Lattice vectors are commonly measured in Å and reciprocal lattice vectors in $Å^{-1}$. Different crystal planes can be produced cleaving a crystal bulk. Miller indices are used to present these crystal facets. $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ are used to define the intercepts (i_1, i_2, i_3) between the crystal axes an planes. The surface orientation of a given facet is then determined by the reciprocals of the intercepts $(i_1^{-1}, i_2^{-1}, i_3^{-1})$. If there is no interception between a plane and an axis, then 0 is defined as its index.

2.6 Blöch's Theorem

Different basis sets can be used to solve the monoelectronic Schrödinger equation. For isolated molecules, basis set composed by molecular orbitals are commonly used to simulate the properties of the chemical system. However, for solid state systems, plane-waves are a better choice as they include the periodicity of the system, significantly decreasing the computing time. Blöch's theorem describes this behavior. A linear combination of the lattice vectors ($\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$) defines the crystal lattice translation $\vec{R}$, under $\vec{R}$ the electron potential energy is invariant (**Eq. 2.19**). Blöch's theorem states that the eigenstates of the monoelectronic wave function can be written as a product of a cell periodic part, **Eq. 2.20** and wave-like part dependent on the wave vectors $\vec{k}$, called k-points, **Eq. 2.21**. An infinite summ over the plane wave is required to extract the exact eigenstate Ψ_j for each wave vector $\vec{k}$. Yet, a cutoff can be applied since the contributions of high kinetic energies plane waves are minor. k-points sampling defines the integration grid in the first Brillouin zone and a high k-points density increases the precision but also computational time.

$$U(\vec{r}) = U(\vec{r} + \vec{T}) \tag{2.19}$$

$$u_j(\vec{r}) = u_j(\vec{r} + \vec{T}) = \sum_{\vec{G}} C(\vec{k} + \vec{G}) e^{i\vec{G}\cdot\vec{r}} \tag{2.20}$$

$$\Psi_j(\vec{r}) = u_j(\vec{r}) e^{i\vec{k}\vec{r}} \tag{2.21}$$

2.7 Pseudopotentials

While Blöch's theorem allows the reduction of simulation time *via* an appropriate choice of k-points sampling and energy cutoff, the amount of plane waves required to take into account the fast oscillations of the valence electrons wave functions in the core region is high and thus, a plane-waves basis expansion over both core and valence electron dramatically increases the required computational power. Since core electrons are strongly bound to the nuclei, valence electrons are usually used to evaluate the chemical behavior of physical systems. Pseudopotential approximation allows the replacement of the core electrons and strong ionic potentials between the nucleus and core electrons with a weaker pseudopotential based on a set of pseudo wave functions. Thus, valence electrons are explicitly assessed for the numerical solution of Kohn-Sham equation while core electrons are estimated by the pseudopotential. r_c is the cutoff radius where all electrons and pseudopotential wave functions overlap. Pseudopotential are transferable and soft: (i)

they reproduce the properties of the inner electrons independently from the atomic valence electron and (ii) the plane-waves expansion of the valence electron must be limited to the lowest energy cutoff to decrease simulation time. Higher cutoff radius improve the softness while decrease the transferability. The most common pseudopotentials are the Vanderbilt Ultrasoft Pseudo-Potentials (US-PP) [114] and the Projected-Augmented Wave (PAW) [115] pseudopotentials. In this work, the PAW pseudopotentials have been employed due to their affinity with transition metals.

2.8 Computing Transition States

For catalyzed reactions, reaction kinetics are crucial to understand the behavior of the studied catalysts. For ab-initio evaluations, kinetic features of the studied system are evaluated *via* obtaining the TS involved in such system. There are a large set of procedures to guess TS for a given reaction. Specifically, in surface-chemistry there are two methods widely used to obtain such states, Nudged-Elastic Band (NEB) [116, 117] and Improved Dimer Method [118, 119]

2.8.1 Nudged-Elastic Band

Let $[\boldsymbol{R}_0, \boldsymbol{R}_1, \boldsymbol{R}_2 \dots \boldsymbol{R}_N]$ be an elastic band with $N+1$ images, where the first $\boldsymbol{R}_0$ and the last $\boldsymbol{R}_N$ images are fixed and their geometry and energy are an energy minima and the rest $N-1$ images are adjusted by an optimization algorithm.

$$F_i = F_i^S|_{\parallel} - \nabla E(\boldsymbol{R}_i)|_{\perp} \tag{2.22}$$

$$\nabla E(\boldsymbol{R}_i)|_{\perp} = \nabla E(\boldsymbol{R}_i) - \nabla E(\boldsymbol{R}_i) \cdot \hat{\boldsymbol{\tau}}_i \tag{2.23}$$

$$F_i^S|_{\parallel} = k(|\boldsymbol{R}_{i+1} - \boldsymbol{R}_i| - |\boldsymbol{R}_i - \boldsymbol{R}_{i-1}|)\hat{\boldsymbol{\tau}}_i \tag{2.24}$$

Then, in the NEB method, the energy of each non-fixed image is the sum of the spring force along the local tangent and the true force perpendicular to the local tangent (**Eq. 2.22**). Then the true force $\nabla E(\boldsymbol{R}_i|_{\perp})$ is calculated as in **Eq. 2.23** where E is the energy of the system and $\hat{\boldsymbol{\tau}}_i$ the normalized local tangent at image i. Finally, the spring force $F_i^S|_{\parallel}$ is defined as in **Eq. 2.24** where k is the spring constant. An optimization algorithm (e. g.: Velocity Verlet) to move the images according to the computed force.

However, the energy of the TS is interpolated using the central images as none of these images contains the exact structure/energy of the transition state. To address this issue, the Climbing-Image Nudged Elastic Band (CI-

NEB) is used.

$$\begin{aligned} \boldsymbol{F}_{i_{max}} &= -\nabla E(\boldsymbol{R}_{i_{max}}) + 2\nabla E(\boldsymbol{R}_{i_{max}})|_{\parallel} \\ &= -\nabla E(\boldsymbol{R}_{i_{max}}) + 2\nabla E(\boldsymbol{R}_{i_{max}}) \cdot \hat{\boldsymbol{\tau}}_{i_{max}} \hat{\boldsymbol{\tau}}_{i_{max}} \end{aligned} \tag{2.25}$$

CI-NEB constitutes a small modification to the NEB method in which after a few iterations, the image with the highest energy i_{max} is identified, and then the computation of its force is substituted as in **Eq. 2.25**. This leads to an image rigorously converging to the saddle point.

2.8.2 Dimer Method

Dimer method is local saddle-point search algorithm developed by Henkelman and Jónsson,[118] that uses only first derivatives of the potential energy, being appropriate for large systems. Dimer methods defines two points in a 3-n dimensional space (n being the number of atoms) sightly displaced by a distance $2\Delta R$. Location of the end points $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are defined as in **Eq. 2.26**.

$$\begin{aligned} \boldsymbol{R}_1 &= \boldsymbol{R}_0 + \Delta R \boldsymbol{N} \\ \boldsymbol{R}_2 &= \boldsymbol{R}_0 - \Delta R \boldsymbol{N} \end{aligned} \tag{2.26}$$

Where, $\boldsymbol{N}$ is the unitary vector along the dimer axis and $\boldsymbol{R}_0$ is the dimer mid point. The dimer method algorithms involves two steps: (i) the dimer axis is rotated into the lowest curvature mode of the potential energy at the midpoint of the dimer, $\boldsymbol{R}_0$ and (ii) the dimer is translated for an arbitrary step on the potential energy surface, moving it towards a saddle point. The curvature of the Potential Energy Surface (PES), C_N is calculated numerically along the direction of the dimer axis (**Eq. 2.27**), and therefore, more accurate than a curvature calculated with an updated Hessian, $\boldsymbol{H}$.

$$C_N = \boldsymbol{N}^T \boldsymbol{H} \boldsymbol{N} \approx \frac{(\boldsymbol{f}_2 - \boldsymbol{f}_1)^T \cdot \boldsymbol{N}}{2\Delta R} \tag{2.27}$$

Where $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are the forces acting on the points 1 and 2, and E_i the energy at the corresponding configuration. The minimum of the curvature in the rotation plane is equivalent to the minimum dimer energy $E = E_1 + E_2$. This plane, is spanned by the dimer axis and a normalized vector Θ, orthogonal to the dimer axis. Optimal direction of Θ is the one that produces the maximum overlap. The steepest descent direction of the rotation can be computed than as in **Eq. 2.28**.

$$\frac{\partial E}{\partial \varphi} = -\Delta R (\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \Theta \tag{2.28}$$

As Θ has to be parallel to $(\boldsymbol{f}_1 - \boldsymbol{f}_2)$, Θ is orthonormalized to $\boldsymbol{N}$. The efficiency of the dimer method algorithm relies in the rotation algorithm into

the lowest curvature spanned by $\boldsymbol{N}$ and Θ. In their work, Henkelman and Jónsson propose that the rotational force F on the dimer can be expressed as in **Eq. 2.29**.

$$-\frac{\partial E}{\partial \varphi} = F \approx A\ sin(2\varphi) \tag{2.29}$$

Where A is a constant. The lowest curvature is then found by rotating the dimer axis bi φ_{min} (**Eq. 2.30**)

$$\varphi_{min} = -\frac{1}{2}\ arctan\left(\frac{2F}{F'}\right) - \frac{\delta\varphi}{2} \tag{2.30}$$

Where F and F' are the rotational force and the rotational curvature respectively, obtained by performing gradient calculation at a dimer configuration rotated by a small angle $\delta\phi$ (**Eq. 2.31**)

$$\begin{aligned} F &= \frac{[(\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \cdot \Theta]_{\phi=\delta\varphi} + [(\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \cdot \Theta]_{\phi=0}}{2} \\ F' &= \frac{[(\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \cdot \Theta]_{\phi=\delta\varphi} + [(\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \cdot \Theta]_{\phi=0}}{(\delta\varphi)} \end{aligned} \tag{2.31}$$

Finally, in the second step of the dimer method, the midpoint is translated along a modified force $\boldsymbol{f}^\dagger$ with the force component along the dimer axis inverted (**Eq. 2.32**). Force at midpoint being computed as the arithmetic mean of the force at positions 1 and 2.

$$\boldsymbol{f}^\dagger = \begin{cases} -(\boldsymbol{f}_0^T \cdot \boldsymbol{N})\boldsymbol{N} & if\ C_N > 0 \\ \boldsymbol{f}_0 - 2(\boldsymbol{f}_0^T \cdot \boldsymbol{N})\boldsymbol{N} & if\ C_N < 0 \end{cases} \tag{2.32}$$

Dimer method was further modified by Heyden et al. [119] by including a series of improvements to increase its stability and performance. I used the improved version of the Dimer Method (Improved Dimer Method) to compute some of the transition states present in this work.

2.9 General computational details

Chemical systems in this work were described *via* the DFT computations implemented in Vienna Ab Initio Simulation Package (VASP) code [120]. Generalized Gradient Approximation with the Perdew-Burke-Enrzerhof functional (GGA PBE-D2),[103, 121] with the C_6 values reparametrized for transition metals [7] were used to obtain the exchange correlation energies. and plane waves with a cutoff energy of 450 eV were used to represent the inner and valence electrons respectively.[115] γ centered k-points mesh was generated by using the Monkhorst-Pack method.[122] The metallic surfaces

presented in this work were usually modeled as a four layer p(3 × 3) slab, while gas phase molecules where modeled in a box of 15 Å × 15 Å × 15 Å. A four layer p(3 × 3)-(111) slab was to model the fcc metals, a p(3 × 3)-(0001) slab to model the hcp metals, a p(3 × 3)-(110) to model Fe and a p(3 × 3)-(100) to model the OD-Cu slab. The vacuum of the slabs was implemented by adding a space in the cell of 15 Å in the z-direction, including dipole corrections due to the symmetry of the system.[123] For convergence, the threshold criteria were set as 10^{-5} eV for the electronic relaxations and 0.02-0.05 eV $Å^{-1}$ for ionic relaxations. The two upper metal layers and the adsorbates were allowed to relax, while the rest of the atoms were fixed. A frequency analysis has performed after each calculation to check the stability of relaxed structures and TS. CI-NEB and Improved Dimer Method were used to identify TS Structures. The Computational Hydrogen Electrode (CHE) was used as the electrochemical model when required.[124, 125]

Chapter 3

Automation and Geometry Manipulation in Computational Chemistry

As Robert Moore stated, the computational power of human-built computers has been increasing exponentially since he formulated his infamous law in 1965.[126] The increasing computational power and the development of new, more efficient algorithms have been two of the cornerstones that allowed the extension of scientific computational simulations.

In theoretical chemistry, the access to such computational power together with the development and popularization of DFT methods drastically decreased the human-time required to perform a chemical simulation. This time reduction has been unavoidable translated in a massive generation of chemical information. However, as data production increases, new problems regarding its generation, analysis and representation appears, requiring new approaches to be solved, leading to the **More is Different** motto.

In this chapter I will explore how computational chemists deal with size and representation problems through the development of computational tools and how these tools are applied in the field of computational heterogeneous catalysis.

3.1 Automation Tools

The normal workflow of a computational chemist starts with the creation of the input values required for a simulation. Usually, computations are performed in an external resource and a scientist needs to periodically supervise the status of the active processes remotely. Run-time errors are common (e. g. non-converging algorithm, power issues, technical problems, etc.) and

for small sets, easy to tackle for a single human being. However, as the number of required evaluations increases, it is almost impossible for a scientist to detect and correct all the possible errors. Moreover, these preparation tasks require highly repetitive workflows that are prone to human errors, decreasing the chances to obtain a successful simulation. Software platforms such as Aiida,[71] or Fireworks [70] provide a complete framework to deal with the problems raised during the calculation time. These packages allow the user to focus on the workflows required by the experiment and how to handle errors, forgetting about procedural errors and thus, preventing an unsatisfactory waste of human time.

Storage of data generated by scientific simulations raises a second problem as commonly this data is heterogeneous and scattered through multiple terminals. There is an increase concern regarding the scientific data handling; scientific databases need to fulfill a series of qualities to be considered robust and trustworthy. These qualities are accurately defined by the FAIR-Data Principles, defining that a fair data needs to include the following qualities: (i) Findable, (ii) Accessible, (iii) Interoperable and (iv) Reusable.[58] Many initiatives have emerged in material science to perform a proper data management, being the Materials Project,[65] NoMaD,[60] Materials Cloud,[61] and ioChem-BD [59] some examples. Most of these platforms provide the users with a great tool that eases the homogenization and handling of data.

Diversity found in the market of simulation packages also presents a challenge when database homogenization comes into play. Each software works with their own molecular abstractions and internal rules, thus, requiring particular input formats and producing unique output files. Attempts to solve issue have been made, beginning with the codification of molecular structures into raw cartesian coordinates (xyz format), topological representations of molecules (Simplified Molecular-Input Line-Entry System (SMILES),[127] and more recently Self-Referencing Embedded Strings (SELFIES) [128]), to more sophisticated techniques that store physical data and metadata obtained from the simulation (Chemical Markup Language (CML) [129]). Fortunately, most of the tools cited in the previous paragraph provide a robust translator to handle different file formats. However, still for condensed matter systems the situation is not fully resolved.

With the combination of molecular building tools, these automation frameworks provide a powerful environment that favors the design, production and storage of complex and large experimental settings, considerably decreasing the experimental time, and allowing more time investment during the data-analysis phase.

3.2 Geometry Manipulation and Visualization Tools

Humans are visual animals, and therefore, they need to envision the system they are working with to correctly operate with it. However, mathematical and thus, machine languages are not visual, and there is a need of an interface to allow the communication between abstract mathematical/machine concepts and humans. First, it is important to remark that since the development of chemical concepts, a lot of efforts have been made to apply these interfaces to chemistry, with examples such as the Lewis formulae,[130] three-dimensional molecular representations or chemical equations. Computer screens and the invention of unique interfaces such as mouse or touchscreen led to the development of complex software that allows the visualization and interaction with virtual objects that are not accessible in the human-scale. These virtual representations allow higher abstraction levels and ultimately help in the search of solutions to real-world problems.

Visualization tools are crucial at the building phase of scientific simulations; they ease the understanding of the problem while helping building the chemical system. There are a wide range of software available that fulfills this need: Avogadro,[131], jmol,[132] 3dmol.js [133] and VESTA [134] are only a small set belonging to a wide range of open-source alternatives. Additionally, two-dimensional plotting libraries that provide easy-to-use experiences such as ggplot2,[135] gnuplot [136] and matplotlib [137] also contributed to problem representation in a non-structural, more mathematical depiction.

These tools are incredibly useful to build fine-tuned, intuition-based systems. However, they are unhandy when they need to be used to create procedural and precise structures as they often require a human operator, prone to exhaustion errors. Therefore, different molecular building frameworks have emerged to manipulate chemical geometries in a more precise, workflow-focused fashion. The most popular alternative, Open Babel,[138] offers an exceptional framework to work with chemical abstractions. However, the growing adoption of Python [139] by the scientific community, followed by the developing of mathematical and scientific libraries such as NumPy [140] and SciPy,[141] and specific development environments (e. g. Spyder [142]) created a powerful ecosystem that motivated the reworking of these building tools to be integrated in the scientific Python ecosystem. One example of this integration process is Pybel,[143] that serves as a wrapper of the Open Babel [138] framework that can be scripted with Python. Also, new Python native alternatives have been emerged during the recent years. The Atomic Simulation Environment (ASE) [144] or pyMol [145] are two interesting packages, as they combine operator-building visualization tools with more precise worflow-driven scripting capabilities.

Before ending this section it is important to remark the impact of Blender

[146] in the chemical community. While it is not a visualization program designed to work with chemical constructs, it has been playing an important role in scientific outreach during the last years thanks to its capability to create more cinematographic representation of scientific systems.

3.3 Building a multi-purpose tool: pyRDTP

Systems modeled in heterogeneous require specific tweaks and procedures that are partially covered by the building tools mentioned in the previous section. To address this caveat, we developed pyRDTP, our own molecular building package.

pyRDTP is written in Python 3, allowing the integration with the packages of the scientific Python environment. The package is focused on the building and manipulation of catalytic systems. It provides an easy to build and manipulate abstract molecular representation, and also includes additional tools to communicate with other chemical packages. The package consist in a set of different modules and scripts that ease the building and exploration of catalytic systems, being the most important:

`molecule` Tools to build and manipulate abstract catalytic systems.

`geomio` Interface to read/write different molecular formats.

`graph` Allows to work with graph representation of molecules.

`networks` Building and exploration tools for reaction networks.

`scripts` Command-line scripts to modify input-files on-the-fly.

As `geomio` allows the conversion of the molecular abstractions to external formats, it provides an interface to work with ASE [144] and Pybel [143]. The capabilities of pyRDTP are documented using pydoc and can be found in its source code, which is freely available in gitlab.[147] pyRDTP is licensed under a MIT license, allowing free use and derivations. Most of the capabilities of pyRDTP rely on the work of external Python packages developed by natural and computer scientists, thus, requiring some dependencies be fulfilled before its use.

All the scientific research presented in this work has been made expanding pyRDTP with ad-hoc modules required by each research Project. Combined with the Fireworks automation tool [70] and the ioChem-BD data-handling platform,[59] pyRDTP has demonstrated to provide an excellent setting to work with heterogeneous catalytic systems.

Chapter 4

Statistical Analysis in Computational Chemistry

Automation tools, building methods and data-handling platforms give an excellent environment that eases the projection of complex computational chemistry experiments. Nevertheless, a higher accessibility to reliable databases and the availability of methods that simplify data mining materializes the need to explore and analyze large datasets.

This chapter traverses the state-of-the art of the statistical analysis tools applied to chemical data. Due to the growing scientific Python community, the tools presented here are available in the Python Package Repository.[148] It is important to remark that traditionally, statistical packages have been developed for the R programming language [149], and it still provides an exceptional resource for statistical analysis tools, although for simplicity they will not be presented here.

4.1 Size and Complexity in Heterogeneous Catalysis

In heterogeneous catalysis, complex systems containing (i) highly connected reaction networks, (ii) large combinatorial spaces and (iii) polymorphism (diverse catalytic phases) are common, and despite their simulation starts to become computationally viable, their analysis stills remains a field of research.

An example of complexity created by highly connected reaction networks can be found in the alcohol decomposition reactions. A simple alcohol network (ethanol) is comprises a total of 55 molecular intermediates and 215 reactions.[5] However, when all the C_3 alcohol decomposition networks are included, the number of compounds and reactions raises to 463 and 2266 re-

spectively,[150] creating a combinatorial explosion rendering the exploration and analysis *via* traditional methodologies implausible.

A different kind of problem raises when there is a need to find a link between computational and experimental data. For a simple catalytic reaction network, unwanted side reactions, surface poisons and adsorptions sites need to be considered to correctly to describe the systems via DFT methods, creating a large set of combinatorial structures. This is particularly true for systems that also present different phases as they contain multiple adsorption positions that are not properly described with experimental characterization. The entanglement that these problems introduce is not directly related to the size of the reaction network but instead to the inability to properly describe the reaction system, requiring an expensive parametric evaluation to find the most relevant chemical steps in the process.

Both these problems require unique solutions. While for highly connected reaction networks a tool ease the exploration of the network in their completeness is needed (preserving the complexity), combinatorial/polymorphic systems require a tool to identify and preserve its most relevant pieces (reducing the complexity).

4.2 The Statistical Learning Approach

Data availability has been attracting the attention of data scientists during the last decades, who have been developing new methods and algorithms to improve data mining, analysis and storage. The agile research performed in the data science field has heighten the availability of manipulation tools in addition to simplifying their usage. Due to the easy-to-learn/easy-to-use philosophy, Python [139] has become the default platform for data science developers. Thus, providing packages for data manipulation/analysis such as pandas,[151] that mimics the famous `data.frame` data structure from R and includes a complete set of statistical tools or scikit-learn,[152] that has become the current state-of-the-art package for the application techniques due to its exhaustive statistical learning toolbox.

Albeit there are a lot of powerful traditional statistical analysis methods, these methods have demonstrated a great potential in pattern searching and learning/prediction power for high-dimensional data. As scientific data falls into the last category, SL methods are a powerful tool to enhance data analysis of scientific data. These methods fall into two different groups.

(i) Supervised Statistical Learning, where an algorithm is trained with known input/output pairs with the aim to predict the output pair of unknown inputs.

$$train : m\ s \to [(X, Y)] \to m\ s \tag{4.1}$$

$$loss : [(Y, Y)] \rightarrow err \tag{4.2}$$

$$predict : m\ s \rightarrow X \rightarrow Y \tag{4.3}$$

Eq. 4.1-4.3 represent the phases followed by a typical machine learning algorithm, where m is the supervised model with a set of arbitrary hyper-parameters, s is the state of the coefficients of the model, X is an n-dimensional vector containing the features (input) of a sample and Y is an n-dimensional vector containing the output values. *train* function generates a model m with a new state s of its internal coefficients that tries to minimize the error err for a function of the form of $loss$ (included in the model). Note that $loss$ takes a pair of outputs (Y, Y) as input, usually these are pairs of experimental and predicted (obtained through *predict*) values. Least Absolute Shrinkage and Selection Operator (LASSO),[153, 154], Random Forest (RF),[155] and Artificial Neural Networks (ANN),[156] are some of the most common methods of this family.

(ii) Unsupervised Statistical Learning, where an algorithm learns patterns from the data. These methods are incredibly useful to extract meaningful information, discover hidden pattern or discard highly correlated features from a dataset.

$$fit : m\ s \rightarrow [X] \rightarrow [Z] \tag{4.4}$$

Eq. 4.4 represents the application of an unsupervised machine learning algorithm, where m is the supervised model with a set of arbitrary hyper-parameters, s is the state of the coefficients of the model, X is an n-dimensional vector containing the features (inputs) of the samples in the data and Z is an n-dimensional vector containing a set of new features collected by the model. k-Means,[157, 158] Principal Components Analysis (PCA) [159] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [160, 161] are the most common methods of this family applied in materials science.

4.3 Dimensionality Reduction Techniques

Unsupervised statistical learning contains a family of methods that aid in the complexity reduction of a given problem. Most of these methods try to reduce the number of features (dimensionality) required to represent a system in its completeness. As unsupervised statistical learning methods do not require a previous knowledge of the output values, these methods are widely used to reject low-score features that vaguely contribute to the description of an observable, thus, speeding up the training and prediction phases of supervised machine learning algorithms and increasing the knowledge about the studied system.

Nonetheless, compared to their supervised counterparts these algorithms have the capability of extract meaningful information from the processed data, whereas supervised statistical learning algorithms are incapable to do so due to their black box nature. Specifically, dimensionality reduction techniques are commonly used to extract descriptors from physical systems. In computational chemistry, PCA methods and t-SNE presented flawless results for classification and even prediction purposes.[82, 83] Subsequently, it is important to explore how they are applied and their intrinsic characteristics.

4.3.1 Principal Component Analysis

PCA performs an orthogonal linear transformation that projects the data into a new coordinate system, resting in its first coordinate (principal component) the scalar projection of the data with the maximum variance, in their second coordinate the scalar projection with the second maximum variance, and so on.

$$\boldsymbol{t}_{k(i)} = \boldsymbol{x}_{(i)} \cdot \boldsymbol{w}_{(k)} \quad i = \{1..n\} \quad k = \{1..d\} \tag{4.5}$$

$$\boldsymbol{w}_{(k)} = arg\,max \left\{ \frac{\boldsymbol{w}^T \hat{\boldsymbol{X}}_k^T \hat{\boldsymbol{X}}_k \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \right\} \quad \boldsymbol{T} = \boldsymbol{XW} \tag{4.6}$$

PCA is then applied as follows: Given a column zero-centered empirical mean matrix $\boldsymbol{x}$ of n samples and p features, there is a set of $\boldsymbol{w}_{(k)}$ p-dimensional vectors that transform the space of each row vector $\boldsymbol{x}_{(i)}$ to a new score vector $\boldsymbol{t}_{(i)}$, as show in **Eq. 4.5**. The weights $\boldsymbol{w}_{(k)}$ of each principal components are obtained as in **Eq. 4.6**, where $\boldsymbol{w}_{(k)}$ is the k-th eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{W}$ is a weights matrix whose columns are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$.

As result, the obtained principal components are sorted depending on their contribution to the variance of the system, allowing to discard the less contributing ones, non-meaningful components. As it allows to reduce the number of relevant features of a given system, PCA is commonly used to pre-process data before training a supervised statistical learning algorithm or to find highly correlated features in a given data series.

4.3.2 t-Distributed Stochastic Neighbour Embedding

Alternatively, t-SNE is an stochastic visualization method that groups the different samples of a given system depending their similarity. t-SNE is applied as follows: Given a matrix $\boldsymbol{x}$ with n samples and p features, t-SNE

first computes the similarity of two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as their probability p_{ij}.

$$\left\{ p_{j|i} = \frac{e^{(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2\sigma_i)^2}}{\sum_{x \neq k} e^{(-||\boldsymbol{x}_i - \boldsymbol{x}_k||^2/2\sigma_i^2)}} \mid p_{i|i} = 0 \ , \ \sum_j p_{j|i} = 1 \ \forall \ i \right\} \tag{4.7}$$

$$\left\{ p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \mid p_{ij} = p_{ji} \ , \ p_{ii} = 0 \ , \ \sum_{i,j} p_{ij} = 1 \right\} \tag{4.8}$$

p_{ij} are calculated as in **Eq. 4.7-4.8**, for σ_i being the bandwidth of the Gaussian kernels adjusted by the perplexity hyper-parameter. As t-SNE aims to learn a d-dimensional projection that reflect the similarities p_{ij} it measures the similarity between two projections $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ as q_{ij}.

$$\left\{ q_{ij} = \frac{(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2)^{-1}}{\sum_k \sum_{x \neq k} (1 + ||\boldsymbol{y}_i - \boldsymbol{y}_k||^2)^{-1}} \mid i \neq j \ , \ q_{ii} = 0 \right\} \tag{4.9}$$

$$KL(P||Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{4.10}$$

Similarities q_{ij} are calculated as in **Eq. 4.9**, following a Cauchy Distribution. Finally, the location of the points $\boldsymbol{y}_i$ in the projection are determined by minimizing the Kullback-Leibler (using gradient descent) divergence of the distribution P from the distribution Q, as shown in **Eq. 4.10**.

t-SNE produces a new projection of the data based in the similarity between the samples, creating groups that are easily recognizable visually. However, the distribution of the points along the new projection highly depends on the hyper-parameters (e. g. perplexity), requiring fine-tuning to represent the projection correctly.

4.4 Bayesian Symbolic Regression

Dimensionality Reduction techniques do not provide a full understanding of the studied system as further interpretations are needed to link the obtained patterns within the studied system. However, there an specific family of techniques that allow the extraction of meaningful information from the studied system. Symbolic regression are a set of methods that try to overcome this drawback with the inference of a mathematical equation that represents the model matching the studied system. As these approaches need to explore an extensive equation space, different algorithms have been proposed to achieve this goal.

1. g. genetic algorithms, greedy search and more recently ANN and Bayesian

methods.[162–165]

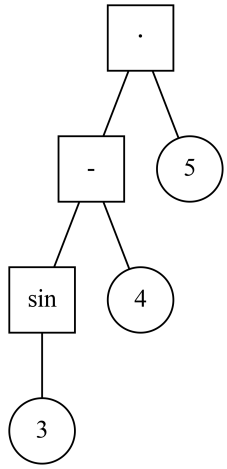


Figure 4.1: Graph representation of the (sin(3)-4) · 5 equation

Recently, Sales and Guimerà have been suggested the Bayesian Machine Scientist (BMS) algorithm for symbolic regression.[166] BMS represents the mathematical equation using a directed graph (**Figure 4.1**) that connects different operation nodes with variables and constants, creating an tree-like structure. It uses a Markov-Chain Monte-Carlo (MCMC) algorithm that mutates the graph abstraction of the mathematical model to explore the functional space. This mutation is perform allowing three different transformations: (i) node replacement, where an operation node is replaced by another operation, (ii) root addition/removal, where a new root is added or the current is removed and (iii) an elementary tree replacement, where a tree consisting in a single operation is replaced by a different one. Using an arbitrary probability distribution one of these transformations is selected and a decision of whether it is applied or not is made using the Metropolis-Hastings algorithm.[167] The calculation of the acceptance probability p_{accept} for each of the proposed transforms is unique and rather complex and the author strongly recommends reading the original article for a better understanding of the acceptance criteria. That said, it is important to remark that three factors affect the value of the 3 function $\mathcal{L}$, linked with the acceptance criteria of each transformation: (i) the length (complexity) of the model equation, (ii) the error in terms of Sum of Squared Errors (SSE) and (iii) a prior obtained from the comparison with known physical equations. The final aim of the algorithm is optimize the likelihood to explore local

minimums inside the functional space.

$$select : rng \rightarrow [(p_{select}, (m\ s \rightarrow m\ c))] \rightarrow (m\ s \rightarrow m\ s) \tag{4.11}$$

$$acceptance : m\ s \rightarrow (m\ s \rightarrow m\ s) \rightarrow prior \rightarrow [(X, y_e)] \rightarrow p_{accept} \tag{4.12}$$

$$decide : rng \rightarrow (m\ s \rightarrow m\ s) \rightarrow p_{accept} \rightarrow Maybe\ (m\ s \rightarrow m\ s) \tag{4.13}$$

$$transform : m\ s \rightarrow Maybe\ (m\ s \rightarrow m\ s) \rightarrow m\ s \tag{4.14}$$

The algorithm is summarized in **Eq. 4.11-4.14**: For a model m with their coefficients in an state s, a function $(m\,s \rightarrow m\,s)$ that performs a transformation to the model and its coefficients is randomly selected from a set of pairs of arbitrary probabilities p_{select} and transformations using a random number rng (**Eq. 4.11**). Then, the acceptance probability p_{accept} of the transform operation is obtained using the model, the transformation, a prior and a set of pairs of n-dimensional vectors X containing the features and their associated experimental outputs y_e (**Eq. 4.12**). The decision of whether the transformation will be applied or not is decided using a random number rng and the obtained p_{accept}, leading to the transformation $(m\,c \rightarrow m\,c)$ or $Nothing$ (**Eq. 4.13**). Finally, the decision is applied to the model (**Eq. 4.14**).

As supervised statistical learning methods presented previously, BMS can be used to find a model linking input/output pairs. Contrary to the supervised machine learning methods, BMS is able to return a model containing substantial information about the physical behavior of the system. These mathematical models are remarkably useful to connect ab-initio and macroscopic data, as they can be used to avoid the use of intermediate multi-scale models while keeping a significant amount of interpretability.

Chapter 5

Graph Theory in Computational Chemistry

Graph theory has been a helpful companion for chemistry; since the discovery of benzene by August Kekulé [168] to modern representations of molecules such as in SMILES,[127] chemists have been building graphs to represent molecules and chemical reactions. In this chapter we will explore the use of such representations to virtually build and manipulate molecules.

5.1 Molecules as Graphs

Interaction between atomic particles is the foundation of chemistry. Traditionally, these interactions have been studied through the solution of the Schrödinger equations. However, finding the solution of these equations for a complex chemical system is not trivial and usually requires costly numerical approximations that are performed by computers. Numerous methods that reduce the cost of solving the equations with more or less have been developed, being DFT approximations the most widely used. Yet, an outstanding number of chemical properties can be explained modeling molecules as graphs. As these models simplify the abstraction of a molecule (See **Figure 5.1a-b**), they are widely used build, understand and manipulate chemical compounds. The use of these representations has become mainstream in fields related to chemistry, being the term "molecular graph" included in different chemical glossaries.[169, 170]

$$G = (V(G), E(G)) \tag{5.1}$$

Mathematically, a graph G is a structure (**Eq. 5.1**) defined as an ordered pair of $(V(G), E(G))$ consisting in a set of objects (vertices, $V(G)$) and a set of connections between them (edges, $E(G)$), disjoint from $V(G)$,

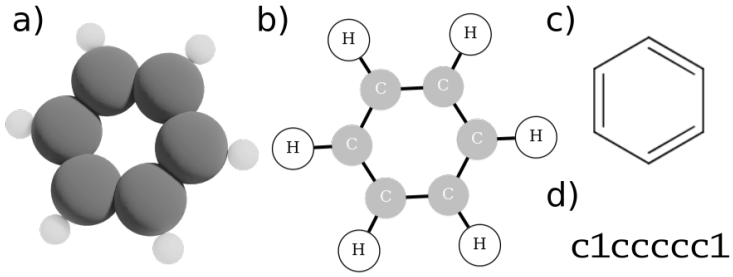


Figure 5.1: Different representations of the C_6H_6 benzene molecule. a) three-dimensional representation, b) graph representation, c) traditional chemical representation and b) SMILES representation.

and an incidence function ψ_G that associates each edge of G with and unordered pair of vertices of G.[171] The most simple graph structure is the undirected graph, in which the edges represent connections without any information about directionality. In chemistry, interactions between different atoms (bonds) are not directional, and such, molecules can be represented as undirected graphs. Graph representation of benzene (**Figure 5.1b**) shows an example of a set of objects V, representing the carbon and hydrogen atoms, being connected by a set of bonds E, representing the C-C and C-H interactions. Although this definition is incredibly useful to model simple compounds, the hybridization states of carbon atoms led organic chemists to include an additional abstraction layer to represent these states. As they constitute different bond strengths, hybridization states can be elegantly introduced adding a strength parameter to the edges of the graph representation. Graphs that include a strength (weight) parameter belong to a particular type of graphs named Weighted Graphs.

$$(G, w) \quad usually \quad w : E \to \mathbb{R} \tag{5.2}$$

Weighted graphs are denoted as (G, w), where w is a function that applied to an edge, e, returns the value of its associated weight, usually a real number (**Eq. 5.2**).[171] An example of the use of weighted graphs in organic chemistry is depicted with the molecular graph of benzene at **Figure 5.1c**, where parallel lines represent higher weighted edges, indicating a stronger interaction than bond represented with simple lines.

Graph theory is a mature and extent field of study [171, 172] and thus, an extensive number of properties, methods and algorithms related with graphs have been discovered, being all this knowledge transferable to molecular graphs. However, all the major ab-initio chemical packages available on

the market tend to use three-dimensional coordinates in their input/output formats. Then, to properly combine graph theory with automation and analysis procedures, conversion between molecular three-dimensional structures and graphs (and *viceversa*) is needed.

$$atom \cong (elem, coords) \tag{5.3}$$

$$pair : (atom, atom) \rightarrow ((elem, elem), (coords, coords)) \tag{5.4}$$

$$d_{exp} : (elem, elem) \rightarrow d \tag{5.5}$$

$$d_{calc} : (coords, coords) \rightarrow d \tag{5.6}$$

$$check : (d, d) \rightarrow Maybe\ bond \tag{5.7}$$

$$getBond : (atom, atom) \rightarrow Maybe\ bond \ \cong \ check \circ (d_{exp}, d_{calc}) \circ pair \tag{5.8}$$

$$toGraph : [(atom, atom)] \rightarrow ([atom], [bond]) \ \cong \ graph \tag{5.9}$$

Eq. 5.3-5.9 represent a simple the algorithm to converts a three-dimensional set of atoms into a graph: let d_{exp} be a function that given a pair of chemical elements $(elem, elem)$ returns an previously known, experimental distance d (**Eq. 5.5**), d_{calc} a function that given a pair of coordinates $(coords, coords)$ returns a distance d (**Eq. 5.6**) and *check* a function, that given a pair of distances returns a *bond* if they are considered bonded or $Nothing$ if they not (**Eq. 5.7**). Using function composition, these steps are wrapped into the function *getbond*, returning for a given pair of atoms $(atom, atom)$ either a *bond* or $Nothing$ (**Eq. 5.8**). Finally, *tograph* function maps *getbond* over a set of atom pairs $[(atom, atom)]$, and returns a pair of $[atom]$ (vertices) and $[bond]$ (edges) (**Eq. 5.9**), that have an equivalent graph structure $G = (V(G), E(G))$.

$$molecule \cong [atom] \tag{5.10}$$

$$getpairs : molecule \rightarrow [(atom, atom)] \tag{5.11}$$

To ease the conceptual understanding of the algorithm, a *molecule* can be defined as a set of atoms (**Eq. 5.10**). Then, a function *getpairs* that given a *molecule* extracts its atomic pairs $[(atom, atom)]$ is defined (**Eq. 5.11**). *getpair* can be as simple as returning the unique atomic pairs or as complex as the Voronoi expansion algorithm to detect adjacent atoms,[173] more reliable for high-bonded chemical systems (e. g. crystals). An implementation of the algorithm can be found in pyRDTP, allowing the conversion between three-dimensional molecules into graphs.

The inverse conversion, from molecular graph to a three-dimensional molecule, is convoluted as it requires previous experimental knowledge regarding bond distances, angles and dihedrals for different atomic combinations. More intricate approaches, such as the initial guess construction based on previous experimental understanding, with the later application of molecular forcefields are found in building packages as Open Babel.[138]

Another popular and extended graph representation is the SMILES codification [127] (**Figure 5.1d**). SMILES transforms the molecular graph of a compound into an string. The major advantages of a string representation over another types such as the adjacency matrix, are that strings are fluently shared and parsed. Thus, SMILES and another string representations have been widely used as input for ANN, as they compress complete bonding information molecules into a single string.[174–176] However, for solids this procedure is more intricate, due to the use periodic cells (using Periodic Boundary Conditions (PBC)) for their modeling.

5.2 Reaction Network as Graphs

$$Cl_2 + H_2 \longrightarrow 2HCl \tag{5.12}$$

As with molecules, chemists have a long tradition representing chemical reactions as graphs. As a simple example, **Eq. 5.12** shows the chemical representation for the HCl formation reaction, in which distinctive graph elements can be identified: Cl_2, H_2 and HCl are the vertices, while the arrow constitute two edges from the reactants to the product.

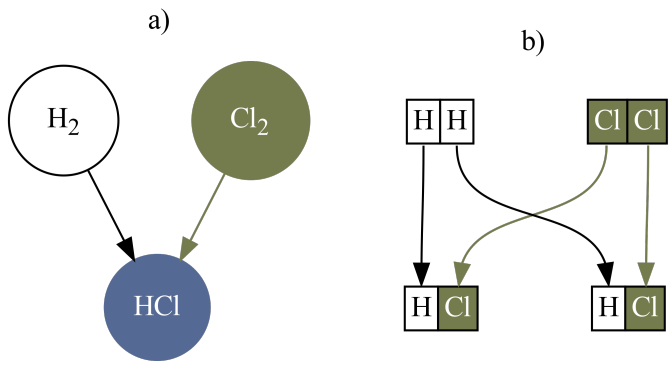


Figure 5.2: Graph representation of the HCl formation reaction. a) Simplified, not considering stoichiometry. b) Considering stoichiometry.

Figure 5.2 represent two possible graph representations of the chemical reaction, a) a simple graph where only the connections between the reactants and the product are considered and b) considering the stoichiometry of the reaction. These graph representations have different uses, the first being useful for exploration purposes, where only the connections of a reaction are important, and the former being effective to model simple reactions where keeping the track of the atoms is crucial.

It is important to point that while chemical reactions are an equilibrium between their components, it is common to find reactions that are greatly displaced to either the reactants or the products. These reactions are traditionally represented with a directional arrow $\longrightarrow$ while reactions in chemical equilibrium are portrayed with bidirectional arrows $\leftrightharpoons$. As directionality is crucial to understand chemical reactions, it must be included in their graph representation. Fortunately, directed graphs allow to solve this issue.

$$D = (V(D), A(D)) \tag{5.13}$$

Similar to undirected graphs, directed graphs D (**Eq. 5.13**) are defined as an ordered pair $(V(D), A(D))$ consisting in a set of vertices $V := V(D)$ and a set of arcs $A := A(D)$, disjoint from $V(D)$, together with an incidence function ψ_D (similar in undirected graphs), that associates with each arc of D an unordered pair of vertices of D. However, for directed graphs, for an arc $a \in D$ with $\psi_D(a) = (u, v)$, a is said to join the vertices u and v, and also that u (tail) dominates v (head).[171] Directed graphs allow a flawless representation of highly displaced reactions, unfortunately, this is not true for reactions close to equilibrium, as they require additional kinetic information to be fully understood.

$$Cl_2 + H_2 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} 2HCl \tag{5.14}$$

The equilibrium of a reaction can be defined by its kinetics constants k_i. For example, the formation of HCl (**Eq. 5.14**) can be described as a directed weighted graph, being the value of these constant associated with the weights of the connection arcs. **Figure 5.3** shows a weighted graph a), where the width of the arcs is linked with the reaction speed (assuming that the reaction is displaced to the products). Yet, in computational chemistry the kinetic constants are obtained through the explicit ab-initio computation of the transition state.

$$Cl_2 + H_2 \rightleftharpoons [Cl_2H_2]^{\ddagger} \rightleftharpoons 2HCl \tag{5.15}$$

Transition state theory [177] assumes the presence of a intermediate complex between reactants and products. The energy difference between reac-

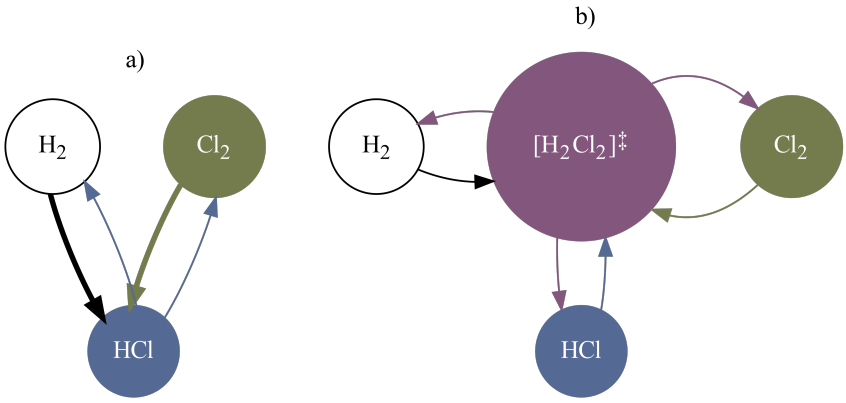


Figure 5.3: HCl formation reaction modeled as a graph, a) without including the transition state, b) including the transition state.

tants/products and their transition state is connected to the kinetic behavior of the reaction. As they contain kinetic and stoichiometric information about the reaction, transition states can be added as an additional participant to the chemical reaction, and thus, included in the reaction graph (**Figure 5.3b**).

$$TS \cong ([molecule], [molecule]) \tag{5.16}$$

$$vertex \cong Either\ molecule\ TS \tag{5.17}$$

$$arc \cong (vertex, vertex) \tag{5.18}$$

$$toGraph : TS \rightarrow ([vertex], [arc]) \cong dgraph \tag{5.19}$$

Algorithmically, a simple representation of a transition state would consist in a pair including two sets of reactants/products $([molecule], [molecule])$. (**Eq. 5.16**). As TS comprises the complete connectivity information of a chemical reaction, a function that converts a TS into a graph can be designed: let a *vertex* be a *molecule* or a TS (**Eq. 5.17**) and an *arc* be a pair of vertices $(vertex, vertex)$ (**Eq. 5.21**), then *tograpgh* function that given a TS returns a pair of vertices and arcs $([vertex], [arc])$ is defined (**Eq. 5.19**). Note that $([vertex], [arc])$ structure is equivalent to a directed graph $D = (V(D), A(D))$.

$$getparams : vertex \rightarrow p \tag{5.20}$$

To include additional details into the reaction graph, a function *getparams* that for a given *vertex* returns a set of arbitrary parameters p is defined (**Eq. 5.20**). As *getparams* allows to extract information from both the transition

states and the reactants/products, it can be used to depict the entire kinetic properties of the reaction, thus, allowing the definition of bidirectional reactions.

$$arc \cong (vertex, vertex) \tag{5.21}$$

$$unique : [TS] \rightarrow [vertex] \tag{5.22}$$

$$reactions : [TS] \rightarrow [arc] \tag{5.23}$$

$$graphnetwork : [TS] \rightarrow ([vertex], [arc]) \ \cong \ (unique, reactions) \tag{5.24}$$

With the defined graph representation for a singular reaction, the extension to a reaction network including n chemical steps (reaction graphs) is straightforward: as a single TS can be converted into a reaction graph (**Eq. 5.19**), there can be defined the *unique* and *reactions* functions that given a set of TS $[TS]$ return their unique $[vertex]$ and $[reaction]$ sets respectively (**Eq. 5.22-5.23**). Combining both functions, the *graphnetwork* function (**Eq. 5.24**) can be defined that given a set of $[TS]$ returns a pair of vertices (molecules and TS) and reactions (connections between the TS and the molecules) $([vertex], [arc])$.

Reaction graphs are remarkably useful during the exploration and analysis phase of a complex reaction network. Graphviz [179] and networkx [180] provide an outstanding set of tools for graph visualization, the former being part of the scientific python ecosystem. **Figure 5.4** shows an example of a reaction network automatically generated using the methods described in this chapter and networkx and graphviz software.

The combination of the graph representation of both molecules and reactions network has been recently used to build tools that automate the generation, analysis and exploration of complex reaction networks, with emphasis in MK modeling. [76–80]

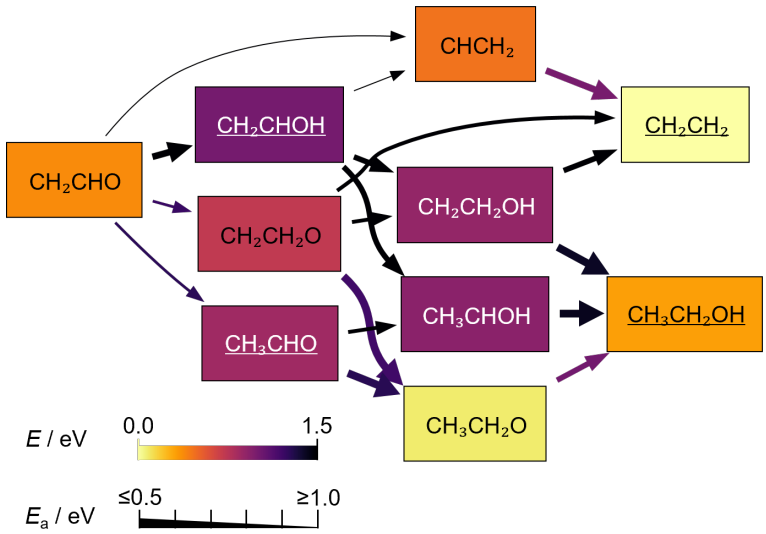


Figure 5.4: Simplified graph representation of a reaction network as a weighted directed graph. The edges portray the transition states while the vertices portray the intermediates. Molecular graphs were used to build the formula scheme for the intermediates. The color of the vertex (vertex weights) depicts the relative DFT energy of the intermediates while the width of the edges (edge weights) depict the activation energy E_a of the transition states. Visualization of the reaction graph was automatically generated using networkx and graphviz. Image adapted from the Supporting Information from Ref.[178].

Part II

Studying Catalytic Systems Using Computational Tools

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Chapter 6

Design of a Simple Workflow

6.1 Background

Catalyst discovery is commonly performed by testing a chemical reaction on a diverse family of catalysts. Therefore, it is needed to survey the behavior of the reaction on all the selected metallic hosts to acquire a complete insight of the setting. These surfaces (hosts) commonly present changes both in their composition and in their structure, leading to different possible metal/site combinations for a molecule to bind. Thus, the number of evaluations needed to cleanly study a catalyzed reaction on a set of metallic surfaces is then a product of three elements ($composition \cdot structure \cdot reactant$), which depending of their size may lead to enormous workload due to a combinatorial explosion. As stated in previous chapters, this tedious work has been traditionally performed by a scientist operator that builds all the possible combinations found in the reaction setting and supervises the calculations. Years ago, this was consider a fair procedure due to the low access to computational resources: the scientist needed to create a precise molecular *ansatz* and fine-tune the input parameters to avoid the waste of valuable computational time. Nonetheless, as the access to computational resources increases, the building of both molecules and input parameters, becomes the limiting step, drastically decreasing the research performance. Fortunately, automation methods can be used to tackle this challenge, decreasing the time wasted during initial guess building and fully automating the calculation process.

In this project, we combined our own molecule building package (pyRDTP), Fireworks [70] and ioChem-BD [59] to create a framework that automates the study of a chemical reaction on a set of different metallic surfaces. In particular, these packages were in charge of (i) generate and check molecular structures, (ii) prepare VASP environments and track the associated calculations and (iii) handle the generated data to follow the FAIR principles.[58]

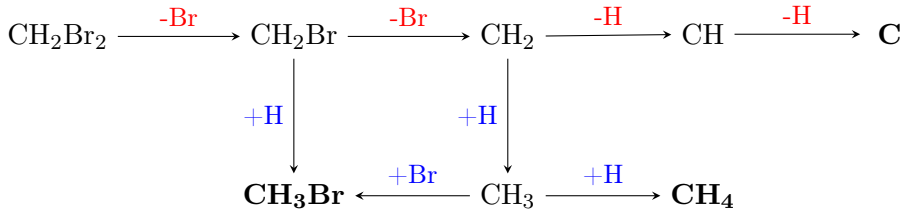


Figure 6.1: Scheme of the CH_2Br_2 hydrodebromination network. Adapted from Ref.[181].

We tested our framework with the dehydrobromination reaction (**Figure 6.1**) on different pristine metallic surfaces. First, we relaxed the intermediates and found the transition states for a single metal to then transfer the obtained structures to a set of nine similar metallic surfaces. For this purpose, two different workflows were designed, a simple one focused in obtaining relaxed structures and a more elaborate one focused finding transition states. Additional optimizations of the *ansatz* structures were made by inheriting already-known information from the original structure over similar catalysts.

6.2 Initial Guess Generation

Figure 6.2 shows the implemented algorithm using pyRDTP to create the initial geometry guesses inheriting structures from a given metal (host) over other metallic surfaces (guests). The algorithm works as follows: given a host slab with the relaxed molecule (1a) and an bare slab (1b), the layers are identified for both structures (2), the two upper layers are then selected (3) and their possible adsorption sites are detected (4), the two lowest atoms of the molecule in the host slab (5, 5o) are used to identify the topology of the adsorption site, the adsorption distance, and angle of the molecule (6, 6o) and finally uses the acquired information to transfer and align the molecule on the guest slab (6, 7).

The algorithm was used to transfer both, the already-relaxed intermediates and the transition states from ruthenium surface to a set of unique metals. Multiple alignments were used during this procedure, while for the relaxed states the geometries of the intermediates were carefully preserved from the relaxed host slab.

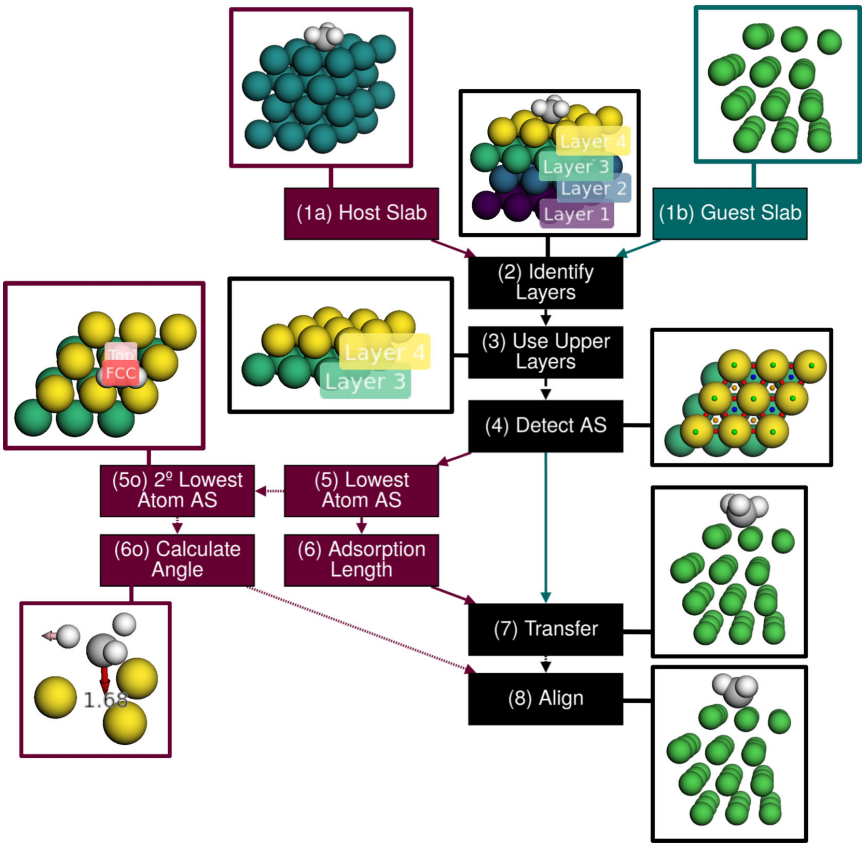


Figure 6.2: Scheme representing the algorithm implemented with pyRDTP to inherit already-relaxed structures from metal surfaces. Adapted from Ref.[182].

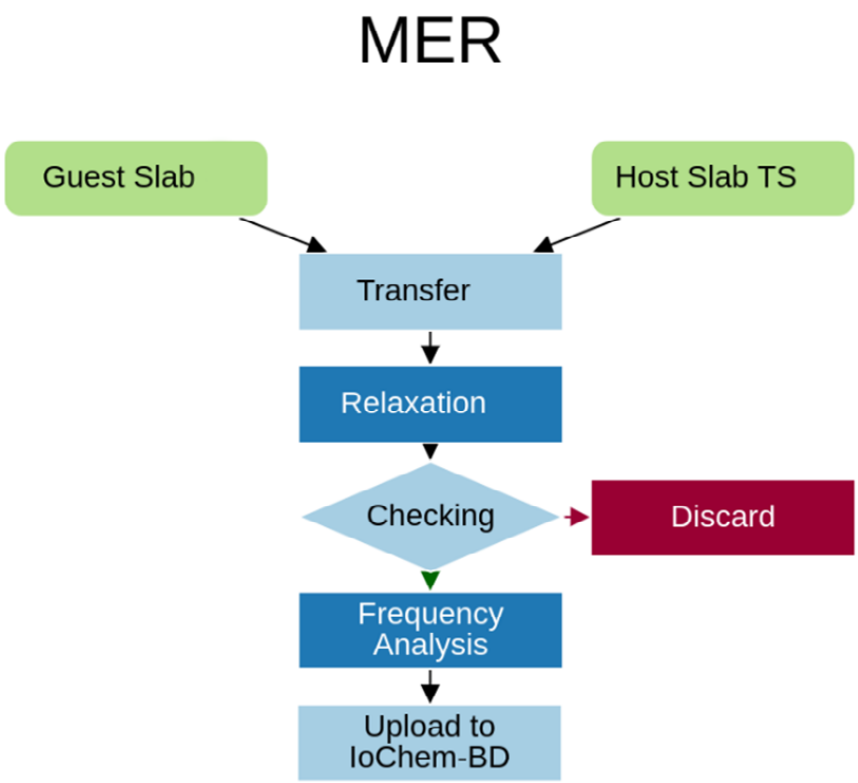


Figure 6.3: Scheme of the workflow used to calculate relaxed states (MER). Adapted from Ref.[182].

6.3 Workflow for Relaxed Structures

Using Fireworks, a simple workflow was created to automate the relaxing procedure of the newly generated *ansatz* of the hydrodebromination reaction (**Figure 6.1**) on additional pristine metallic surfaces. The workflow is shown in **Figure 6.3** and works as follow: Given a host pristine metallic slab with a relaxed intermediate, the intermediate is relocated on a similar slab using the transfer algorithm presented in **Figure 6.2**. Then, an input setup is deployed and sent to a computational resource, that performs the relaxation of the given input using VASP. Once the computation finishes, Fireworks checks (using pyRDTP) if the geometry presents any anomaly, if found, the calculation is discarded and an alert is sent to an human-operator. Contrary, if no errors are detected, another input is prepared and sent to the computational resource to calculate the vibration frequencies using VASP. Once the calculation finishes, both the frequency analysis and the relaxed structures are uploaded to ioChem-BD. Note that after each step the exit code is checked for anomalies. If a non-zero code is found, Fireworks will create a create a report and set the workflow to an stop status, waiting to be resumed.

6.4 Workflow for Transition-States

Figure 6.4 shows the procedure followed to shape TS guesses. Compared with the workflow used for relaxations (**Figure 6.3**), it generates multiple guesses and tries to select the one with the lower energy as the best candidate. Additionally, it performs an extra step to confirm the nature of the transition state. The routine works as follows: a transition state already obtained for a host pristine metal slab is relocated on a similar metal slab (guest) using the transfer algorithm previously described (**Figure 6.2**). Moreover, the last step (align) is modified, generating of three different candidates depending on how the applied rotation: (i) no align as applied, only the lowest atom is used, (ii) a thirty-degree rotation through the axis perpendicular to the surface is applied using the lowest atom as rotation point (this step is arbitrary and decided due to the hcp/fcc nature of the slabs) and (iii) using a second atom to align the transition state, as described in the algorithm. The three structures are deployed and sent to a computational resource and carefully relaxed with VASP using a small number of ionic steps. The output is collected and the energy of the three candidates is compared, selecting the one with the lowest energy. The selected geometry is then prepared for an improved dimer method using VASP and sent to a computational resource. When the calculation finishes, the integrity of

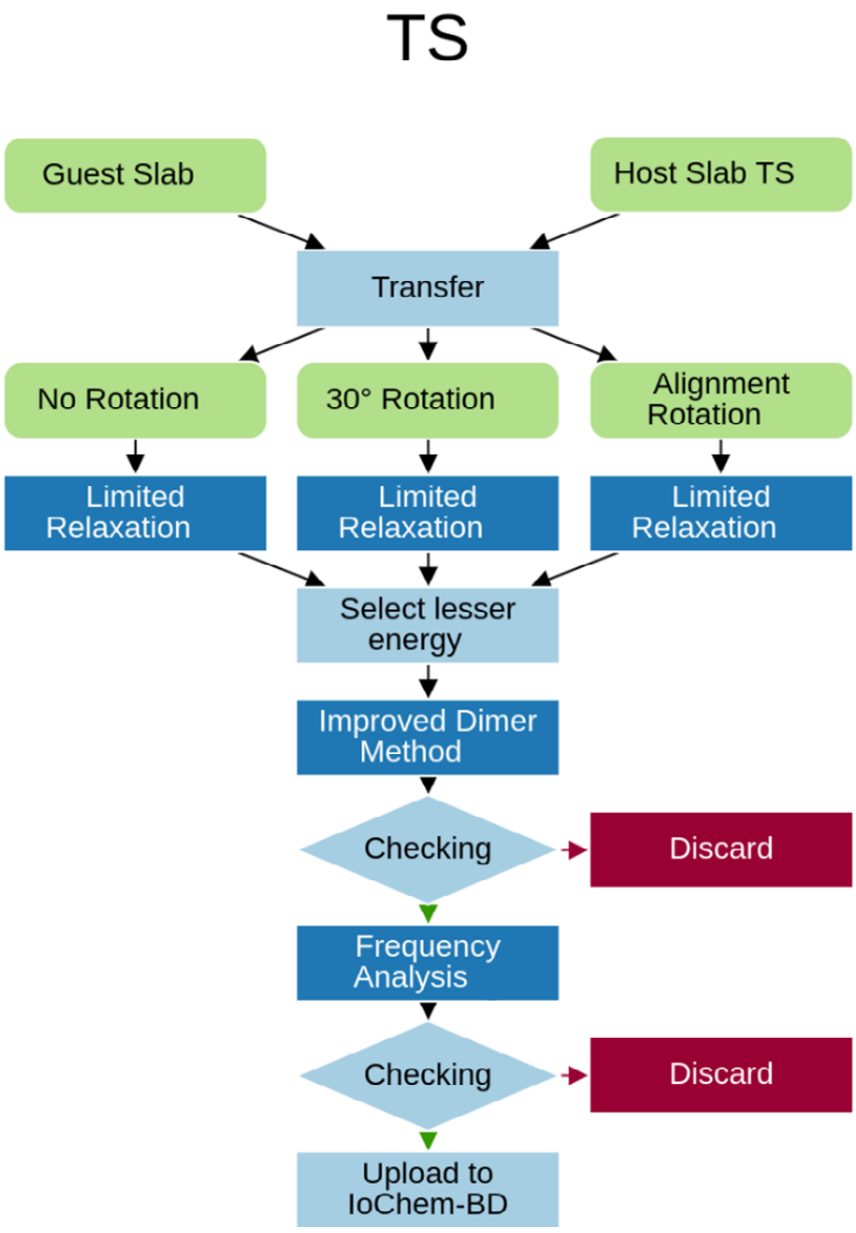


Figure 6.4: Scheme of the workflow used to calculate Transition State (TS). Adapted from Ref.[182].

the transition state is tested, if any anomaly is found, then the calculation is discarded and an alert is sent to a human-operator. If there are no anomalies, a frequency analysis is performed using VASP. After the analysis, the output is retrieved and an additional control is performed to search for a single imaginary frequency. If a number of imaginary frequencies distinct than one is detected, then Fireworks stops and sends an alert. Contrary, if only one imaginary frequency is found, the result of the improved dimer method and the frequency analysis are uploaded to ioChem-BD. As for the relaxed structures, the exit codes of the scripts executed are verified after the end of each step, creating a log and stopping the procedure if a non-zero exit is found.

6.5 Benchmark

Figure 6.5 shows the success rate among the two different workflows used. "Good" calculations (the ones that finished without any issue) represent the 88% of the relaxations and the 56% of the transition state searches. "Additional Steps" were required for 0.1% of the relaxations and 27% of the transition states. For the relaxed structures, the needed additional steps were mainly a 10% of additional ionic steps, while for the transition states two different approaches were used: (a) adding additional ionic steps and (ii) inherit from a different host metal. The rest of the calculations did not finish using the main workflow or additional steps and "Manual Preparation" by a scientist was needed in order to obtain them.

6.6 Integration with ioChem-BD

Both relaxation and transition state workflows share a last step that sends the generated outputs to ioChem-BD [59] to perform the data management. ioChem-BD automatically process the received data and stores it in a repository. Active projects that need dynamic data handling (adding or removing calculations) are stored in a private repository. Once the project is finished, the data can be set into an embargo state, not allowing any changes to be made. Embargoed projects are marked with a DOI [183] and a link to allow external connections is generated. Embargo state is commonly used to allow reviewers visit the dataset associated with a manuscript without making a final publication of it. Finally, the dataset is sent to a public repository and associated with the published article, making it accessible for any researcher, satisfactorily fulfilling the FAIR principles.[58, 184]

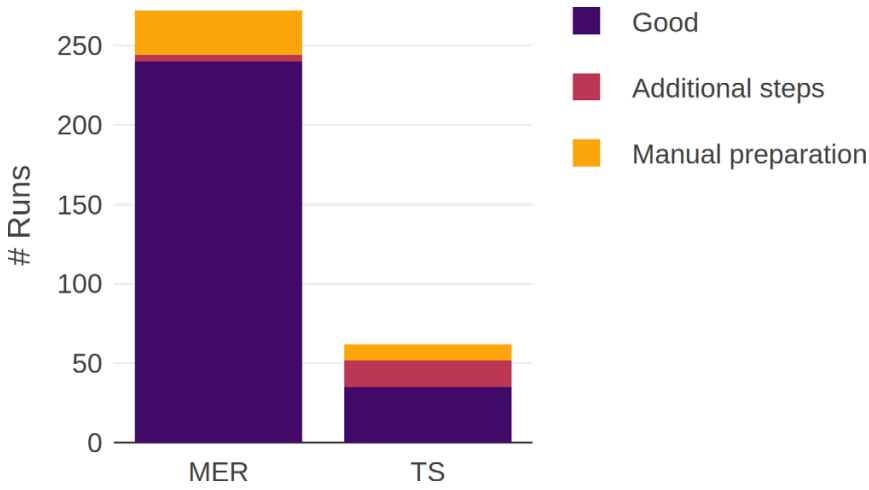


Figure 6.5: Benchmark for the success rate of the relaxed and transition state workflows. "Good" stands for workflows that finished flawlessly, "Additional Steps" "Additional Steps" stands for calculations that needed additional relaxation steps to converge or being inherit from a different host and "Manual Preparation" stands for evaluations that were unable to to be obtained using the presented method and required the preparation from an operator. Adapted from Ref.[182].

6.7 Authorea

This work was published using Authorea [182, 185] an online platform that allows the creation of interactive preprints. The original figures of this article are interactive and integrated with ioChem-BD. The author strongly recommends the reader to visit the online version of the article as it serves to show how ioChem-BD can be integrated with research articles to quickly access and visualize experimental data.[182]

6.8 Final Remarks

The automation workflow presented in this chapter was successfully used to automate part of the DFT calculations made in Refs.[81, 181, 186].

Chapter 7

Descriptors Search Using Statistical Learning

Previously, we defined a workflow that automates the data production for an arbitrary catalytic study. We used the provided automation framework to generate all the data needed for the study presented in this chapter. Nonetheless, this Chapter is dedicated to the analysis of the already-generated and labeled data, and the author redirects the reader to the previous chapter for more information about the automation procedure.

7.1 Background

Research of efficient and innovative methods to on-site transform of natural gas into fuels and chemicals has become an strategic growing field during the last decades.[187–190] Halogen mediated processes have emerged to transform CH_4, the main component of natural gas, into halogenated transportable liquid [191, 192]. Preferring bromine over chlorine as it provides higher selectivities to the desired product, bromomethane.[193–195] Whereas dehybromination reaction has been studied over Ru, Rh, Pd, Ag, Pt and Au, other know dehydrogenation catalysts such as Fe, Co, Ni, Cu and Ir were never studied in this reaction and the stability performance.

Catalysts development has been linked to the Sabatier principle,[13, 196] which is the "not-too-strong-not-too-weak" rule for the optimization within a family of catalytic materials. This rule was formulated deriving the volcano-shaped functions of a single energy value acting as a catalytic activity descriptor that populate the research works in heterogeneous catalysis. Nonetheless, identification of chemical descriptors is not straightforward, and are chosen, to some extent, by arbitrary heuristics.[14–16]

MK and KMC models are the gold standard in heterogeneous cataly-

sis to predict experimental behavior by using the DFT energies relations of the intermediates found in a certain reaction network over different metallic surfaces.[5, 197–209] Still, coverage effects, catalyst phase or surface transformations and large reactions networks where elementary steps grow exponentially, highly prevent the use of both methodologies. However, the growing development of SL, provided the needed tools to surpass this barrier. SL techniques are able to both, search for the descriptors of a given catalytic system while linking its experimental behavior with DFT descriptors,[85, 210, 211] avoiding the use of MK and KMC models. Unfortunately, SL models present serious drawbacks in terms of interpretability as they provide little information about the nature of the link between DFT and experimental dimensions. Recently, a series of Bayesian Learning methods have emerged to overcome this issue.[113, 212–214]

In this work, we use a toolkit consisting in unsupervised and Bayesian learning method to try to, first, identify the descriptors and predict the experimental behavior of a single reaction network (CH_2Br_2) hydrodebromination (**Figure 6.1**), and finally extend this methodology to a family of CH_2X_2 dehalogenation reactions.

7.2 Descriptor Identification

As they can be combine to assemble the famous "*vulcano plot*" [13, 196, 215, 216] chemical descriptor are one of the best known methods to connect experimental and theoretical data in heterogeneous catalysis. These plots allow to qualitative infer the behavior of a non-tested catalysts by the solely use of their intrinsic descriptors. However, finding chemical descriptors is challenging and remains an active field of research.[82]

7.2.1 Principal Component Analysis Method

García-Muelas and López presented a novel method to find descriptors in squared catalytic systems (systems composed by a group of intermediates adsorbed on a set unique metallic surfaces).[83] This procedure uses a PCA to extract and rank the principal components of a chemical system as well as their associated scores and contributions by the use of the associated DFT binding energies (surface/intermediate). Thus, descriptors can be identified by three singular qualities, (i) they predominantly contribute to only one component, (ii) this contribution is made unambiguously to one of the main components and (iii) their contributions remain stable among all observables. In the same work, it was exposed that components may be associated with physical properties and the two main components are connected with the covalent and redox potential of both, species and metals.

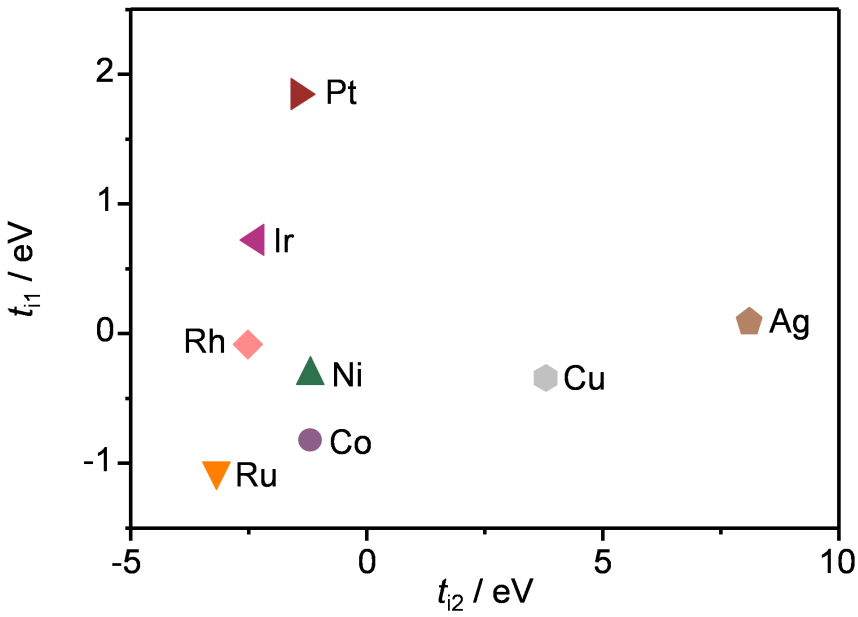


Figure 7.1: Contribution to the two main components for a different set of transition metals obtained using the PCA. Adapted from Ref.[181].

7.2.2 Obtaining the Descriptors

Therefore, we applied the PCA procedure to a squared system, composed by a group of metallic pristine surfaces (Ru, Rh, Pt, Ir, Ni, Co, Fe, Cu, Ag) and the set of intermediates that form the dehybromination reaction. We evaluated the results of the PCA, elucidating that the obtained components and the distribution of their contributions among the metals and adsorbed species are similar to the ones obtained in Ref.[217]. Specifically, we identified three main components contributing by 92.8%, 5.2% and 1.2% to the variance of the system. Therefore, the two main components account for 98% of the variance of the adsorption energies, and are thus taken as relevant. Consequently, seeking for the qualities previously presented, we disclosed two principal descriptors for these components, the CH molecule and the Br atom respectively. Notice that CH fragment presents a strong covalency while Br present strong redox potential, which is in agreement with the physical characteristics presented in Ref.[217].

As a final remark, only the binding energies were included in the PCA, avoiding the inclusion of the energies of the transition states. Despite we tried to include them during our first benchmarks, we obtained similar results in both cases. Additionally, we found that most of the transition states contribute to the system as a combination of their associated species (reac-

tants and product), while their robustness is lesser than for their associated isolated species, creating spurious noise.

7.3 Experimental Activities and the Bayesian Machine Scientist

The final aim of this work was to predict the experimental values for conversion X_{CH_2Br}, selectivity to CH_3Br, S_{CH_3Br}, and yield to CH_3Br ,Y_{CH_3Br}, coupling them with DFT energies. Usually, this procedure is made through the use of MK modeling. However, MK models are not suitable to model chemical systems that present surface poisoning, requiring the use of alternative methods.[43] Particularly, CH_2Br_2 hydrodebromination presents two problematic behaviors hard to tackle using a MK model: (i) surface presents poisoning and (ii) possible phase changes of the catalyst. Experimental results to train and contrast our models were provided by our collaborators, Saadun et al. from ETH Zürich.

7.3.1 Random Forest Regressor

To bypass the use of MK modeling, we trained a random forest model [155] using pairs of (i) the energy descriptors identified *via* PCA and (ii) their associated experimental outputs, conversion, selectivity and yield (**Figure 7.2**, left). In a first qualitative analysis we were able to identify four clusters among the metallic catalysts: (i) poor hydrodebromination activity (Fe, Co, Cu and Ag), (ii) intermediate activity and selectivity to CH_3Br (Ni and Rh), (iii) great propensity to CH_4 besides producing CH_3Br (Pt and Ir) and finally (iv) high selectivity to CH_3Br (Ru). Although the amount of data was small, RF presented different trends across the E_{ads} of the descriptors, metals presenting high values of $E_{ads}(CH)$ have low activity and are selective to coke, metals with high $E_{ads}(Br)$ and low $E_{ads}(CH)$ have high activity and are selective to CH_4 and finally metals with low $E_{ads}(Br)$ and high $E_{ads}(CH)$ have different selectivities to CH_3Br that decrease as they move away from Ru (peak of the vulcano).

7.3.2 Bayesian Machine Scientist

RF methods present three important drawbacks: (i) their black-box nature masks the physical interpretability of the trained model, (ii) their accuracy and mapped area are closely related to the number of samples employed during their training and (iii) extraction a functional form is difficult due to their stochastic, swarm nature. Thus, we employed the BMS method [166] to extract a functional form able to map the experimental behavior of the

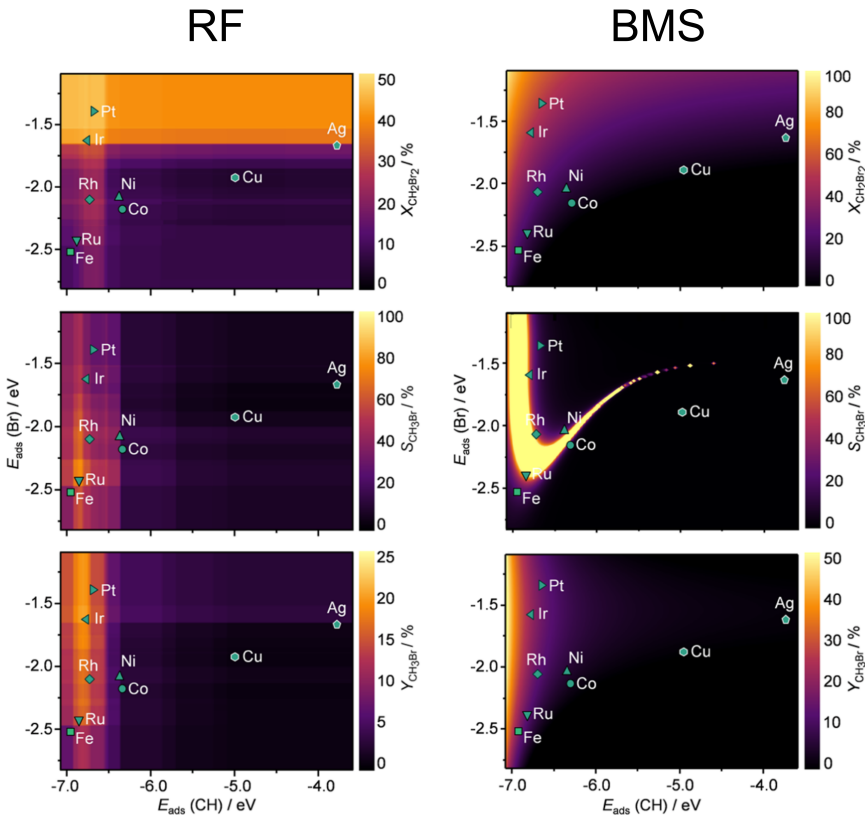


Figure 7.2: Predicted experimental values for conversion (X_{CH_2Br}), selectivity (S_{CH_3Br}) and yield (Y_{CH_3Br}) obtained with a trained RF model (left) and a trained BMS model (right). Both models were trained by the use of the $E_{ads}(Br)$ and $E_{ads}(CH)$ as input values. Adapted from Ref.[181].

system with our descriptors. We explored a diverse range of equations from which we selected the ones presenting the best accuracy/complexity ratios:

$$X_{CH_2Br_2} = -(E_{ads}(Br) - c_{c_1}) \cdot c_{c_2} + \frac{c_{c_3}}{E_{ads}(CH) + c_{c_4}} \tag{7.1}$$

$$\begin{aligned} S_{CH_3Br} = &- ((E_{ads}(Br) - c_{s_1}) \cdot (E_{ads}(CH) + c_{s_3})^{E_{ads}(CH)+c_{s_3}+1} \\ &+ (E_{ads}(CH) + c_{s_3} + c_{s_2}))^{-2} \end{aligned} \tag{7.2}$$

$$Y_{CH_3Br} = -(E_{ads}(Br) + c_{y_1})^2 \cdot c_{y_2} + \frac{c_{y_2}}{(E_{ads}(CH) + c_{y_3})^2} \tag{7.3}$$

Eq. 7.1-7.3 show the selected functional forms, while **Figure 7.2**, right, maps the area defined by the functions for a range of E_{ads}. For conversion and yield the areas have a similar shape than the ones obtained using the RF model, contrary to selectivity, which presents a cliff area. **Table 7.1** shows the difference between SSE errors found for the predictions over the training set using both RF and BMS. We found that BMS outperforms RF in all categories.

Table 7.1: Comparison between the prediction error among the training set for the Random Forest algorithm and the Bayesian Machine Scientist. Iron was excluded from the training set due to its identification outlier. Adapted from Ref.[181].

Observable	SSE_{RF}	SSE_{BMS}
$X_{CH_2Br_2}$	260	98
S_{CH_3Br}	1347	271
Y_{CH_3Br}	42	16

7.4 Generalization to Reaction Families

We demonstrated that the PCA+BMS methodology is able to identify descriptors of a given reaction network while linking them with experimental activities. However, testing different reactions from a given family of compounds is a common process when exploring chemical space, and thus extension of the predictions is a powerful tool to acquire knowledge about the complete chemical space. Yet, extending to reaction networks adds a new dimension to the problem, requiring fine-tuning to be correctly deployed. We benchmarked these capabilities by extending the PCA+BMS methodology to generalize the CH_2X_2 hydrodehalogenation reaction family for $X \in \{F, Cl, Br, I\}$ on eight unique metallic surfaces (Ni, Ru, Rh, Co, Cu, Ag, Ir, Pt) with the purpose of identify common DFT descriptors and

bound them with experimental activities: conversion $X_{CH_2X_2}$, selectivity to different products S_i, yield to different products Y_i and rate $r_{CH_2X_2}$.

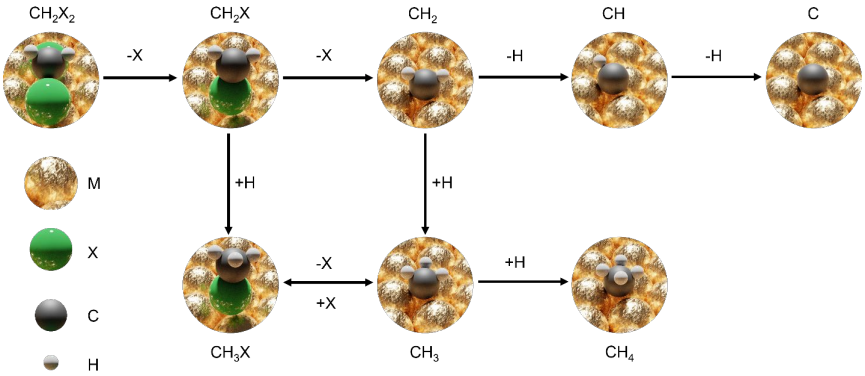


Figure 7.3: Scheme of the CH_2X_2 hydrodebromination network. Adapted from Ref.[81].

It is important to remark some limitations of this procedure; from a practical point of view dihalomethanes with $X \in \{F, I\}$ present handling issues: HF formation and high boiling point of CH_2I_2 respectively, and consequently, we were unable to obtain experimental activities for their reaction networks. Accordingly, we decided to proceed using an hybrid approach, we performed the PCA over the complete species found in reaction family and the predictive approach (SL methods) using a reduced $\{CH_2X_2 : X \in \{Cl, Br\}\}$ reaction set.

7.4.1 Reaction Families Descriptors

To compare the scores of the principal components and their distribution with their analogous obtained for the single CH_2Br_2 hydrodebromination reaction, we performed a PCA including the complete set of binding energies of the intermediates found in our hydrodehalogenation family ($\{CH_2X_2 : X \in \{F, Cl, Br, I\}\}$) adsorbed on each of the metallic surfaces previously mentioned. **Figure 7.4** depicts the obtained results. The three significant principal components cover 89.9%, 7.6% and 1.6% of the total variance. Therefore, when expanding the size of the sampling space by a factor of four, we observe that the contributions are robust but the relative weight between first and the second component rebalance sightly (89.9% and 7.6%, with respect to 92.8% and 5.2%). As with the $X = Br$ set, PCA highlight CH and halogen binding energies as the best descriptors for the first and second components respectively. Notice that the contribution of the halogen species to the two main components follow the trends according to their position in

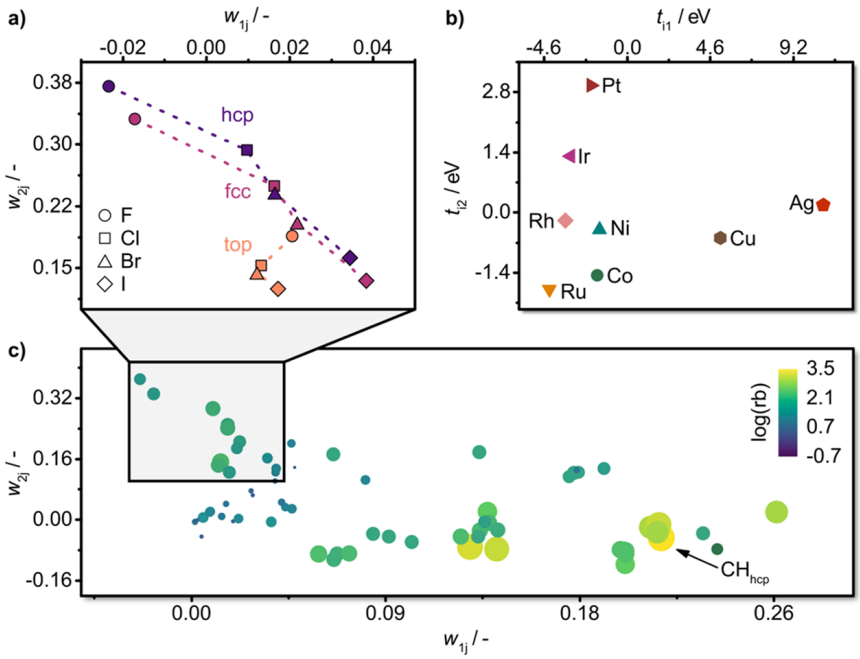


Figure 7.4: Contribution to the main component of a) halogen species, b) metal surfaces, c) included intermediates. Adapted from Ref.[81].

the periodic table (**Figure 7.4a**), exposing additional clues about their link with covalency and redox properties.

7.4.2 Bayesian Learning to find Generalized Equations

With the descriptors identified, we applied the BMS methodology as described for hydrodehalogenation reaction. In this case, we included the experimental observables and the DFT values obtained for the CH_2X_2 hydrodehalogenation with $X \in \{Br, Cl\}$. Due to the technical difficulties clarified earlier, we were unable to include the F and I species.

Before trying to obtain additional equation forms, we benchmarked the equations found for $X = Br$ with the CH_2Cl_2 hydrodechlorination by fitting their constants to the activity values of the reaction. We found that although we do estimate these values correctly, the differences in the fitting constants remarkable change the shape of the predicted surface (**Figure 7.5a-f**). The latter presenting an artificial compression along the x-axis, corresponding to the CH adsorption.

Thus, the different $\boldsymbol{w}$ values obtained for halogens are not enough for the Br equations to be representative of Cl properties as the halogen presence modifies the CH binding. The adsorbed halogens drag intensity from the

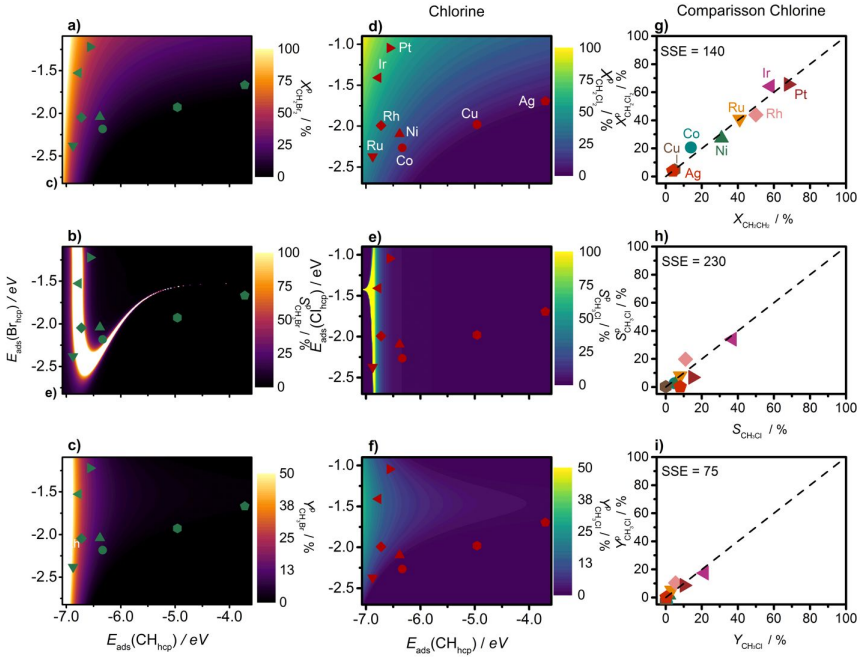


Figure 7.5: Predictions obtained after fitting the constants of the prediction functions obtained for the CH_2Br_2 hydrodebromination to CH_2Cl_2 hydrodechlorination. Adapted from Ref.[81].

surface and modify the Work Function (WF) of the metals, which severely affects covalent contributions to the adsorption energy of coadsorbates. To include this phenomenon, we re-scaled the $\boldsymbol{w}$ parameters corresponding to the CH fragments. The difference between the metal-only WF and the electron affinity, E_{ea} of the isolated halogen atom is taken as a proxy ω_1 for the penalty of the CH fragment adsorption energy, leading to $\omega_1 = E_{ads}(CH_{hcp}) + WF + E_{ea}$ and hence allowing an alignment of the energies of different halogens. After this alignment, we stream the new proxy values ω_1 and $\omega_2 = E_{ads}(CH) + c_3$ with their respective activity pairs (one *per* unique run) to the BMS, finally obtaining a new set of generalized equations:

$$X_{CH_2Br_2} = -\omega_2 c_1 + \frac{c_2^2}{w_1} \quad SSE = 260 \tag{7.4}$$

$$S_{CH_4} = \left[\omega_2 + \omega_1 + \omega_2\left(\omega_2 - \frac{1}{c_1}\right)(\omega_1 + c_2)^2\right]^2 \quad SSE = 284 \tag{7.5}$$

$$S_{coke} = \frac{\omega_2 c_1}{\omega_2 c_2 - \omega_1}(\omega_2 + c_2)\left[\omega_2 + c_2 + \frac{c_2}{\omega_1 cos\left(\frac{\omega_2^2}{c_2}\right)}\right] \quad SSE = 510 \tag{7.6}$$

$$S_{CH_3X} = 100 - S_{coke} - S_{CH_4} \quad SSE = 856 \tag{7.7}$$

$$Y_{i\in\{CH_3X,coke,CH_4\}} = \frac{X_{CH_2X_2} S_i}{100} \quad SSE = \{95, 193, 119\} \tag{7.8}$$

$$r_{CH_2X_2} = \left(\frac{\omega_1 + c_2\omega_2}{c_1}\right)^2 + \omega_1 \quad SSE = 1687 \tag{7.9}$$

Eq. 7.7 presents the generalized equation for CH_3X product selectivity. This equations was not directly obtained though the BMS but rather deduced from the other possible selectivity equations (S_{coke} and S_{CH_4}). A direct attempt to predict S_{CH_3X} leads to complex and difficult to interpret functional forms. Yet, it can be obtained from the carbon balance. The main contribution to its SSE value comes from the error propagation of the S_{coke} and S_{CH_4}. A similar approach is present in **Eq. 7.8**, where $Y_{i\in\{CH_3X,coke,CH_4\}}$ is deduced from $X_{CH_2Br_2}$ and S_i.

7.4.3 Comparison Between Statistical Learning Methods

Figures 7.6-7.7 show the predicted values using BMS, RF and Gaussian Regressor (GR) [218] methods for all the experimental observables. RF and GR divide the CH_2X_2 conversion, product selectivity, yield and rate surfaces qualitatively analogous to BMS. Excluding conversion, RF is able to draw the frontiers of the regions with similar ω_2, although its interpretability is tough as no equation is provided. GR is even less interpretable, as discerning

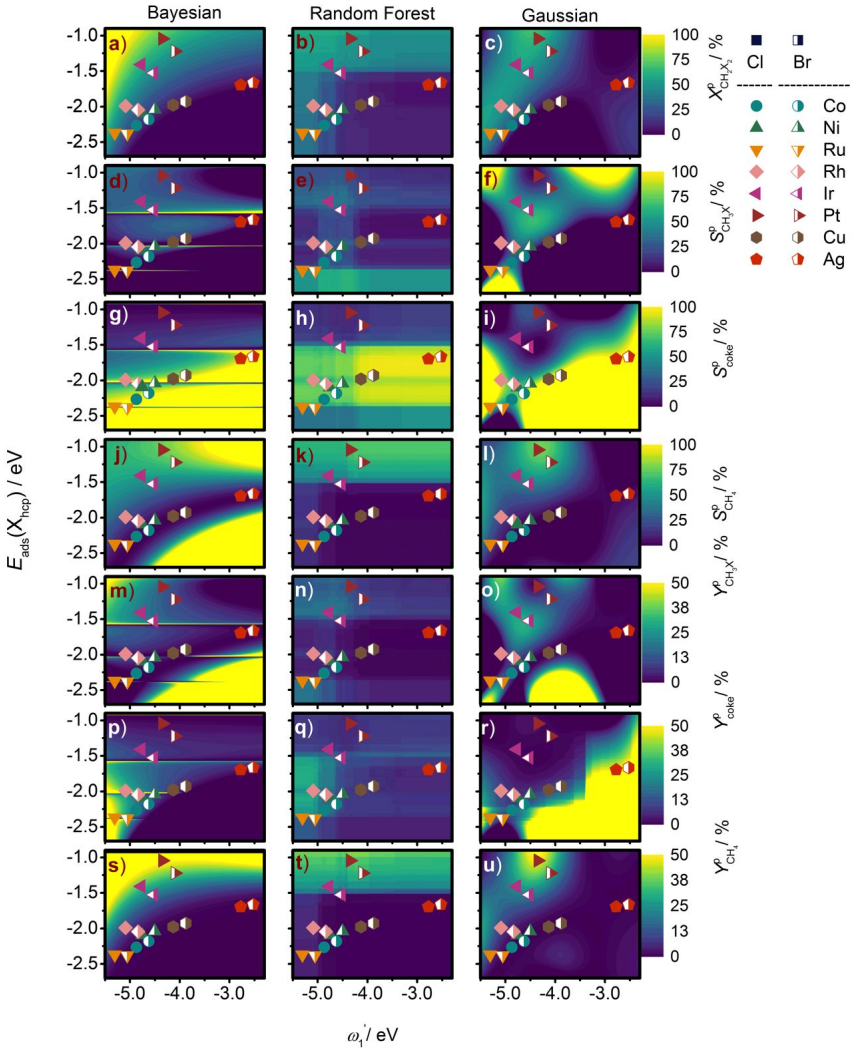


Figure 7.6: Prediction surfaces for the values of $X_{CH_2X_2}$, S_i and Y_i for $i \in \{CH_4, coke, CH_3X\}$ obtained using BMS, RF and GR methods. Adapted from Ref.[81].

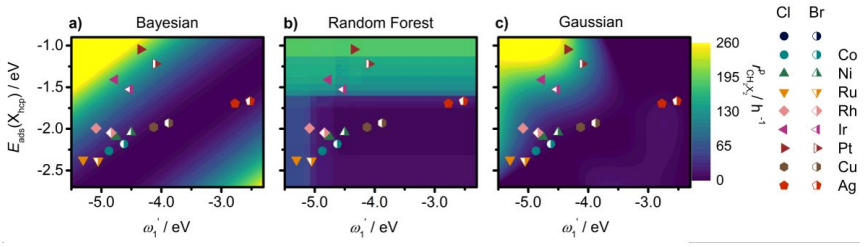


Figure 7.7: Prediction surfaces for the value of $r_{CH_2X_2}$ obtained using BMS, RF and methods. Adapted from Ref.[81].

selectivity with Gaussian functions is confusing due to the sharp nature of selectivity cliffs. Comparison of SSE errors of each method can be found in **Table 7.2**, where BMS presents much better accuracy for the selectivity of CH_3X and coke than RF and GR.

Table 7.2: Comparison between the prediction error among the training set for the BMS, BMS and GR. Adapted from Ref.[81].

Observable	SSE_{BMS}	SSE_{RF}	SSE_{GR}
$X_{CH_2X_2}$	260	629	61
S_{CH_4}	842	811	77
S_{coke}	510	3513	2925
S_{CH_3X}	856	3971	3745
Y_{CH_4}	95	196	20
Y_{coke}	193	378	185
Y_{CH_3X}	119	172	130
$r_{CH_2X_2}$	1687	2924	71

7.5 Conclusions

Classical MK models are no longer able to handle the issues associated with different phase, dynamic rearrangements, site blocking due to poisoning and generally problems related to material gaps. As an alternative, combination of Principal Components Analysis and the Bayesian Machine Scientist has demonstrated to be a outstanding tool to identify descriptors and infer functional forms for limited datasets, bypassing issues that classical methods struggle with. While the accuracy of RF and GR methods is tightly bonded to the number of samples during the training phase, BMS is able to infer an equation even for small datasets. Moreover, the black-box nature of commonly used statistical learning methods made them fall apart when

there is a need of a physical interpretation of the system. BMS is able to handle this interpretability problem by providing functional forms comparable to already found physical models. Although this work is only focused on pristine surfaces, its extension to more complex systems is the next step.

Chapter 8

Analysis of the Carbon Dioxide Electroreduction Network

8.1 Background

Catalysis development for Electroreduction of Carbon Dioxide (eCO_2R) is a prominent field of research due to their capability of converting CO_2 into high value fuels and commodity chemicals.[219, 220] Specifically, copper-based electrocatalysts occupy a pivotal role due to their unique ability to form the C_2 precursors. Structure factors (that vary depending on the preparation) and applied electropotential influence the distribution of the obtained products (**Figure 8.1**).[221–224] Among other products of interest, 1-butanol (C_4), 1-propanol and propylene (C_3) have been reported.[225, 226] However, the last has only been detected as a trace product and 2-propanol has not been found.[178, 227] Formation of C_3 have been reported to require mild potentials (-0.36 to -0.56 V vs. RHE) and asymmetric sites on OD-Cu.[226, 228] Notwithstanding, the reason for the low generation of C_3 products at molecular level is still unknown, as the number of elementary steps involved (10^3) prevents the use of high power-consuming DFT methods to unravel the network reactivity in its completeness.

In this work, we present a novel methodology to model the C_{1-4}, allowing the access and exploration of the reaction network in its completeness. First, we start creating the needed tools to explore and visualize the butanol C_{1-4} network. Then we improved our methodology to apply an hybrid divide and conquer strategy to unravel the most important components and paths involved in the production of propylene and propanol C_3 products.

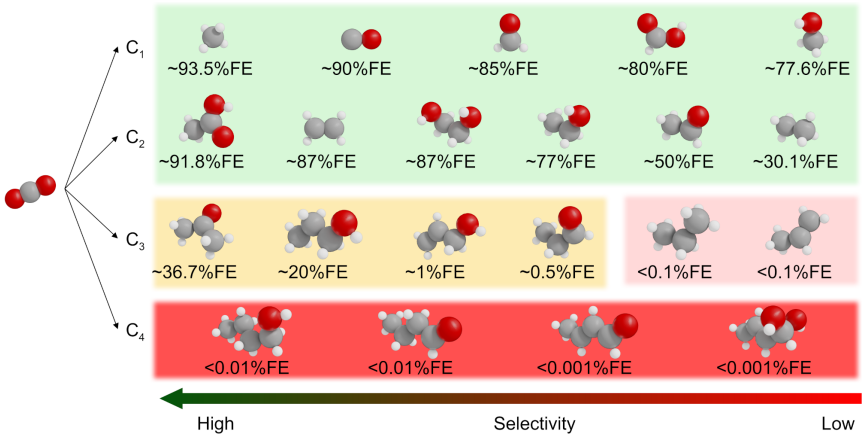


Figure 8.1: List of eCO_2R products reported, depending on their FE: common products (green), uncommon products (yellow), scarce products (pink and red).

8.2 Self-Building Reaction Network

8.2.1 Codifying the intermediates

Exploring large reaction networks is an error prone task; they are composed of a large number of intermediates which in turn are connected to each other. Chemical language does not correspond to machine language, and thus, there is a need to chose a suitable encoding to alleviate the communication between the operator and the virtual network. This step is crucial in terms of interpretability, as the operator demands a quick and smooth understanding when performing exploratory tasks within included reactions. As the final aim of this project was the exploration the eCO_2R, we decided

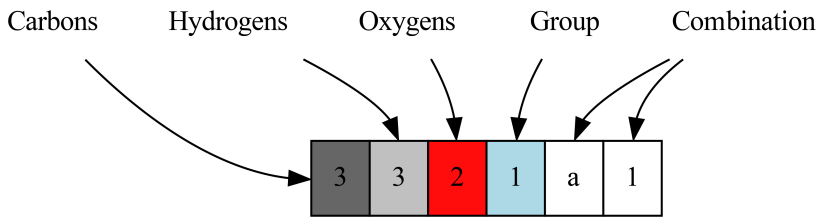


Figure 8.2: Codification of molecular intermediates.

to apply an specific codification depicted in **Figure 8.3**. This representation labels each fragment with six hexadecimal values (to avoid overflow), storing chemical information of the molecule: (i) number of carbon atoms,

(ii) number of hydrogen atoms, (iii) number of oxygen atoms, (iv) the generation group and (v-vi) the combination label. Generation group label (iv) serve to identify molecules that were built during different project stages and combination labels (v-vi) allows the differentiation between isomers. This labeling is easy to interpret and generate by a human operator and thus, is preferred above more complex codifications that retain additional structural information.[174]

8.2.2 Building the Intermediates

Using pyRDTP, we implemented a methodical algorithm to build all the possible molecular fragments inside an organic reaction network.

$$getHydrogens : mol \rightarrow [atom] \tag{8.1}$$

$$delAtom : mol \rightarrow [atom] \rightarrow [mol] \tag{8.2}$$

$$genDescendents : mol \rightarrow (mol, [mol]) \tag{8.3}$$

The algorithm works as follows: Given a *getHydrogens* function that returns the unique hydrogen atoms [*atoms*] of a molecule *mol* (**Eq. 8.1**) and a second function *delAtom* that for each atom in [*atom*] found in *mol*, generates a new molecule without this atom (leading to a set of [*mol*]) (**Eq. 8.2**), they are combined in a *genDescendents* function that given a molecule *mol* returns a pair composed by the molecule (*id*) and a set of unique molecules [*mol*], each with one hydrogen less than the parent molecule (**Eq. 8.3**). This pair structure have two purposes: (i) obtain an offspring generation composed by all the possible 1-less-hydrogen combinations and (ii) as the parent and the offspring are in the same pair, store connectivity information of (de)hydrogenation reactions. This procedure is systematically applied to the generated offspring until a fully dehydrogenated family is obtained.

We used this methodology to generate the complete butanol network, starting from fully saturated C_{1-4} alcohols and hydrocarbons and assuring the integrity of the network. The starting molecules were first relaxed using DFT, allowing to use their three-dimensional forms to generate the *ansatz* of their descendants by the simple deletion of an hydrogen from the parent structure, making them a perfect candidates for the initial guess of DFT relaxations.

8.2.3 Molecular Graphs

The previous automation mechanism presents a serious drawback: chemically identical isomers may appear. Yet graph isomorphism is able to discern between equivalent molecular structures, the graph representation of

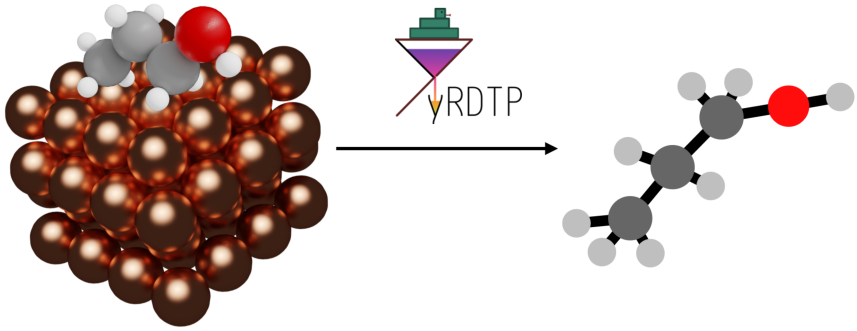


Figure 8.3: Three dimensional representation of n-propanol over a copper surface (left) and its graph representation (right).

the molecule is needed to perform the test. Thus, we implemented an algorithm in pyRDTP to obtain the graph form of an arbitrary molecule (**Figure 8.3**). Hence, we applied isomorphism checks as implemented in networkx [180] over the entire set of (graph form) intermediates, dumping chemically equivalent structures. **Figure 8.4** shows an example of a curated offspring using n-propanol as the primordial structure. Note that group labeling only include unique structures (fragments are labeled after the elimination). After the filtering (ensuring chemically unique structures), we computed the DFT relaxation energies for all the moieties included in the network on a pristine metallic Cu(100) slab.

8.2.4 Integrating Dehydrogenations

While generating the intermediates, we stored the connectivity between parent molecules and the generated offspring (**Eq. 8.3**). This allows the generation of a network graph that includes the dehydrogenation reactions as TS nodes, where the DFT energies can be stored as weights of the intermediate nodes (see **Section 5.2**). As all the included dehydrogenations are elementary steps, they inevitably undergo the following general formula:

$$\forall x, y, z \in \mathbb{N} \quad C_x H_y O_z \longrightarrow C_x H_{y-1} O_z + \frac{1}{2} H_2 \tag{8.4}$$

LSR [229] can be applied to estimate the energy of their associated transition states. With this simple, yet effective, methodology, visual representations of the dehydrogenation network (or sub-graphs) becomes straightforward; it can be automatically generated piping networkx and graphviz.[179, 180] DFT energies are blended into the visual representation of the graph by the

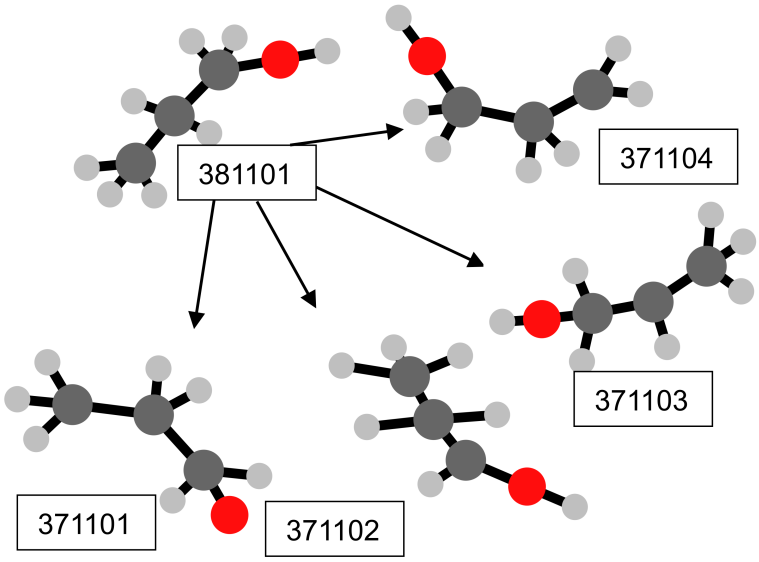


Figure 8.4: Graph representation of the first offspring generation using *n*-propanol as the original *ansatz* and discarding chemically equivalent structures using graph isomorphism.

use of color gradients. An example of a graphical representation generated using this procedure is found in **Figure 8.5**.

8.2.5 Missing Reactions

While dehydrogenations sub-networks are able to partially describe the reactivity of molecular families sharing the same number of O and C atoms, they are meaningless when it comes describing the full reactivity of a network containing a variable number of non-H atoms. As this work aims to describe the C_{1-4} networks associated with butanol and propanol production, the inclusion of these reactions is required to properly describe the network.

Graph representation of the intermediates can be used to detect the sub-graphs produced by the excision of an arbitrary bond (edge). As molecular graphs contain information about elements inside the nodes, these ruptures can be carefully selected to match the $C-C$ and $C-O$ bond-breaking reactions. Ultimately, sub-graphs produced after the fracture can be matched with intermediates already included in the reaction network *via* graph isomorphism and, if found, linked through the creation of a transition state node (**Figure 8.6**).

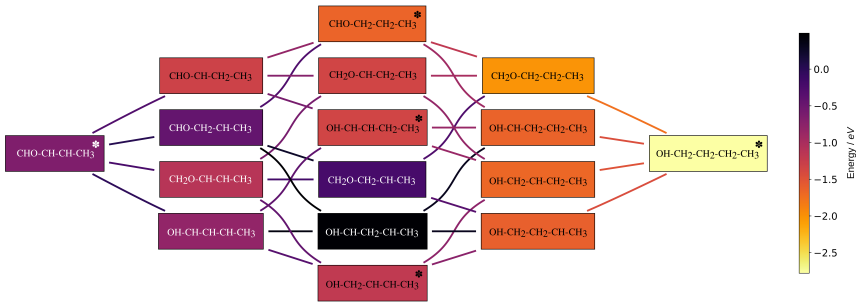


Figure 8.5: Graph representation of the dehydrogenation reactions found starting from the butanal ($CHO - CH_2 - CH_2 - CH_3$). Intermediates are depicted with boxes and their color is associated with their DFT energy, while lines represent the dehydrogenation reactions with color representing their associated LSR energy. Adapted from Ref.[178].

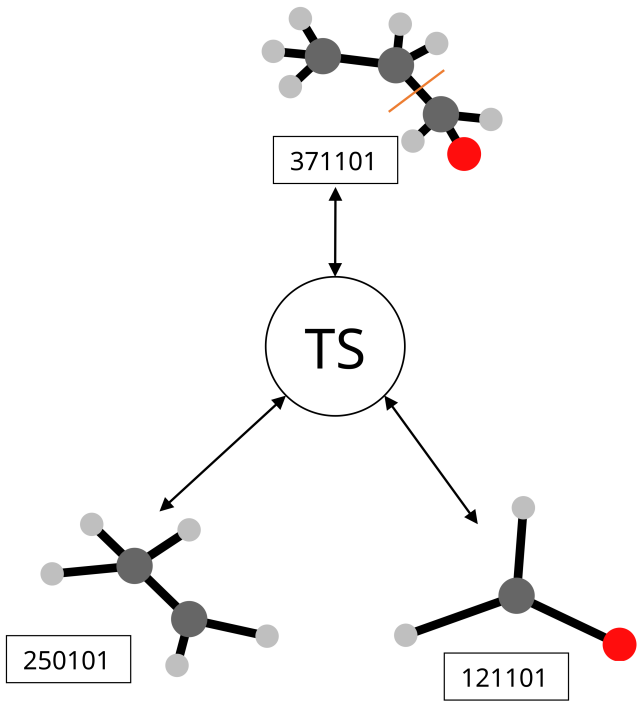


Figure 8.6: $C - C$ bond breaking transition state of butanol ($CH_3 - CH_2 - CH_2O$) to form ethyl ($CH_3 - CH_2$) and formaldehyde (CH_2O) fragments.

Figure 8.7: Representation of the C_{1-4} reaction network. Circles represent, intermediates and lines TS connecting the compounds. Color gradient of both, nodes and lines, is related with their DFT/LSR energy.

8.2.6 Final Remarks

In brief, we defined the complete butanol/propanol C_{1-4} reaction network, composed by >500 intermediates and >3000 TS $C-H$, $C-C$ and $C-O$ TS. **Figure 8.7** shows a representation of the entire network. Tools presented here were used to visualize the butanol eCO_2R reaction network [178], and lately upgraded to deeply explore the propylene/propanol formation from the eCO_2R.

8.3 Exploration of the Propylene/Propanol Formation

We extracted the C_{1-3} network from the global network graph to explore the formation of propylene and n-propanol during the eCO_2R on OD-Cu. Some improvements were needed to understand the network in its completeness.

8.3.1 Including Electrochemical Steps

When building the reaction network, we considered that all hydrogenations occur from adsorbed H^*, in Tafel-like elementary steps. However, as all reactions take place under electrochemical conditions, to strip O and OH groups, one additional step was taken into account. This consideration was

made by explicitly adding new TS nodes to the network containing the Heyrovsky-like steps. As information can be stored in the nodes of the network, using this methodology we were able to give an special treatment to these reactions, changing their E_a according to the applied potential.

$$\forall x, y, z \in \mathbb{N} \quad C_xH_yO_z^* + H^* \longrightarrow C_xH_{y+1}O_z^* + ^* \tag{8.5}$$

$$\forall x, y, z \in \mathbb{N} \quad C_xH_yO_z^* + H^+ + e^- + H^* \longrightarrow C_xH_yO_{z-1}^* + H_2O + ^* \tag{8.6}$$

$$\forall x, y \in \mathbb{N} \quad C_xH_yOH^* + H^+ + e^- \longrightarrow C_xH_y^* + H_2O \tag{8.7}$$

8.3.2 Transition State Search

The energies of the TS included in the network may be approximated by the use of LSR. While they provide a qualitative insight during the exploration phase, LSR are not precise enough when exploring critical paths of the network. Therefore, to extract knowledge of these critical reaction paths (involved in the propanol/propylene formation), multiple DFT searches of TSs were required. Combining the reaction information included in the TS nodes and the structural details contained in the fragment nodes (see **Section 5.2** for additional information of the graph model), we were able to automate the production of the *ansatz* structures of the TS. The algorithm is straightforward: (i) the bond to be broken is detected and the molecule is divided into two moieties, (ii) an arbitrary distance is set between the two fragments, (iii) the structure is relaxed and (iv) a NEB calculation is then performed using the parent already-relaxed structure as the initial state and the two-fragments structure as the final state. Using this procedure we were able ease the obtainment of the DFT energy of unique TS.

8.3.3 High-throughput Filtering

To unravel the possible paths through propylene/propanol, we started using a brute force approach to obtain all the possible $CO_2 \longrightarrow CH_3CHCH_2$ and $CO_2 \longrightarrow CH_3CH_2CH_2OH$ paths to then extract the most feasible ones through the calculation of the E_a of their transition states using LSR. During this path-search we used graph-tool [230, 231] as it provides a high-performance library to work with large graphs. Despite that, the number of possible routes was overwhelming ($> 10^7$) and the computed values of the E_a were similar among the most promising candidates. Henceforth, we decided to adopt a hybrid approach combining available reports, experiments and theory.

We started screening the possible $C_1 - C_2$; we discarded molecular fragments with carboxylate, carboxylic acid, esthers or cyclic backbones as they

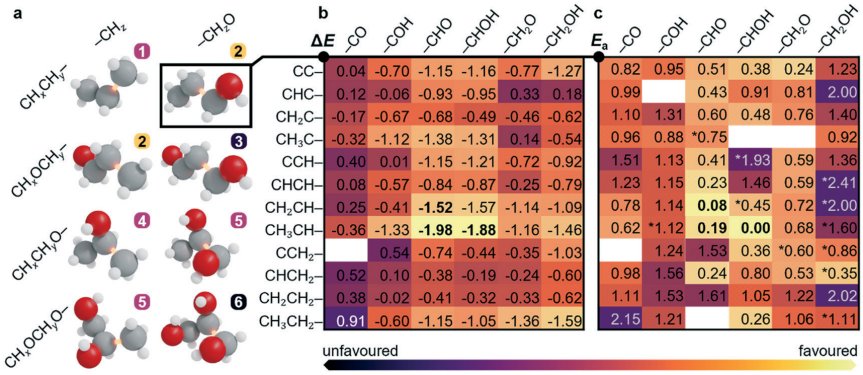


Figure 8.8: a) C_3 families reported in literature for the eCO_2R b) ΔE and E_a DFT values obtained for a set of critical $C_1 - C_2$ C-C couplings. Adapted from Ref.[150].

were not found experimentally in the pool of C_3 products (**Figure 8.8a**). Thus, we considered all the possible couplings emerging from the combination of 10 C_1 ($-CH_xO^*$ and $-CH_x^*$) and 70 C_2 ($CH_yCH_z^*-$) precursors. To filter the best candidates among the couplings, we benchmarked their feasibility by evaluating their thermodynamic, ΔE, and kinetic, E_a, DFT values in combination with experiments tracking the products obtained from singular $C_1 - C_2$ couplings.

8.3.4 Finding the C_3 Precursor

We started discarding the $1,2,3-C_3O_3H_x$ backbone due to its unfavorable kinetic values to form the CO-trimer and its impossibility to go through an alternative dimerisation path (CO then $OCCO^-$ then $COCOCHO$) as no glycerol product is reported on literature. $1,2-C_3O_2H_x$ and $1,3-C_3O_2H_x$ were rejected due to the low stability of their C_2 reactants and be unfavored under eCO_2R conditions respectively. As for the $2-C_3OH_x$ backbone, it expectedly yield to 2-propanol, thermodynamically more stable than *n*-propanol whereas intermediates leading to these two species show similar stabilities, 2-popanol is not found experimentally, provoking the rejection of this backbone. Combining a revision of the literature and our experimental/DFT results, we found that the $1-C_3OH_x$ backbone is the common precursor among the C_3 products, particularly, the CH_2CHCHO^*.

To unravel the routes starting in the common precursor to the desired products, we tracked the possible reaction paths from the $H_2CCHCO*$ common precursor to propanol and propylene (**Figure 8.9**), explicitly computing *via* DFT the TS associated with $C-O(H)$ bond ruptures and some

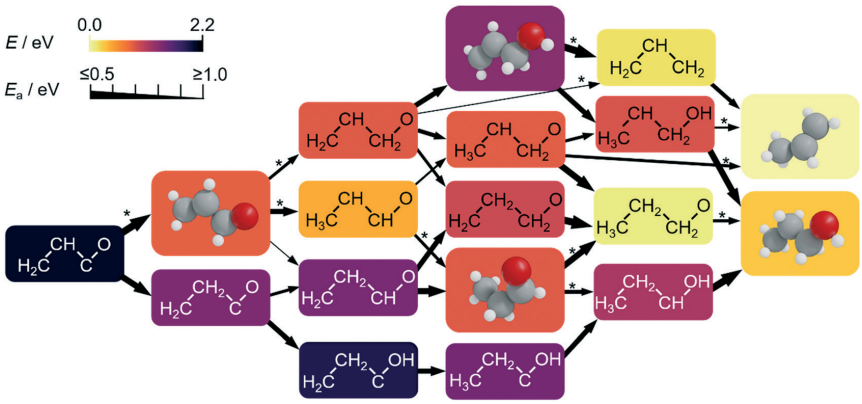


Figure 8.9: Possible reaction routes from $H_2CCHCHO*$ to propylene CH_3CHCH_2 and propanol $CH_3CH_2CH_2OH$. Boxes depicts the intermediates and arrows the transition states. The color of the boxes represent the DFT energy of the intermediates while the width of the arrows represent the E_a of the of the TS. Adapted from Ref.[150].

critical hydrogenations. We found that propionaldehyde CH_3CH_2CHO is the common precursor of propanol while allyl alcohol CH_2CHCH_2OH. Yet, formation of propionaldehyde is highly favored with respect to allyl alcohol production if starting from the common precursor, displacing the overall reaction to 1-propanol and avoiding the formation of propylene. These paths were finally confirmed experimentally through the electroreduction of allyl alcohol and propionaldehyde.

8.4 Conclusions

We were able to create a self-building C_{1-4} network that for a given a set of initial structures it is able to generate all the possible elementary transition states and an *ansatz* of both the fragments and TS composing the network. Moreover, this network has the capability to store arbitrary values ($E_{DFT}(X)$, structures, ...) in their nodes, making its visualization and exploration straightforward.

Once created, we extracted a portion of the reaction graph C_{1-3} to devise the possible routes and mechanisms involving the formation of propylene and propanol products. Combined with experiments, the exploration of the network was able to identify the presence of a common precursor of both propylene and proponal products and to fully describe all the possible reaction routes between the precursor and the products. Finally, our predictions were confirmed by experimental results.

Chapter 9

Convolutional Graph Neural Networks

In the previous chapters I explored the capabilities of molecular and networks graphs and how to build them from a raw molecule or a reaction network. In this chapter, I examine the prediction power of the Convolutional Graph Neural Network (GNN), which use graphs as input values, as well as how they can be used to obtain prediction of different chemical observables.

9.1 Graph Neural Networks

9.1.1 Artificial Neural Network Regressor

Artificial Neural Networks (ANN) are an supervised learning method that aims at learning from a vector $\boldsymbol{X}$ with p components and a target $\boldsymbol{Y}$. Let $\boldsymbol{\omega}_m, m = 1, 2, \ldots, M$ be units p-vectors of unknown parameters. Then, a nonparametric multiple regression Projection Pursuit Regression (PPR) [232] model is defined as in **Eq. 9.1**.[233]

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} g_m(\boldsymbol{\omega}_m^T \boldsymbol{X}) \tag{9.1}$$

$$\sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{M} g_m \left(\boldsymbol{\omega}_m^T \boldsymbol{x}_i \right) \right] \tag{9.2}$$

Where $g_m(\boldsymbol{\omega}_m^T \boldsymbol{X})$ function is called a *ridge function* in $\mathbb{R}^p$, and the model can be trained by approximating the minimizers of an error function. For a set of training data $(x_i, y_i, i = 1, 2, \ldots, N)$ the error function takes the form of **Eq. 9.2**. For a simplified regression model, thus aiming to a single output value ($k = 1$), the target Y_k can be modeled as a function of linear

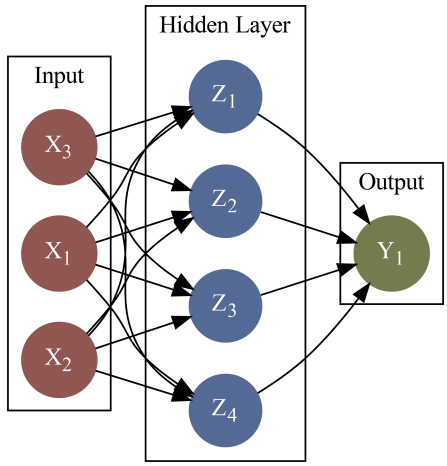


Figure 9.1: Scheme of a simple ANN presenting a vector of features X_p, a hidden layer Z_m and the output value Y_k

combinations of derived features Z_m (hidden layers) that are created from linear combinations of the input (**Eq. 9.3**).

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^T \boldsymbol{X}), m = 1, \ldots, M \\ T_k &= \beta_{0k} + \beta_k^T Z, x = 1, \ldots, K \\ f_k(\boldsymbol{X}) &= g_k(T) = T_k, k = 1, \ldots, K \end{aligned} \tag{9.3}$$

$\sigma(x)$ function is known as the *activation function* and usually corresponds to a sigmoid of the form $\sigma(x) = \frac{1}{1+e^{-x}}$ (**Figure 9.2**). However, modern ANN packages such as Tensorflow or pyTorch [234, 235] implement alternative functions such as the widely used Rectified Linear Unit (ReLU) (**Figure 9.2**) function (**Figure 9.2**).[236] ReLU applies the function $ReLU(x) = max(0, x)$, meaning that it only activates a node of (Z_m) if its value is positive, and then scales linearly.

$$g_m(\boldsymbol{\omega}_m^T \boldsymbol{X}) = \beta_m \sigma(\alpha_{0m} + \|\alpha_m\|(\boldsymbol{\omega}_m^T \boldsymbol{X})) \tag{9.4}$$

Thus, for a single layer neural network the model can be viewed as a PPR model (**Eq. 9.1**) with the ridge function defined as presented in 9.4

$$\begin{aligned} \{\alpha_{0m}, \alpha_m; m = 1, 2, \ldots, M\} M(p+1) \\ \{\beta_{0k}, \beta_k; k = 1, 2, \ldots, K\} K(M+1) \end{aligned} \tag{9.5}$$

To fit a neural network, a set θ of weights is defined (**Eq. 9.5**) then, a *LOSS* Function, for example the SSE (**Eq. 9.6**) is used to compute the error of

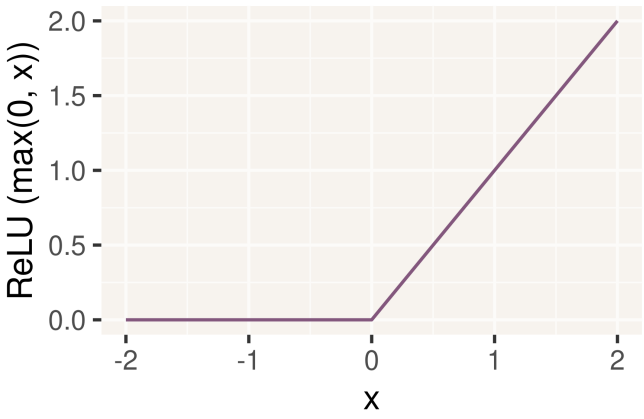


Figure 9.2: The ReLU activation function.

the prediction.

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(\boldsymbol{x}_i))^2 \tag{9.6}$$

A backpropagation algorithm,[237] that usually combines gradient descent and automatic differentiation techniques [238] is then used to adjust the weights to minimize $R(\theta)$, trying to evade global minima, to avoid overfitting towards the training set.

9.1.2 Convolutional Graph Neural Networks

Convolutional Graph Neural Network (GNN) [239, 240] are a particular type of ANN including [241, 242] one or more convolutional graph layers. These layers use a graph structure as input, and they produce an output based on its topology and, in commonly also its weights. Each vertex of the graph is associated with a $\boldsymbol{v}_i$ vector and each vertex with a $\boldsymbol{e}_{(i,j)_k}$ vector both holding weights related to the properties of the graph. The feature vector of each node is updated r times through a $Conv$ function that uses information from its neighbors (**Eq. 9.7**), learning from its environment.

$$\boldsymbol{v}_i^{(t+1)} = Conv(\boldsymbol{v}_i^{(t)}, \boldsymbol{v}_j^{(t)}, \boldsymbol{e}_{(i,j)_k}) \quad (i,j)_k \in G(V,E) \tag{9.7}$$

$$\boldsymbol{v}_g = Pool(\boldsymbol{v}_0^{(0)}, \boldsymbol{v}_1^{(0)}, \ldots, \boldsymbol{v}_n^{(0)}, \ldots, \boldsymbol{v}_n^{(r)}) \tag{9.8}$$

Once the update phase ends, all the feature vectors are directed to a pooling layer $Pool$, creating a complete $\boldsymbol{v}_c$ feature vector for the overall graph. Feature vectors may contain hidden weights $\boldsymbol{W}$ that influence the value of

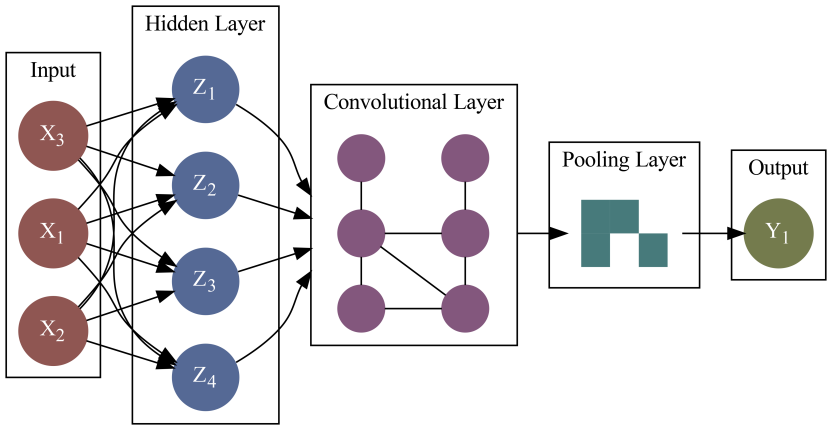


Figure 9.3: Representation of a simple convolutional graph neural network.

the features associated with each node and edge. These weights are updated upon a prediction during the backpropagation phase of the network based on the error of the estimation.

9.2 Applications of GNN in Chemistry

9.2.1 From Functional Groups to Molecular Graphs

As extensively discussed in previous chapters, molecules can be simplified into molecular graphs, condensing information about its atoms and their connections in a simple datum. In organic chemistry, environmental information of a set of atoms can be used to infer the thermodynamical properties of the molecule they form.[243–246] These estimates were made before the popularization of machine learning methods, instead, they were approximated by using a divide-and-conquer strategy, where a molecule is divided into a set of functional groups that lately will add their contributions to the target property to estimate the overall value for the molecule. These contributions were, in fact, tabulated, and they come from fitting experimental results to thermodynamical properties.

$$\forall g_f \in m \quad T(m) = \sum_{i=1}^{N} g_{if} w_f + c_m \tag{9.9}$$

Eq. 9.9 shows a representation of the general formula proposed by Benson et al. [245] to estimate a thermodynamic property T for a given molecule m of N atoms. Let g_f be a set of atoms that forms the functional group f

included in m, w_f the weight (traditionally tabulated) associated with the functional group and c_m a series of corrections based on some properties of the molecule, (e. g.: optical activity, symmetry, ...). Then, T can be obtained from the sum of the functional groups found in m pondered by their associated weights, plus an additional correction. This equation is similar to the general PPR equation described for an artificial neural network (**Eq. 9.1**).

With this approach, functional groups act as the building block unit of the model. However, GNN allows to reduce the functional group to bare atoms by including their environments (bonds). Also, it allows to obtain information about the whole molecule (c_m in **Eq. 9.9**) by applying a convolutional graph layer (**Eq. 9.7-9.8**). Thus, training a GNN network with molecular graphs potentially speeds the learning process and the accuracy of the predictions toward the target properties as it is able to decompose the functional groups in smaller units while getting closer to the chemical concepts.

9.2.2 Applications

Using the previous approach, GNNs have been successfully applied in computational chemistry to predict DFT energy values of gas-phase molecules with exceptional performance. [247] For solid-state chemistry, specific convolutional graph layers have been successfully developed to estimate properties of bulk metals allowing the inclusion of periodicity.[248] Extending to surface chemistry, graph representations have been used to automatically obtain all the possible adsorption combinations of multiple bonding adsorbants on a surface [72], and convolutional networks demonstrated to be an excellent method to predict DFT binding energies in metallic surfaces.[249, 250]. Notwithstanding, GNN is an active field of research and new GNN architectures considering additional chemical properties,[251] improved graph models [252] and additional prediction targets [253] are emerging.

9.3 Proposed use of GNN to Predict DFT energy of High-Order Organic Molecules

9.3.1 Background

Exploring the catalytic behavior of hydrocarbon decomposition on different metals is a time-consuming task due to the amount of species that might appear in a single process and the size of the evaluated molecules.[254] High-order hydrocarbon not only increase the DFT computation time but also require special treatment to be properly modeled, for example extension of

the unit cell. For fuel catalysis and plastic decomposition reactions, these computations are required, and thus, they become a serious bottleneck in catalysis research.

ANN accuracy shares the data size dependence with the other supervised machine learning methods, and thus, they need to be trained with large datasets storing high quality data. In this work we developed the needed framework to create such datasets, as we were able to generate different sets of molecular families and automate their DFT calculation, easing the generation of the input/output pairs.

Hence, we propose the use of a *atomic-splitting* strategy, training a GNN with a full-featured dataset including the most prominent functional groups found in hydrocarbons of industrial interest. Precisely, this dataset should contain geometries (to be converted into graphs) and their E_{DFT} including a diverse set of functional groups adsorbed on a representative set of metallic surfaces. The adsorbant/surface pair and the DFT energies may then be used as the input/target pairs to train a GNN to learn the environmental properties of their functional groups and backbones. This strategy follows the same basis as the one proposed by Benson, but reducing the building block and increasing the learning speed through the use of modern computational techniques.

9.3.2 Proposed Molecular Model and Network Architecture

The proposed graph model to be used as input of the GNN is depicted in **Figure 9.4**. As the target for prediction is the binding energy, only the atoms of the surface that directly bond with the molecule are included.

$$E(X) = E_{DFT}(X^*) - E_{DFT}(^*) \tag{9.10}$$

During the training process, the energy of an adsorbed molecule used as input value $E(X)$ is calculated as in **Eq. 9.10**, where $E_{DFT}(X^*)$ and $E_{DFT}(^*)$ are the DFT energies of the adsorbed molecule and the surface respectively. To avoid the inclusion of the complete surface (including PBC) surface atoms included in the GNN graph input are not considered in terms of energy. The simplified model retains all the information of the molecule while eases the training phase of the GNN.

The architecture of the proposed GNN is shown in **Figure 9.5**, it has been built with Geometric pyTorch. [255] It consists in two linear layers, activated by a ReLU function followed by a GCNConv GNN layer.[248] The output is then feed to a recurrent GRU layer linked to 2 additional GCNConv layers looped three times. Finally, a set2set pooling layer [256] is applied and a final linear layer reduces the output values to a single output value.

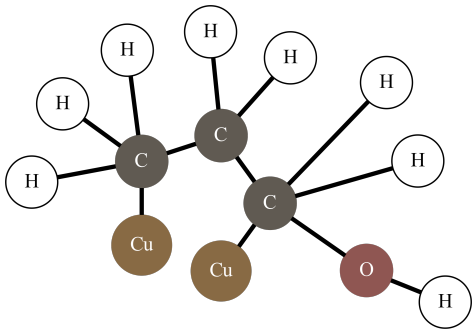


Figure 9.4: Example of the proposed molecular model. Propylene molecular graph adsorbed on a Copper surface through two bonds C1-Cu and C2-Cu of two different copper atoms.

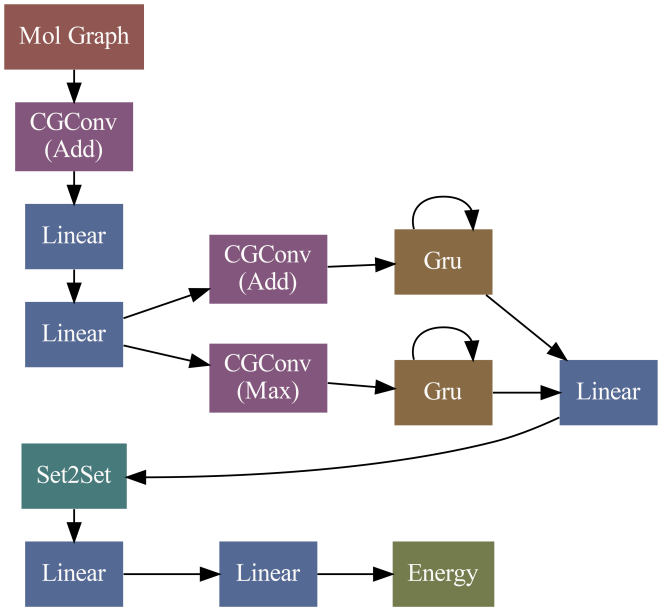


Figure 9.5: Proposed architecture for the GNN, composed by linear layers (blue), graph convolutional layers (purple) GRU layers (ocre) and a pooling layer (teal).

ADAM [257] is used as the optimization algorithm and SSE is used as the LOSS function due to its robustness for regressions.

9.4 Testing

Santiago M. has been building a dataset composed by molecules containing common functional groups: aromatic molecules, alkanes, alcohols, thiols, amines, imines, carbonates, carboxylic acids, esters, amides, oximes and amidines. The molecules have been relaxed on a set of metallic surfaces (Ag, Au, Cd, Cu, Ir, Ni, Os, Pd, Pt, Rh, Ru and Zn), building a dataset of ≈ 2500 molecular structures adsorbed on a surface. The relaxed structures and their DFT energies have been converted into the graph model previously presented and fed to the GNN aiming to predict their energetic DFT values. With a random split of 60/20/20 % (training/validation/test) the Mean Absolute Error (MAE) obtained for the test set is *circa* 0.23 eV. Still, the model is under development and further improvements are needed to enhance its precision (modifications in its architecture of the model and optimization of hyperparameters), being the final aim the prediction of the DFT energy of a set of macromolecules adsorbed on the same set of metals.

Chapter 10

Conclusions and Outlook

This thesis shows the use of different computational tools to handle common problems present in computational heterogeneous catalysis (**Chapter 1**). For this purpose, I combined diverse techniques leading to the development of a complete framework to deal with the high complexity that inhabits catalytic systems (**Part I**). This framework consists in the traditional DFT procedures (**Chapter 2**), automation tools (**Section 3**), statistical learning techniques (**Chapter 4**) and applied graph theory (**Chapter 5**). I started with the deployment of an automated DFT workflow to study the reactivity of pristine catalytic surfaces (**Chapter 6**). I employed this workflow to generate a complete DFT data of the CH_2Br_2 hydrodebromination reaction network. The dataset was used to find the descriptors of the system and find and generalize output equations through statistical learning techniques (**Chapter 7**). Using applied graph theory, I improved the automation framework to extend its usability to complex reaction networks, allowing the exploration of the eCO_2R reaction network (**Chapter 8**). Finally, taking advantage of the improved framework, I proposed a GNN to predict binding energies of large organic molecules by training the network with group fragments (**Chapter 9**).

- In **Chapter 6** I combined an automation framework with my own software (pyRDTP) to create a complete framework able to automate the calculation of ionic relaxations and transition state searches over a set of pristine metallic surfaces. Moreover, this section explores the integration of ioChem-BD in the automation workflow to store and catalog the data in a freely accessible database. The complete automation is tested with the CH_2Br_2 hydrodebromination network, allowing to explore the issues and challenges of the framework.

- In **Chapter 7** I apply dimensionality reduction techniques (PCA) to

extract chemical descriptors from the CH_2Br_2 hydrodebromination reaction over a set of pristine metallic surfaces. These descriptors were used to train a RF model that connects the obtained DFT values with experimental activity. Due to the intrinsic limitations of RF and the small size of the dataset, BMS is applied to infer the equations connecting the descriptors of the system and the experimental values. To generalize the inferred formulas, the studied reaction is expanded to the $CH_2X_2 : X \in \{F, Cl, Br, I\}$ reactions. During the generalization I found that the formulas inferred for the CH_2Br_2 are not able to describe the entire family. Thus, a term including the influence of the halogen on the covalent interactions is added to the formula. Using this correction the equations are inferred again, obtaining a new set of generalized activity equations for Br and Cl. Finally, these equations are benchmarked against classical Microkinetic models and other statistical learning models, specifically, Random Forest and Gaussian Regressor.

- In **Chapter 8** I improved the automation framework to implement the automatic generation of reaction networks *via* the use of graph theory. This new feature is exploited to automate the generation, calculation and analysis of the eCO_2R reaction network on Oxide Derived Copper. After gathering all the data, a divide-and-conquer strategy is used to unravel the critical paths from CO_2 to the studied C_3 products (propylene and 1-propanol). This strategy uses reported data, experiments and DFT evaluations to identify the critical paths of the formation of a common C_3 precursor, refining (*via* DFT) the network paths considered critical. After applying this procedure, CH_2CHCHO^* was identified as the common precursor of both molecules and allyl alcohol and propionaldehyde as key intermediates present in the reaction route. Finally, experiments confirmed our hypothesis.

- In **Chapter 9** I take advantage from the automation framework developed during the previous chapter to create and propose a model and a GNN architecture to predict the binding energy of complex hydrocarbons. The proposed technique exploits the capability of converting molecules into graphs to train the GNN network with a simplified model consisting in the connectivity of the molecule and the elements of their atoms.

Combination of automation, big data and statistical learning presents the current state-of-the-art to study heterogeneous catalysis systems. Still, this set of tools exhibits further limitations that need to be assessed to

assure its robustness: (i) extracting kinetic information from catalytic systems remains a challenge due to the difficult of automating the generation of transition state *ansatz* structures using procedural algorithms, (ii) while data handling platforms improve the accessibility to scientific data, users are responsible of the quality of control of uploaded data, being prone to contain incomplete or low quality datasets and (iii) scarcity of homogenized chemical data restricts the use of unsupervised and supervised statistical learning tools to predict or extract information from catalytic systems. Therefore, and taking into account the problem associated with public databases mentioned above, applying these studies to chemical systems with scarce data is sometimes impossible, and alternative techniques such as Bayesian methods must be used. Finally, the graph representation of molecules and reaction networks seems to be the next logical step in the study of catalysts, since it is able to condense all the complexity of a chemical system into a well-defined data structure, prone to be automated and analyzed.

The methodologies presented in this work are receiving an increasing acceptance by the scientific community, which is making a great effort to improve both the algorithms used and the quality of the available databases to mitigate their drawbacks. This gives rise to a new perspective, in which large and complex chemical systems will be analyzed with ease due to the support given by the SL methods and the trust given by FAIR data platforms.

Bibliography

(1) Chorkendorff, I.; Niemantsverdriet, J. W., *Concepts of Modern Catalysis and Kinetics*; Wiley: Weinheim, 2003, DOI: 10.1002/3527602658.

(2) Gandhi, H.; Graham, G.; McCabe, R. *J. Catal.* **2003**, *216*, 433–442, DOI: 10.1016/s0021-9517(02)00067-2.

(3) Heck, R. M.; Farrauto, R. J.; Gulati, S. T., *Catalytic Air Pollution Control*; John Wiley & Sons, Inc.: 2009, DOI: 10.1002/9781118397749.

(4) Mullins, D. R. *Surface Science Reports* **2015**, *70*, DOI: 10.1016/j.surfrep.2014.12.001.

(5) Li, Q.; García-Muelas, R.; López, N. *Nat. Commun.* **2018**, *9*, 526, DOI: 10.1038/s41467-018-02884-y.

(6) Ahmadi, M.; Mistry, H.; Roldan Cuenya, B. *J. Phys. Chem. Lett.* **2016**, *7*, 3519–3533, DOI: 10.1021/acs.jpclett.6b01198.

(7) Almora-Barrios, N.; Novell-Leruth, G.; Whiting, P.; Liz-Marzán, L. M.; López, N. *Nano Lett.* **2014**, *14*, 871–875, DOI: 10.1021/nl404661u.

(8) García-Muelas, R.; Dattila, F.; Shinagawa, T.; Martín, A. J.; Pérez-Ramírez, J.; López, N. *J. Phys. Chem. Lett.* **2018**, *9*, 7153–7159, DOI: 10.1021/acs.jpclett.8b03212.

(9) Mitchell, S.; Michels, N.-L.; Pérez-Ramírez, J. *Chem. Soc. Rev.* **2013**, *42*, 6094, DOI: 10.1039/c3cs60076a.

(10) Hargreaves, J. S. J.; Munnoch, A. L. *Catal. Sci. Technol.* **2013**, *3*, 1165, DOI: 10.1039/c3cy20866d.

(11) Sholl, D. S.; Steckel, J. A., *Density Functional Theory*; John Wiley & Sons, Inc.: Hoboken, N.J, 2009, DOI: 10.1002/9780470447710.

(12) Sabatier, *La Catalyse en chimie organique*; Nouveau Monde: 2013, DOI: 10.14375/np.9782369430186.

(13) Balandin, A. In *Advances in Catalysis*; Advances in Catalysis; Elsevier: 1969, pp 1–210, DOI: 10.1016/s0360-0564(08)60029-2.

(14) Falsig, H.; Hvolbæk, B.; Kristensen, I. S.; Jiang, T.; Bligaard, T.; Christensen, C. H.; Nørskov, J. K. *Angew. Chem. Int. Ed.* **2008**, *47*, 4835–4839, DOI: 10.1002/anie.200801479.

(15) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. *ACS Catal.* **2019**, *9*, 2752–2759, DOI: 10.1021/acscatal.8b04478.

(16) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. *ACS Catal.* **2019**, *9*, 6624–6647, DOI: 10.1021/acscatal.9b01234.

(17) Pablo-García, S.; García-Muelas, R.; Sabadell-Rendón, A.; López, N. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1540, DOI: 10.1002/wcms.1540.

(18) Hammett, L. P. *J. Chem. Phys.* **1936**, *4*, 613–617, DOI: 10.1063/1.1749914.

(19) Hammond, G. S. *JACS* **1955**, *77*, 334–338, DOI: 10.1021/ja01607a027.

(20) Brönsted, J. N.; Pedersen, K. *Zeitschrift für Physikalische Chemie* **1924**, *108U*, 185–235, DOI: 10.1515/zpch-1924-10814.

(21) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1936**, *32*, 1333, DOI: 10.1039/tf9363201333.

(22) Sutton, J. E.; Vlachos, D. G. *Chem. Eng. Sci.* **2015**, *121*, 190–199, DOI: 10.1016/j.ces.2014.09.011.

(23) Nørskov, J.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. *J. Catal.* **2002**, *209*, 275–278, DOI: 10.1006/jcat.2002.3615.

(24) Mazeau, E. J.; Satpute, P.; Blöndal, K.; Goldsmith, C. F.; West, R. H. *ACS Catal.* **2021**, *11*, 7114–7125, DOI: 10.1021/acscatal.0c04100.

(25) Majumdar, P.; Greeley, J. *Phys. Rev. Materials* **2018**, *2*, 045801, DOI: 10.1103/physrevmaterials.2.045801.

(26) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. *J. Catal.* **2015**, *328*, 36–42, DOI: 10.1016/j.jcat.2014.12.033.

(27) Valter, M.; Santos, E. C. d.; Pettersson, L. G. M.; Hellman, A. *ACS Catal.* **2021**, *11*, 3487–3497, DOI: 10.1021/acscatal.0c04186.

(28) Kopač, D.; Huš, M.; Ogrizek, M.; Likozar, B. *Phys. Chem. C* **2017**, *121*, 17941–17949, DOI: 10.1021/acs.jpcc.7b04985.

(29) Rankin, R. B.; Greeley, J. *ACS Catal.* **2012**, *2*, 2664–2672, DOI: 10.1021/cs3003337.

(30) Wu, H.; Sutton, J. E.; Guo, W.; Vlachos, D. G. *Phys. Chem. C* **2019**, *123*, 27097–27104, DOI: 10.1021/acs.jpcc.9b08662.

(31) Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. *Science* **2015**, *350*, 185–189, DOI: 10.1126/science.aab3501.

(32) Batchelor, T. A.; Pedersen, J. K.; Winther, S. H.; Castelli, I. E.; Jacobsen, K. W.; Rossmeisl, J. *Joule* **2019**, *3*, 834–845, DOI: 10.1016/j.joule.2018.12.015.

(33) Busch, M.; Fabrizio, A.; Luber, S.; Hutter, J.; Corminboeuf, C. *Phys. Chem. C* **2018**, *122*, 12404–12412, DOI: 10.1021/acs.jpcc.8b03935.

(34) Fernández, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martínez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. K. *Angew. Chem. Int. Ed.* **2008**, *47*, 4683–4686, DOI: 10.1002/anie.200705739.

(35) Moser, M.; Czekaj, I.; López, N.; Pérez-Ramírez, J. *Angew. Chem. Int. Ed.* **2014**, *53*, DOI: 10.1002/anie.201483371.

(36) Bell, R. P. *Proc. R. Soc. London A - Math Phys. Sci.* **1936**, *154*, 414–429, DOI: 10.1098/rspa.1936.0060.

(37) Evans, M. G.; Polanyi, M. *Trans. Faraday Soc.* **1938**, *34*, 11, DOI: 10.1039/tf9383400011.

(38) Bligaard, T.; Nørskov, J.; Dahl, S.; Matthiesen, J.; Christensen, C.; Sehested, J. *J. Catal.* **2004**, *224*, 206–217, DOI: 10.1016/j.jcat.2004.02.034.

(39) Zaffran, J.; Michel, C.; Delbecq, F.; Sautet, P. *Phys. Chem. C* **2015**, *119*, 12988–12998, DOI: 10.1021/acs.jpcc.5b01703.

(40) Sutton, J. E.; Vlachos, D. G. *ACS Catal.* **2012**, *2*, 1624–1634, DOI: 10.1021/cs3003269.

(41) Loffreda, D.; Delbecq, F.; Vigné, F.; Sautet, P. *Angew. Chem. Int. Ed.* **2009**, *48*, 8978–8980, DOI: 10.1002/anie.200902800.

(42) García-Muelas, R.; Li, Q.; López, N. *ACS Catal.* **2015**, *5*, 1027–1036, DOI: 10.1021/cs501698w.

(43) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. *Nat. Catal.* **2019**, *2*, 659–670, DOI: 10.1038/s41929-019-0298-3.

(44) López, N.; Almora-Barrios, N.; Carchini, G.; Błoński, P.; Bellarosa, L.; García-Muelas, R.; Novell-Leruth, G.; García-Mota, M. *Catal. Sci. Technol.* **2012**, *2*, 2405, DOI: 10.1039/c2cy20384g.

(45) Vlachos, D. G. *AlChE J.* **2012**, *58*, 1314–1325, DOI: 10.1002/aic.13803.

(46) Motagamwala, A. H.; Dumesic, J. A. *Chem. Rev.* **2020**, *121*, 1049–1076, DOI: 10.1021/acs.chemrev.0c00394.

(47) Pérez-Soto, R.; Besora, M.; Maseras, F. *Org. Lett.* **2020**, *22*, 2873–2877, DOI: 10.1021/acs.orglett.0c00367.

(48) Brezny, A. C.; Landis, C. R. *ACS Catal.* **2019**, *9*, 2501–2513, DOI: 10.1021/acscatal.9b00173.

(49) Rebarchik, M.; Bhandari, S.; Kropp, T.; Mavrikakis, M. *ACS Catal.* **2020**, *10*, 9129–9135, DOI: 10.1021/acscatal.0c01642.

(50) Linic, S *J. Catal.* **2003**, *214*, 200–212, DOI: 10.1016/s0021-9517(02)00156-2.

(51) Chatterjee, A.; Vlachos, D. G. *J. Comput.-Aided Mater. Des.* **2007**, *14*, 253–308, DOI: 10.1007/s10820-006-9042-9.

(52) Stamatakis, M.; Vlachos, D. G. *ACS Catal.* **2012**, *2*, 2648–2663, DOI: 10.1021/cs3005709.

(53) Chutia, A.; Thetford, A.; Stamatakis, M.; Catlow, C. R. A. *Phys. Chem. Chem. Phys.* **2020**, *22*, 3620–3632, DOI: 10.1039/c9cp05476f.

(54) Mahlberg, D.; Groß, A. *ChemPhysChem* **2020**, *22*, 29–39, DOI: 10.1002/cphc.202000838.

(55) Reuter, K.; Scheffler, M. *Phys. Rev. B* **2006**, *73*, 045433, DOI: 10.1103/physrevb.73.045433.

(56) Pogodin, S.; López, N. *ACS Catal.* **2014**, *4*, 2328–2332, DOI: 10.1021/cs500414p.

(57) Vorobyeva, E.; Gerken, V. C.; Mitchell, S.; Sabadell-Rendón, A.; Hauert, R.; Xi, S.; Borgna, A.; Klose, D.; Collins, S. M.; Midgley, P. A.; Kepaptsoglou, D. M.; Ramasse, Q. M.; Ruiz-Ferrando, A.; Fako, E.; Ortuño, M. A.; López, N.; Carreira, E. M.; Pérez-Ramírez, J. *ACS Catal.* **2020**, *10*, 11069–11080, DOI: 10.1021/acscatal.0c03164.

(58) Wilkinson, M. D. et al. *Sci. Data* **2016**, *3*, 160018, DOI: 10.1038/sdata.2016.18.

(59) Álvarez-Moreno, M.; de Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. *J. Chem. Inf. Model.* **2014**, *55*, 95–103, DOI: 10.1021/ci500593j.

(60) NoMaD Repository https://nomad-coe.eu (accessed 04/01/2022).

(61) Materials Cloud https://www.materialscloud.org/home (accessed 04/01/2022).

(62) Computational Materials Repository https://cmr.fysik.dtu.dk (accessed 04/01/2022).

(63) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. *Nat. Rev. Mater.* **2018**, *3*, 5–20, DOI: 10.1038/s41578-018-0005-z.

(64) Council of European Union Council regulation (EU) no 2021/695 https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32021R0695 (accessed 04/01/2022).

(65) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013**, *1*, 011002, DOI: 10.1063/1.4812323.

(66) Gromski, P. S.; Granda, J. M.; Cronin, L. *Trends in Chemistry* **2020**, *2*, 4–12, DOI: 10.1016/j.trechm.2019.07.004.

(67) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. *Trends in Chemistry* **2019**, *1*, 282–291, DOI: 10.1016/j.trechm.2019.02.007.

(68) Coley, C. W.; Eyke, N. S.; Jensen, K. F. *Angew. Chem. Int. Ed.* **2020**, *59*, 22858–22893, DOI: 10.1002/anie.201909987.

(69) Coley, C. W.; Eyke, N. S.; Jensen, K. F. *Angew. Chem. Int. Ed.* **2020**, *59*, 23414–23436, DOI: 10.1002/anie.201909989.

(70) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; Gunter, D.; Persson, K. A. *Concurr Comput* **2015**, *27*, 5037–5059, DOI: 10.1002/cpe.3505.

(71) Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. *Nato. Sc. S. Ss. Iii. C. S.* **2016**, *111*, 218–230, DOI: 10.1016/j.commatsci.2015.09.013.

(72) Boes, J. R.; Mamun, O.; Winther, K.; Bligaard, T. *J. Phys. Chem. A* **2019**, *123*, 2281–2285, DOI: 10.1021/acs.jpca.9b00311.

(73) Montoya, J. H.; Persson, K. A. *npj Comput. Mater.* **2017**, *3*, 14, DOI: 10.1038/s41524-017-0017-z.

(74) Tran, K.; Palizhati, A.; Back, S.; Ulissi, Z. W. *J. Chem. Inf. Model.* **2018**, *58*, 2392–2400, DOI: 10.1021/acs.jcim.8b00386.

(75) Kahle, L.; Marcolongo, A.; Marzari, N. *Energ. Environ. Sci.* **2020**, *13*, 928–948, DOI: 10.1039/c9ee02457c.

(76) Pfaendtner, J.; Broadbelt, L. J. *Ind. Eng. Chem. Res.* **2008**, *47*, 2897–2904, DOI: 10.1021/ie071481z.

(77) Rangarajan, S.; Bhan, A.; Daoutidis, P. *Comput. Chem. Eng.* **2012**, *45*, 114–123, DOI: 10.1016/j.compchemeng.2012.06.008.

(78) Goldsmith, C. F.; West, R. H. *Phys. Chem. C* **2017**, *121*, 9970–9981, DOI: 10.1021/acs.jpcc.7b02133.

(79) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. *Chem. Sci.* **2018**, *9*, 825–835, DOI: 10.1039/c7sc03628k.

(80) Vernuccio, S.; Broadbelt, L. J. *AlChE J.* **2019**, *65*, e16663, DOI: 10.1002/aic.16663.

(81) Pablo-García, S.; Sabadell-Rendón, A.; Saadun, A. J.; Morandi, S.; Pérez-Ramírez, J.; López, N. *ACS Catal.* **2022**, *12*, 1581–1594, DOI: 10.1021/acscatal.1c04345.

(82) Wodrich, M. D.; Fabrizio, A.; Meyer, B.; Corminboeuf, C. *Chem. Sci.* **2020**, *11*, 12070–12080, DOI: 10.1039/d0sc04289g.

(83) García-Muelas, R.; López, N. *Nat. Commun.* **2019**, *10*, 4687, DOI: 10.1038/s41467-019-12709-1.

(84) Morán-González, L.; Pedregal, J. R.-G.; Besora, M.; Maseras, F. *Eur. J. Inorg. Chem.* **2021**, *2022*, e202100932, DOI: 10.1002/ejic.202100932.

(85) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. *Nat. Catal.* **2018**, *1*, 531–539, DOI: 10.1038/s41929-018-0094-5.

(86) Artrith, N. *Matter* **2020**, *3*, 985–986, DOI: 10.1016/j.matt.2020.09.012.

(87) Bo, C.; Maseras, F.; López, N. *Nat. Catal.* **2018**, *1*, 809–810, DOI: 10.1038/s41929-018-0176-4.

(88) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. *ACS Catal.* **2019**, *10*, 2260–2297, DOI: 10.1021/acscatal.9b04186.

(89) Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**, *361*, 360–365, DOI: 10.1126/science.aat2663.

(90) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018**, *559*, 547–555, DOI: 10.1038/s41586-018-0337-2.

(91) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. *Sci. Adv.* **2017**, *3*, e1701816, DOI: 10.1126/sciadv.1701816.

(92) Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. *npj Comput. Mater.* **2019**, *5*, 46, DOI: 10.1038/s41524-019-0181-4.

(93) Naik, R. R.; Tiihonen, A.; Thapa, J.; Batali, C.; Liu, Z.; Sun, S.; Buonassisi, T. *CoRR* **2021**, DOI: 10.48550/arXiv.2106.10951.

(94) Computation and Machine Learning for Chemistry https://www.nature.com/collections/gcijejjahe (accessed 04/01/2022).

(95) Schröndinger, E. *Ann. Phys.* **1926**, *384*, 361–376, DOI: 10.1002/andp.19263840404.

(96) Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *389*, 457–484, DOI: 10.1002/andp.19273892002.

(97) Rayner Hartree, D.; Hratree, W *Proc. R. Soc. London A - Math Phys. Sci.* **1935**, *150*, 9–33, DOI: 10.1098/rspa.1935.0085.

(98) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871, DOI: 10.1103/physrev.136.b864.

(99) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138, DOI: 10.1103/physrev.140.a1133.

(100) Ceperley, D. M.; Alder, B. J. *Phys. Rev. Lett.* **1980**, *45*, 566–569, DOI: 10.1103/physrevlett.45.566.

(101) Jones, R. O.; Gunnarsson, O. *Rev. Mod. Phys.* **1989**, *61*, 689–746, DOI: 10.1103/revmodphys.61.689.

(102) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249, DOI: 10.1103/physrevb.45.13244.

(103) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868, DOI: 10.1103/physrevlett.77.3865.

(104) Carrasco, J.; Santra, B.; Klimeš, J.; Michaelides, A. *Phys. Rev. Lett.* **2011**, *106*, 026101, DOI: 10.1103/physrevlett.106.026101.

(105) Błoński, P.; López, N. *Phys. Chem. C* **2012**, *116*, 15484–15492, DOI: 10.1021/jp304542m.

(106) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799, DOI: 10.1002/jcc.20495.

(107) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463–1473, DOI: 10.1002/jcc.20078.

(108) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104, DOI: 10.1063/1.3382344.

(109) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004**, *92*, 246401, DOI: 10.1103/physrevlett.92.246401.

(110) Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. *Phys. Rev. B* **2010**, *82*, 081101, DOI: 10.1103/physrevb.82.081101.

(111) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005, DOI: 10.1103/physrevlett.102.073005.

(112) Almora-Barrios, N.; Carchini, G.; Błoński, P.; López, N. *J. Chem. Theory Comput.* **2014**, *10*, 5002–5009, DOI: 10.1021/ct5006467.

(113) Hibbert, D.; Armstrong, N. *Chemometr. Intell. Lab.* **2009**, *97*, 211–220, DOI: 10.1016/j.chemolab.2009.03.009.

(114) Vanderbilt, D. *Phys. Rev. B* **1990**, *41*, 7892–7895, DOI: 10.1103/physrevb.41.7892.

(115) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*, 17953–17979, DOI: 10.1103/physrevb.50.17953.

(116) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901–9904, DOI: 10.1063/1.1329672.

(117) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978–9985, DOI: 10.1063/1.1323224.

(118) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **1999**, *111*, 7010–7022, DOI: 10.1063/1.480097.

(119) Heyden, A.; Bell, A. T.; Keil, F. J. *J. Chem. Phys.* **2005**, *123*, 224101, DOI: 10.1063/1.2104507.

(120) Kresse, G.; Furthmüller, J. *Nato. Sc. S. Ss. Iii. C. S.* **1996**, *6*, 15–50, DOI: 10.1016/0927-0256(96)00008-0.

(121) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465, DOI: 10.1002/jcc.21759.

(122) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *13*, 5188–5192, DOI: 10.1103/physrevb.13.5188.

(123) Neugebauer, J.; Scheffler, M. *Phys. Rev. B* **1992**, *46*, 16067–16080, DOI: 10.1103/physrevb.46.16067.

(124) Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L.; Kitchin, J. R.; Bligaard, T.; Jónsson, H. *The Journal of Physical Chemistry B* **2004**, *108*, 17886–17892, DOI: 10.1021/jp047349j.

(125) Peterson, A. A.; Abild-Pedersen, F.; Studt, F.; Rossmeisl, J.; Nørskov, J. K. *Energ. Environ. Sci.* **2010**, *3*, 1311, DOI: 10.1039/c0ee00071j.

(126) Moore, G. *Proc. IEEE* **1998**, *86*, 82–85, DOI: 10.1109/jproc.1998.658762.

(127) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

(128) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024, DOI: 10.1088/2632-2153/aba947.

(129) Murray-Rust, P.; Rzepa, H. S. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 757–772, DOI: 10.1021/ci0256541.

(130) Lewis, G. N. *JACS* **1916**, *38*, 762–785, DOI: 10.1021/ja02261a002.

(131) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. *J. Cheminf.* **2012**, *4*, 17, DOI: 10.1186/1758-2946-4-17.

(132) Jmol: An open-source Java viewer for chemical structures in 3D http://www.jmol.org/ (accessed 04/01/2022).

(133) Rego, N.; Koes, D. *Method. Biochem. Anal.* **2014**, *31*, 1322–1324, DOI: 10.1093/bioinformatics/btu829.

(134) Momma, K.; Izumi, F. *J. Appl. Crystallogr.* **2011**, *44*, 1272–1276, DOI: 10.1107/s0021889811038970.

(135) Wickham, H., *ggplot2*; Springer International Publishing: 2016, DOI: 10.1007/978-3-319-24277-4.

(136) Williams, T.; Kelley, C.; et al. Gnuplot 4.6: An interactive plotting program http://gnuplot.sourceforge.net/ (accessed 04/01/2022).

(137) Hunter, J. D. *Comput. Sci. Eng.* **2007**, *9*, 90–95, DOI: 10.1109/mcse.2007.55.

(138) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33, DOI: 10.1186/1758-2946-3-33.

(139) Lurie, S. J.; Winker, M. A., *Reference*; Oxford University Press: Scotts Valley, CA, 2009, DOI: 10.1093/jama/9780195176339.021.273.

(140) Harris, C. R. et al. *Nature* **2020**, *585*, 357–362, DOI: 10.1038/s41586-020-2649-2.

(141) Virtanen, P. et al. *Nat. Methods* **2020**, *17*, 261–272, DOI: 10.1038/s41592-019-0686-2.

(142) Spyder: The Scientific Python Development Environment https://www.spyder-ide.org/ (accessed 04/01/2022).

(143) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. *Chem. Cent. J.* **2008**, *2*, 5, DOI: 10.1186/1752-153x-2-5.

(144) Hjorth Larsen, A. et al. *J. Phys. Condens. Matter* **2017**, *29*, 273002, DOI: 10.1088/1361-648x/aa680e.

(145) Schrödinger, LLC The PyMOL Molecular Graphics System, Version 1.8 https://pymol.org/2/ (accessed 04/01/2022).

(146) Blender Online Community Blender - a 3D modelling and rendering package Blender Foundation, http://www.blender.org (accessed 04/01/2022).

(147) Pablo-García, S. pyRDTP https://gitlab.com/iciq-tcc/nlopez-group/pyrdtp (accessed 04/01/2022).

(148) Bommarito, E.; Bommarito, M. J. *SSRN Electronic Journal* **2019**, DOI: 10.2139/ssrn.3426281.

(149) R Core Team, R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, https://www.R-project.org/ (accessed 04/01/2022).

(150) Pablo-García, S.; Veenstra, F. L. P.; Ting, L. R. L.; García-Muelas, R.; Dattila, F.; Martín, A. J.; Yeo, B. S.; Pérez-Ramírez, J.; López, N. *Catal. Sci. Technol.* **2022**, *12*, 409–417, DOI: 10.1039/d1cy01423d.

(151) McKinney, W. In *Proceedings of the Python in Science Conference*, SciPy: 2010, pp 56–61, DOI: 10.25080/majora-92bf1922-00a.

(152) Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. *Mob. Comput. Commun. Rev.* **2015**, *19*, 29–33, DOI: 10.1145/2786984.2786995.

(153) Tibshirani, R. *J R Stat Soc Series B Stat Methodol* **1996**, *58*, 267–288, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

(154) Santosa, F.; Symes, W. W. *SIAM J. Sci. Stat. Comput* **1986**, *7*, 1307–1330, DOI: 10.1137/0907087.

(155) Ho, T. K. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844, DOI: 10.1109/34.709601.

(156) Rosenblatt, F. *Psychol. Rev.* **1958**, *65*, 386–408, DOI: 10.1037/h0042519.

(157) Steinhaus, H. *Bull. Acad. Pol. Sci., Cl. III* **1957**, *4*, 801–804.

(158) MacQueen, J. Some methods for classification and analysis of multivariate observations, English, Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967). 1967.

(159) Pearson, K. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572, DOI: 10.1080/14786440109462720.

(160) Hinton, G.; Roweis, S. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, MIT Press: Cambridge, MA, USA, 2002, pp 857–864.

(161) Van der Maaten, L.; Hinton, G. *Mach. Learn.* **2011**, *87*, 33–55, DOI: 10.1007/s10994-011-5273-4.

(162) Schmidt, M.; Lipson, H. *Science* **2009**, *324*, 81–85, DOI: 10.1126/science.1165893.

(163) Olivetti de França, F. *Inform. Sciences* **2018**, *442-443*, 18–32, DOI: 10.1016/j.ins.2018.02.040.

(164) Udrescu, S.-M.; Tegmark, M. *Sci. Adv.* **2020**, *6*, eaay2631, DOI: 10.1126/sciadv.aay2631.

(165) Jin, Y.; Fu, W.; Kang, J.; Guo, J.; Guo, J. *CoRR* **2019**, DOI: 10.48550/arXiv.1910.0889208892.

(166) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. *Sci. Adv.* **2020**, *6*, eaav6971, DOI: 10.1126/sciadv.aav6971.

(167) Hastings, W. K. *Biometrika* **1970**, *57*, 97–109, DOI: 10.1093/biomet/57.1.97.

(168) Benfey, O. T. *J. Chem. Educ.* **1958**, *35*, 21, DOI: 10.1021/ed035p21.

(169) International Union of Pure and Applied Chemistry (IUPAC): 2014, DOI: 10.1351/goldbook.mt07069.

(170) Minkin, V. I. *Pure Appl. Chem.* **1999**, *71*, 1919–1981, DOI: 10.1351/pac199971101919.

(171) Bondy, J. A.; Murty, U. S. R., *Graph Theory*; Springer London: 2008, DOI: 10.1007/978-1-84628-970-5.

(172) Balaban, A. T. *J. Chem. Inf. Model.* **1985**, *25*, 334–343, DOI: 10.1021/ci00047a033.

(173) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. *Nat. Commun.* **2017**, *8*, 15679, DOI: 10.1038/ncomms15679.

(174) Pinheiro, G. A.; Mucelini, J.; Soares, M. D.; Prati, R. C.; Da Silva, J. L. F.; Quiles, M. G. *J. Phys. Chem. A* **2020**, *124*, 9854–9866, DOI: 10.1021/acs.jpca.0c05969.

(175) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. *J. Chem. Inf. Model.* **2020**, *60*, 1175–1183, DOI: 10.1021/acs.jcim.9b00943.

(176) Krasnov, L.; Khokhlov, I.; Fedorov, M. V.; Sosnin, S. *Sci. Rep.* **2021**, *11*, 14798, DOI: 10.1038/s41598-021-94082-y.

(177) Eyring, H. *J. Chem. Phys.* **1935**, *3*, 107–115, DOI: 10.1063/1.1749604.

(178) Ting, L. R. L.; García-Muelas, R.; Martín, A. J.; Veenstra, F. L. P.; Chen, S. T.-J.; Peng, Y.; Per, E. Y. X.; Pablo-García, S.; López, N.; Pérez-Ramírez, J.; Yeo, B. S. *Angew. Chem. Int. Ed.* **2020**, *59*, 21072–21079, DOI: 10.1002/anie.202008289.

(179) Ellson, J.; Gansner, E. R.; Koutsofios, E.; North, S. C.; Woodhull, G. In *Graph Drawing Software*, Springer-Verlag: 2003, pp 127–148.

(180) Hagberg, A. A.; Schult, D. A.; Swart, P. J. In *Proceedings of the 7th Python in Science Conference*, ed. by Varoquaux, G.; Vaught, T.; Millman, J., Pasadena, CA USA, 2008, pp 11 –15.

(181) Saadun, A. J.; Pablo-García, S.; Paunović, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J. *ACS Catal.* **2020**, *10*, 6129–6143, DOI: 10.1021/acscatal.0c00679.

(182) Pablo-García, S.; Álvarez-Moreno, M.; López, N. *Int. J. Quantum Chem.* **2020**, *121*, e26382, DOI: 10.1002/qua.26382.

(183) *Information and documentation — Digital object identifier system*; Standard; Geneva, CH: International Organization for Standardization, 2012.

(184) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. *Nat. Chem.* **2021**, *13*, 505–508, DOI: 10.1038/s41557-021-00716-z.

(185) Authorea – Open Research Collaboration and Publishing https://www.authorea.com/ (accessed 04/01/2022).

(186) Saadun, A. J.; Kaiser, S. K.; Ruiz-Ferrando, A.; Pablo-García, S.; Büchele, S.; Fako, E.; López, N.; Pérez-Ramírez, J. *Small* **2021**, *17*, 2005234, DOI: 10.1002/smll.202005234.

(187) McFarland, E. *Science* **2012**, *338*, 340–342, DOI: 10.1126/science.1226840.

(188) Horn, R.; Schlögl, R. *Catal. Lett.* **2014**, *145*, 23–39, DOI: 10.1007/s10562-014-1417-z.

(189) Lunsford, J. H. *Catal. Today* **2000**, *63*, 165–174, DOI: 10.1016/s0920-5861(00)00456-9.

(190) Tang, P.; Zhu, Q.; Wu, Z.; Ma, D. *Energy Environ. Sci.* **2014**, *7*, 2580–2591, DOI: 10.1039/c4ee00604f.

(191) Lin, R.; Amrute, A. P.; Pérez-Ramírez, J. *Chem. Rev.* **2017**, *117*, 4182–4247, DOI: 10.1021/acs.chemrev.6b00551.

(192) Lange, J.-P.; Tijm, P. *Chem. Eng. Sci.* **1996**, *51*, 2379–2387, DOI: 10.1016/0009-2509(96)00094-2.

(193) Zichittella, G.; Paunović, V.; Amrute, A. P.; Pérez-Ramírez, J. *ACS Catal.* **2017**, *7*, 1805–1817, DOI: 10.1021/acscatal.6b03600.

(194) Paunović, V.; Zichittella, G.; Moser, M.; Amrute, A. P.; Pérez-Ramírez, J. *Nat. Chem.* **2016**, *8*, 803–809, DOI: 10.1038/nchem.2522.

(195) Paunović, V.; Lin, R.; Scharfe, M.; Amrute, A. P.; Mitchell, S.; Hauert, R.; Pérez-Ramírez, J. *Angew. Chem. Int. Ed.* **2017**, *56*, 9791–9795, DOI: 10.1002/anie.201704406.

(196) Paul, S. The Method of Direct Hydrogenation By Catalysis, 1912.

(197) Hansen, M. H.; Nørskov, J. K.; Bligaard, T. *J. Catal.* **2019**, *374*, 161–170, DOI: 10.1016/j.jcat.2019.03.034.

(198) Stegelmann, C.; Andreasen, A. *Nature Precedings* **2011**, DOI: 10.1038/npre.2011.6076.2.

(199) Lian, Z.; Ali, S.; Liu, T.; Si, C.; Li, B.; Su, D. S. *ACS Catal.* **2018**, *8*, 4694–4704, DOI: 10.1021/acscatal.8b00107.

(200) Jørgensen, M.; Grönbeck, H. *JACS* **2019**, *141*, 8541–8549, DOI: 10.1021/jacs.9b02132.

(201) Huš, M.; Grilc, M.; Pavlišič, A.; Likozar, B.; Hellman, A. *Catal. Today* **2019**, *338*, 128–140, DOI: 10.1016/j.cattod.2019.05.022.

(202) Singh, S.; Li, S.; Carrasquillo-Flores, R.; Alba-Rubio, A. C.; Dumesic, J. A.; Mavrikakis, M. *AlChE J.* **2014**, *60*, 1303–1319, DOI: 10.1002/aic.14401.

(203) Alexopoulos, K.; Vlachos, D. G. *Chem. Sci.* **2020**, *11*, 1469–1477, DOI: 10.1039/c9sc05944j.

(204) Teschner, D.; Novell-Leruth, G.; Farra, R.; Knop-Gericke, A.; Schlögl, R.; Szentmiklósi, L.; Hevia, M. G.; Soerijanto, H.; Schomäcker, R.; Pérez-Ramírez, J.; López, N. *Nat. Chem.* **2012**, *4*, 739–745, DOI: 10.1038/nchem.1411.

(205) Nikbin, N.; Caratzoulas, S.; Vlachos, D. G. *ChemCatChem* **2012**, *4*, 504–511, DOI: 10.1002/cctc.201100444.

(206) Frei, M. S.; Mondelli, C.; García-Muelas, R.; Kley, K. S.; Puértolas, B. n.; López, N.; Safonova, O. V.; Stewart, J. A.; Curulla Ferré, D.; Pérez-Ramírez, J. *Nat. Commun.* **2019**, *10*, 3377, DOI: 10.1038/s41467-019-11349-9.

(207) Piccinin, S.; Stamatakis, M. *Top. Catal.* **2016**, *60*, 141–151, DOI: 10.1007/s11244-016-0725-5.

(208) Huš, M.; Hellman, A. *ACS Catal.* **2018**, *9*, 1183–1196, DOI: 10.1021/acscatal.8b04512.

(209) Ovesen, C.; Clausen, B.; Hammershøi, B.; Steffensen, G.; Askgaard, T.; Chorkendorff, I.; Nørskov, J.; Rasmussen, P.; Stoltze, P.; Taylor, P. *J. Catal.* **1996**, *158*, 170–180, DOI: 10.1006/jcat.1996.0016.

(210) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. *Chem. Sci.* **2018**, *9*, 7069–7077, DOI: 10.1039/c8sc01949e.

(211) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631, DOI: 10.1126/science.aau5631.

(212) Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. *Nat. Chem.* **2016**, *8*, 331–337, DOI: 10.1038/nchem.2454.

(213) Felton, K. C.; Rittig, J. G.; Lapkin, A. A. *Chemistry–Methods* **2021**, *1*, 116–122, DOI: 10.1002/cmtd.202000051.

(214) Xiong, J.; Shi, S.-Q.; Zhang, T.-Y. *Mater. Design* **2020**, *187*, 108378, DOI: 10.1016/j.matdes.2019.108378.

(215) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. *Nat. Chem.* **2009**, *1*, 37–46, DOI: 10.1038/nchem.121.

(216) Pérez-Ramírez, J.; López, N. *Nat. Catal.* **2019**, *2*, 971–976, DOI: 10.1038/s41929-019-0376-6.

(217) Garcia-Ratés, M.; López, N. *J. Chem. Theory Comput.* **2016**, *12*, 1331–1341, DOI: 10.1021/acs.jctc.5b00949.

(218) Williams, C. K. I.; Rasmussen, C. E. In *Advances in Neural Information Processing Systems 8*, MIT press: 1996, pp 514–520.

(219) Nitopi, S.; Bertheussen, E.; Scott, S. B.; Liu, X.; Engstfeld, A. K.; Horch, S.; Seger, B.; Stephens, I. E. L.; Chan, K.; Hahn, C.; Nørskov, J. K.; Jaramillo, T. F.; Chorkendorff, I. *Chem. Rev.* **2019**, *119*, 7610–7672, DOI: 10.1021/acs.chemrev.8b00705.

(220) Kuhl, K. P.; Cave, E. R.; Abram, D. N.; Jaramillo, T. F. *Energ. Environ. Sci.* **2012**, *5*, 7050, DOI: 10.1039/c2ee21234j.

(221) Schmid, B.; Reller, C.; Neubauer, S.; Fleischer, M.; Dorta, R.; Schmid, G. *Catalysts* **2017**, *7*, 161, DOI: 10.3390/catal7050161.

(222) Kwon, Y.; Lum, Y.; Clark, E. L.; Ager, J. W.; Bell, A. T. *ChemElectroChem* **2016**, *3*, 1012–1019, DOI: 10.1002/celc.201600068.

(223) Wang, X.; Xu, A.; Li, F.; Hung, S.-F.; Nam, D.-H.; Gabardo, C. M.; Wang, Z.; Xu, Y.; Ozden, A.; Rasouli, A. S.; Ip, A. H.; Sinton, D.; Sargent, E. H. *JACS* **2020**, *142*, 3525–3531, DOI: 10.1021/jacs.9b12445.

(224) Birdja, Y. Y.; Pérez-Gallent, E.; Figueiredo, M. C.; Göttle, A. J.; Calle-Vallejo, F.; Koper, M. T. M. *Nat. Energy* **2019**, *4*, 732–745, DOI: 10.1038/s41560-019-0450-y.

(225) Gao, D.; Sinev, I.; Scholten, F.; Arán-Ais, R. M.; Divins, N. J.; Kvashnina, K.; Timoshenko, J.; Roldan Cuenya, B. *Angew. Chem. Int. Ed.* **2019**, *58*, 17047–17053, DOI: 10.1002/anie.201910155.

(226) Li, J. et al. *Nat. Commun.* **2018**, *9*, 4614, DOI: 10.1038/s41467-018-07032-0.

(227) Mandal, L.; Yang, K. R.; Motapothula, M. R.; Ren, D.; Lobaccaro, P.; Patra, A.; Sherburne, M.; Batista, V. S.; Yeo, B. S.; Ager, J. W.; Martin, J.; Venkatesan, T. *ACS Appl. Mater. Interfaces* **2018**, *10*, 8574–8584, DOI: 10.1021/acsami.7b15418.

(228) Wang, X. et al. *Nat. Commun.* **2019**, *10*, 5186, DOI: 10.1038/s41467-019-13190-6.

(229) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. *Phys. Rev. Lett.* **2007**, *99*, 016105, DOI: 10.1103/physrevlett.99.016105.

(230) Peixoto, T. P. *figshare* **2014**, DOI: 10.6084/m9.figshare.1164194.

(231) Siek, J., *The boost graph library : User guide and reference manual*; Addison-Wesley: Boston, 2002.

(232) Friedman, J. H.; Stuetzle, W. *J. Amer. Statist. Assoc.* **1981**, *76*, 817–823, DOI: 10.1080/01621459.1981.10477729.

(233) Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning*; Springer New York: New York, 2009, DOI: 10.1007/978-0-387-84858-7.

(234) Abadi, M. et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems https://www.tensorflow.org/ (accessed 04/01/2022).

(235) Paszke, A. et al. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: 2019, pp 8024–8035.

(236) Agarap, A. F. *CoRR* **2018**, DOI: 10.48550/arXiv.1803.08375.

(237) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature* **1986**, *323*, 533–536, DOI: 10.1038/323533a0.

(238) Manzyuk, O.; Pearlmutter, B. A.; Radul, A. A.; Rush, D. R.; Siskind, J. M. *J. Funct. Program.* **2019**, *29*, DOI: 10.1017/s095679681900008x.

(239) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. *AI Open* **2020**, *1*, 57–81, DOI: 10.1016/j.aiopen.2021.01.001.

(240) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. *Proc. IEEE* **1998**, *86*, 2278–2324, DOI: 10.1109/5.726791.

(241) Hubel, D. H.; Wiesel, T. N. *J Physiol* **1968**, *195*, 215–243, DOI: 10.1113/jphysiol.1968.sp008455.

(242) Fukushima, K. *Scholarpedia* **2007**, *2*, 1717, DOI: 10.4249/scholarpedia.1717.

(243) Cohen, N.; Benson, S. W. *Chem. Rev.* **1993**, *93*, 2419–2438, DOI: 10.1021/cr00023a005.

(244) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. *J Phys Chem* **1973**, *77*, 1687–1691, DOI: 10.1021/j100632a019.

(245) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546–572, DOI: 10.1063/1.1744539.

(246) Benson, S. W. *J. Chem. Educ.* **1965**, *42*, 502, DOI: 10.1021/ed042p502.

(247) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264, DOI: 10.1021/acs.jctc.7b00577.

(248) Xie, T.; Grossman, J. C. *Phys. Rev. Lett.* **2018**, *120*, 145301, DOI: 10.1103/physrevlett.120.145301.

(249) Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. *J. Phys. Chem. Lett.* **2019**, *10*, 4401–4408, DOI: 10.1021/acs.jpclett.9b01428.

(250) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. *J. Phys. Chem. Lett.* **2020**, *11*, 3185–3191, DOI: 10.1021/acs.jpclett.0c00634.

(251) Klicpera, J.; Becker, F.; Günnemann, S. *CoRR* **2021**, DOI: 10.48550/arXiv.2106.0890.

(252) Flam-Shepherd, D.; Wu, T. C.; Friederich, P.; Aspuru-Guzik, A. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 045009, DOI: 10.1088/2632-2153/abf5b8.

(253) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. *CoRR* **2019**, DOI: 10.48550/arXiv.1910.10685.

(254) Bohnet, M., *Ullmann's encyclopedia of industrial chemistry*; Wiley-VCH: Weinheim, Germany, 2002.

(255) Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; Grohe, M. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019**, *33*, 4602–4609, DOI: 10.1609/aaai.v33i01.33014602.

(256) Vinyals, O.; Bengio, S.; Kudlur, M. *CoRR* **2015**, DOI: 10.48550/arXiv.1511.06391.

(257) Kingma, D. P.; Ba, J. *CoRR* **2014**, DOI: 10.48550/arXiv.1412.6980.

Appendix A

Appendix: Algorithm notation

Most of the algorithms found in this work are presented by using Haskell type signatures. For example the Sum of Squared Errors may have the following signature:

$$sse : [a] \rightarrow a$$

Where *sse* is the name of the function, a is a data type and $[a]$ represents a `list` of objects of type a. *sse* then takes as argument a collection of values of type a ($[a]$) and returns a value of the type a.

The following nomenclature is used in the signatures found in this work:

- $f : a \rightarrow b$: A function f that takes as argument a value of type a and returns a value of type b.
- $g : a \rightarrow b \rightarrow c$: A function g that takes as argument a value of type a and a value of type b and returns a value of type c.
- $[a]$: `list` of items of type a.
- $a \cong b$: a and b are isomorphic, and thus both contain the same information. e. g. $molecule \cong [atoms]$
- (a, b): `tuple` containing a left value of type a and a right value of type b.
- *Either a b*: Either a value of type a or a value of type b.
- *Maybe a*: Either a value of type a or *Nothing*.
- $f \circ g$: Function composition. For f and g being two functions, f after g $f(g(x))$.

#+LATEX

Publications

UNIVERSITAT ROVIRA I VIRGILI
MORE IS DIFFERENT: MODERN COMPUTATIONAL MODELING FOR HETEROGENEOUS CATALYSIS
Sergio Pablo García Carillo

Turning chemistry into information for heterogeneous catalysis

Pablo-Garcia S.[1], Alvarez M.[1,2], and Lopez N.[1]

[1]Institute of Chemical Research of Catalonia, ICIQ, Av. Paisos Catalans 16, 43007 Tarragona, Catalonia, Spain
[2]Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, C/Marcelli Domingo s/n, 43007 Tarragona, Catalonia, Spain

February 8, 2020

Abstract

The growing generation of data and their wide availability has led to the development of tools to produce, analyze and store this information. Computational chemistry studies and especially catalytic applications often yield a vast amount of chemical information that can be analyzed and stored using these tools. In this manuscript we present a framework that automatically performs a full automated procedure consisting in the transfer of an adsorbate from a known metal slab to a new metal slab with similar packing. Our method generates the new geometry and also performs the required calculations and analysis to finally upload the processed data to an online database (ioChem-BD). Two different implementations have been built, one to relocate minimum energy point structures and the second to transfer transition states. Our framework show a good performance for the minimum point location and a decent performance for the transition state identification. Most of the failures occurred during the transition state searches needed additional steps to fully complete the process. Further improvements of our framework are required to increase the performance of both implementations. These results point to the *avoidhuman* path as a feasible solution for studies on very large systems that require a significant amount of human resources and in consequence are prone to human errors.

1 Introduction

Computational chemistry is nowadays ubiquitous and has applications in Chemistry, Biology, Physics, Materials Science and Nanotechnology. As the access to massive computers and robust codes (Lejaeghere et al., 2016) extends worldwide, databases for molecules, nanostructures and materials containing structural data, spectroscopic fingerprints (Grimme et al., 2017) and general properties can be easily generated. The ultimate applications of these databases can vary from environmental detection through spectroscopy to data mining for materials in Catalysis and Electrocatalysis (Kitchin, 2018). And yet, most of the purpose-oriented calculations are not saved (general case in Materials and Heterogeneous Catalysis) or are only presented as lengthy xyz coordinate listings in error-prone Supplementary files. Only lately, the relevance of keeping this data in the form of databases has been realized (Bo et al., 2018). Most of the systems though have emerged in Materials Science in projects such as the Materials Project (Jain et al., 2013; *Materials Project*, n.d.), NoMaD (*NoMaD Repository*, n.d.), Materials Cloud (*Materials Cloud*, n.d.) and Computational Materials Repository (*Computational Materials Repository*, n.d.). Data are mostly unlinked to the corresponding works and thus the traceability (who, when, what) and fairness (findable, accessible, interoperable, reusable) are lost (Wilkinson et al., 2016).

Human factors pose an additional problem. Researchers have to perform very routine tasks, therefore they can make mistakes and the generated dense datasets often show/have multiple deficiencies. Factors of humans

include, for example: cognitive functions (such as attention, detection, perception, judgement and reasoning (including heuristics and biases), decision making - each of these is further divided into sub-categories). These issues are particularly problematic when statistical learning techniques (Garca-Muelas & Lpez, 2019; Bruix et al., 2019; Turcani et al., 2018; Butler et al., 2018; Gryn'ova et al., 2018; Ulissi et al., 2017; Gómez-Bombarelli et al., 2018; Schlexer Lamoureux et al., 2019; Meyer et al., 2018; Nandy et al., 2019; Moghadam et al., 2019) are applied; sparse datasets are biased towards a particular type of successful event, exactly what statistical learning algorithms need to avoid to ensure their robustness. Repetitiveness has been addressed by different groups by using scripts with different level of sophistication. However, the emergence of new frameworks that can steer the tedious tasks and generate/check/upload to a database significant data blocks of the phase spaces provides the right tools into the automation concepts (Larsen et al., 2017; Pizzi et al., 2016; *Open Babel: Te Open Source Chemistry Toolbox*, n.d.; Gromski et al., 2019). Generation is still one of the key steps. Some very recent efforts have been made to automate the identification of adsorption sites in metal surfaces, allowing to create new structures with different combinations of sites and adsorbates (Boes et al., 2019; Tran et al., 2018; Montoya & Persson, 2017). Statistical learning techniques have also been applied to get an estimate for the adsorption energies (Ulissi et al., 2017; Tabor et al., 2018). Our work tries to use structure inheritance between different metals not only to automate the study of reaction networks in different metals, and alike materials (like oxides) but also to reduce the explicit Density Functional Theory calculation time not only for adsorbates (a relatively straightforward task) but most relevantly for transition states for a real catalytic problem.Diagnose exceptions and analysis will be the focus of the research in computational chemistry in the coming years. This will increase our abilities to find outliers that can be crucial to performance (and identification of new families of molecules and compounds with particularly appealing properties), to refine the analytics, to incorporate graph theory (Bjørn Jørgensen et al., 2019) or other encodings like SMILES (Weininger, 1988) to be able to transfer active patterns irrespective of the nature of the compound (solid, enzymatic, molecular).

Complementary to data generation and analysis the forms of acquiring information have changed. The new editorial platforms like Authorea can also integrate the advances of these systematic approaches. The process of reading documents has severely changed since the establishment of the world-wide web and the availability of more than one instance simultaneously running. Reading in the 21st century is a completely different experience than it has been for at least 500 years as the meta- and linked data are accessible and are consulted almost simultaneously with the primary source. New ways of reading can thus now benefit from interactive viewers that can integrate the content and improve the visualization of complex information (for instance 3D).

Our manuscript tries to address all these new challenges in computational chemistry and with the overall idea of transforming the results into a true seamless information science that can be interactively read, dynamically searched, analyzed and mined, while ensuring the transferability among different fields through metadata storage. To this end the combination of Fireworks (Jain et al., 2015), ioChem-BD (Álvarez-Moreno et al., 2014; *ioChem-BD*, n.d.) and Authorea (*Open Research Collaboration and Publishing - Authorea*, n.d.) constitutes a privileged platform.

2 Computational Details

Density functional theory (DFT) simulations have been performed using Vienna ab-initio Simulation Package (VASP) (Kresse & Furthmüller, 1996; Kresse & Furthmüller, 1996). Generalized Gradient Aproximation with the Perdew-Burke-Ernzerhof functional (GGA-PBE) have been used to obtain the exchange-correlation energy term (Perdew et al., 1996). Valence electrons have been represented using the Projector Augmented Wave (PAW) (Blöchl, 1994) with a kinetic energy cutoff of 450 eV. To generate the k-points a Γ-centered mesh has been built using the Monkorst-Pack method (Monkhorst & Pack, 1976). Improved dimer method has been employed to locate the transition states (Heyden et al., 2005).

Fireworks (Jain et al., 2015) is an open source software package that allows to managing building and

running workflows. In this project it has been used to built and configure the ties between the different steps of our calculations. It also allows to track the status of active workflows and stops the process if one of the steps fails, allowing to restart the process completely or to continue the step after the application of the needed changes. Software written in Python and Bash has been developed in our lab and used to script the preparation, transfer and checking processes. Developed Python libraries and scripts focus on geometry manipulation and input/output parsing, while Bash scripts intend to manage the files related to the calculations, and control the execution of VASP.

The framework has been tested using a reaction network composed of 9 different metals, 7 of them with p(3x3)-(111) fcc packing and 2 with p(3x3)-(0001) hcp packing. One of the hcp metals serves as the ansatz host for the rest of the metals. Both packings share similar adsorption sites in their surfaces. The two lowest layers have been frozen and a vacuum of 15 Å has been applied to all metal slabs.

A total of 12 organic intermediates with the chemical formula $C_1Br_xH_y$ and 8 transition states have been evaluated with our method. All the intermediates belong to the same reaction network, being the transition states all the possible elemental steps involving the intermediates. For the minimum energy relaxations, different adsorption sites have been calculated for every metal/species combination whereas only one adsorption site has been tested for the transition states. The species were first obtained manually for the hcp metal and then recalculated for the other metals by using our framework.

The figures included in this paper were made with Cytoscape, 3Dmol and Plotly (Shannon et al., 2003; Rego & Koes, 2014; Inc., 2015).

3 Results and Discussion

Full mechanistic DFT studies on the decomposition of organic molecules are among the most challenging processes in heterogeneous catalysis, due to the amount of elemental steps and intermediate species comprised in the network as well as the subsequent screening that require the calculation of these reactions over a set of metal slabs (Muelas et al., 2017). The size of these networks are tightly bonded with the number of carbons of the organic species and for a significant number of carbons the system require a massive amount of time and resources to be evaluated. To address this issue, some databases have been designed to store and ease the access to these steps with the aim to recycle old calculations and reduce the required steps when studying these networks (*CatApp database COMPUTATIONAL MATERIALS REPOSITORY*, n.d.).

We propose an automated procedure to overcome the drawbacks associated with the repetitive processes required for the study of large reaction networks. We have built a framework that combines Fireworks, VASP, ioChem-BD and ad-hoc developed software to fully automate the study of reaction networks over different metals. This framework allows transferring all the relaxed species and transition states obtained for a metal slab to the slab of a different metal as well as preparing and performing the DFT calculations using the generated geometries for the new slab automatically. The simplicity of both, the C_1 species and the pure metal slabs together with the complexity degree added by the inclusion of a halogen give us a optimal scenario to benchmark the framework.

3.1 Transfer algorithm

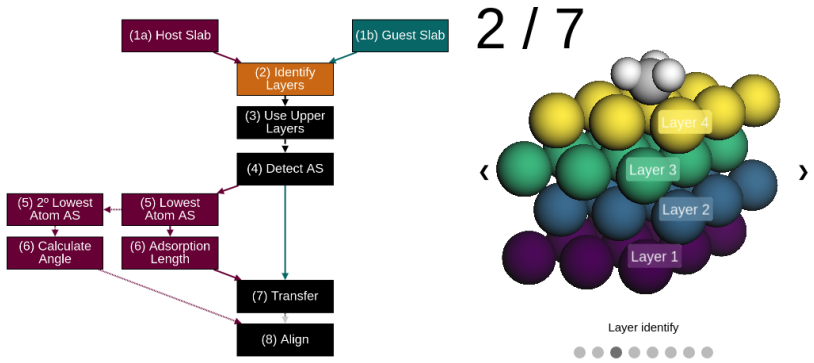


Figure 1: Interactive diagram of the steps that the transfer algorithm performs. While some of the steps only use the data from the host slab, the first (1)-(5) steps are applied concurrently to both slabs.

The process of generating guess geometries for intermediates is one of the most tedious and time-consuming tasks particularly when dealing with large reaction networks, e.g. C_6 sugar alcohol decomposition consists of 10^6 intermediates (Sutton & Vlachos, 2015). Inheriting obtained structures from similar calculations eases the problem, but still it results in a repetitive routine problem that can be fully automated.

A transfer algorithm that performs a relocation of an adsorbed molecule in a metal slab to a similar metal slab has been developed. Automatic transfer of adsorbed species between similar metallic surfaces allows not only saving a considerable amount time during the geometry production but also classifying the generated geometries accurately.

When there are two different metal slabs, a host with an adsorbed molecule (1a) and a guest consisting on an empty metal slab (1b), the transfer algorithm works as follow: First, identifies the layers (2) for both slabs and selects the highest layer (3). Then, searches for all the possible adsorption sites (4) on both surfaces. The site is found by triangulation of the different metal centers. Once the adsorption sites have been identified, the algorithm associates the nearest adsorption site with the lowest atom of the adsorbed molecule in the host slab. (5) In the next step, the adsorption length is computed employing the distance between the assigned site and the lowest atom of the adsorbate. However, for some molecules the lowest atom (z-axis position) is not perfectly aligned with the adsorption site. To overcome this issue, the deviation of the z vector between both is also computed in this step (6). Lastly, the algorithm transfers the molecule to a similar site in the guest surface, taking into account the adsorption site type and maintaining the adsorption length (7). As an optional step, the algorithm can be set to identify the nearest adsorption site of the second lowest (5o) atom and compute the angle between both (6o), this information is then used to rotate the molecule around the z-axis of the lowest atom to preserve the original alignment of the molecule (8). Figure 1 depicts the procedure of the algorithm.

The transfer algorithm applies different methods to find the possible adsorption sites. For the fcc and hcp holes detection, the Voronoi tetrahedron method (Isayev et al., 2017) is used to compute the bonds between the atoms of the upper layer, to then search for cycles of three atoms. Differentiation between hcp and fcc

holes is achieved projecting the triangle formed by the cycles in the lower layer and searching for atoms inside this space. Once the bonds are defined, the bridge and top positions are easy to find.

3.2 Workflow design

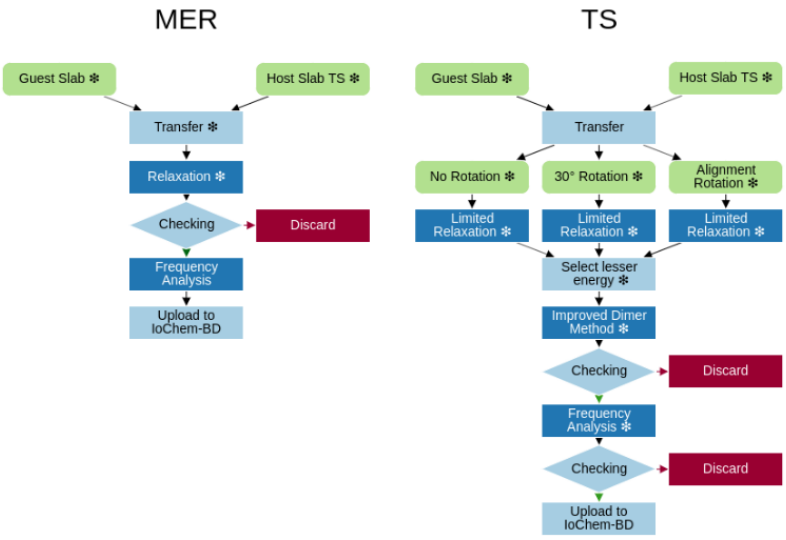


Figure 2: Interactive flux diagram showing the required steps for the relax and TS processes. DFT calculations are colored in dark blue and scripts light blue. Clicking the different points marked with () will show the status of the structure at this point.

The geometries generated using the transfer algorithm are not optimized and require further DFT calculations to obtain a full description of the reaction network. To address this problem two different workflows have been developed and embedded in the framework: one for the minimum energy relaxation (**MER**) and another for the transition state search (**TS**). Both workflows use the transfer algorithm as first step to generate a new metal slab with the desired adsorbate, then, different Bash scripts prepare and run the pertinent DFT calculations using VASP.

While the workflow to search relaxed geometries only generates a single geometry, the transition state (TS) workflow generates three unique geometries with different rotations along the z-axis of the lowest atom. A partial minimum energy search for a small number of ionic steps is performed for the three structures, the atoms of the molecule with high initial forces are allowed to relax thus preventing the TS to fall to a minimum energy point. The best candidate among the relaxed structures is selected using a lowest energy

criteria; in the next step, the chosen candidate is used as the starting geometry in an improved dimer method calculation. Selection steps are unnecessary for the MER workflow due to the efficiency of the minimum energy point relax algorithms in front of the transition search ones. Thus, the generated geometry for MER workflow is directly used as starting point for the minimum energy search calculations.

Errors of different nature can occur during the calculation steps; besides the geometry obtained after the calculation may be chemically meaningless, therefore, it is important to manage errors with care. To address these issues, a checking algorithm is applied to the results of the calculations in order to search for inconsistencies through the calculation steps. Additionally, a bond identification algorithm is used to verify that no bond breaking occurs during the relaxations.

This algorithm, that detects spurious non-valid broken adsorbates was implemented at the checking stage. The process splits the structure between the (metal) surface and the adsorbed molecules. Then the bonds between the atoms of the adsorbates are detected using Voronoi tetrahedra algorithm (Isayev et al., 2017) (with a cutoff radius) and converted to a graph and the number of disconnected graphs is computed. The difference between the initial and final number of disconnected graphs illustrates if the adsorbate has broken during optimization. If the result passes the checking, a frequency calculation is launched.

After a few tests, an improvement has been integrated to the transfer algorithm. The bond recognition was reused to analyze the bonds of the geometries obtained from the MER workflows. As a result, a list of the average distance between the metal surface and the lower atom element was obtained for each metal surface. The difference between these distances was used to correct the adsorption distance of some of the failed MER workflows and all of the TS workflows.

3.3 Storing data

Due to the large number of individual results that compose complex reaction networks it is mandatory to compile, sort and store the results. ioChem-BD (Álvarez-Moreno et al., 2014; *ioChem-BD*, n.d.) provides the essential tools to perform the last step of our project: to convert our results into organized data. The shell client of ioChem-BD allows an easily upload of the generated output files to a private server. Once uploaded, it transforms, parses and store the results to ensure a clear representation and an easy online access to the data and can generate molecular labels as SMILES.

Consequently, as the final step of both previously described workflows, the results of the DFT calculations as well as the frequency calculations performed at the end of the workflows are uploaded to ioChem-BD. For the TS workflow, the data generated by the dimer method is uploaded whereas for the MER workflow the data from the relaxation method is uploaded.

Once all the data are processed, they can be published in the public repository of the ioChem-BD server and then the obtained information can be shared with other researchers keeping the FAIR principles. Moreover, ioChem-BD generates interactive figures that improve the understanding of geometries and reactivity for complex adsorbed molecules when working in multidisciplinary environments (f. ex. with experimental groups).

Datasets can also be published with an embargo option, publishing the dataset in a private repository but generating a DOI and a Reviewer Link. This link allows the coworkers, editor and reviewers to inspect the dataset before making it public. Once the associated manuscript is published, the dataset is synchronized to the public repository, making it accessible to everyone through the dataset DOI. The data then can be accessed via DOI, or by searching directly in the platform by Author, date, SMILES, chemical formula.

When a calculation is uploaded to ioChem-BD additional metadata is generated and stored within the calculation, a trustworthy fingerprint of every calculation is created and deposited in the system. Table 1 contains the minimum parameters that are saved when a calculation is uploaded to ioChem-BD. Two different schemes are used to generate the content tree: the Dublin Core standard (*DCMI*, n.d.) and a the Chemical Custom format created for ioChem-BD to append additional data.

Metadata	
Dublin Core (dc)	*Chemical Custom*
contributor.author	program.name
contributor.other	program.version
date.accessioned	program.other
date.available	method
date.created	shelltype
date.issued	energy.value
identifier.uri	energy.units
description	formula.generic
publisher	hassolvent
relation.ispartof	hasvibrationalfrequencies
rights	numberofjobs
rights.uri	hasmolecularorbital
title	
type	
date.updated	

Table 1: Minimum metadata stored in ioChem-BD for each calculation.

While he output file from VASP is processed and converted to our custom .cml format, the raw input files are stored as they are. To optimize space usage, the particular electronic and ionic iterations are not parsed and only the final structure is stored. Inclusion of the raw input files and the VASP version used allows to accurately reproduce every calculation of the dataset.

Additional information for clarity or on the structure of the different calculations can also be added. ioChem-BD offers several options to include this information so that notes can be added to its description. However, in some cases this information is rather complex and requires a different format. For this instances, ioChem-BD offers the possibility to attach raw files together with the input files. Both the description and the attachments will be available once the calculation is published in the ioChem-BD server.

3.4 Test and results

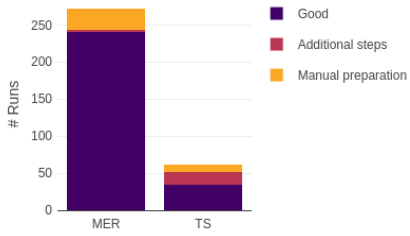


Figure 3: Efficiency of the MER and TS workflows. Additional steps required implies that error checking routines have idenfied a deficiency in the calculation (i.e. maximum convergence steps reached), instead Manual preparation points out that the calculation failed and requires human intervention (i.e. bond breaking).

Both workflows yields reasonable results and apply correctly the procedure for the different species (see Figure 3). Almost 88% of the minimum energy relaxations and 56% of the transition states converge at the first attempt without issues. Additional steps are required for 0.1% of the relaxed models and 27% of the transition states. The rest require manual preparation and supervision to terminate successfully. Among the failures for MER the exhaustion of the number of cycles was the main cause, the solution was to add about 10% ionic cycles (not completely used in all cases). This can thus be directly introduced in alternative network studies for metals and alloys. For TS there were two issues: (i) the same as for MER, insufficient ionic steps in the algorithm, this was sorted in the same manner and was by far the most common; (ii) alternative convergence was not achieved and thus the initial seed was not taken from the reference host (Ru) but instead from a metal with chemical properties closer to the running TS search (Pt was inherited from Ir).

In the case of relaxations, most of the unobtainable structures fall into a different adsorption sites during the calculation, while only few of them end the calculation with a broken bond. On the other hand, all the unobtainable transition states end with a broken bond and more precise methods such as the Nudged-Elastic Band (NEB) (Henkelman & Jnsson, 2000; Henkelman et al., 2000) are required to obtain the geometries.

3.5 Integration with Authorea

The Authorea (*Open Research Collaboration and Publishing - Authorea*, n.d.) platform allows an easy integration of the data stored in ioChem-BD. For instance, the reaction networks and particular structures can be directly linked in the database. In our case, the structure for the CH_3 on two different surfaces Ni and Ru can be retrieved from Host Slab and Guest Slab and the transition state for decomposition from Transition State.

4 Conclusions

We have proved that our framework automates two different kinds of molecular transfers through similar metals. Although our method is not yet able to fully automate the entire process successfully, it is possible to classify the different error cases obtained during our study and incorporate the solutions as additional steps for our workflows. In addition, coupled to the Authorea and ioChem-BD tools allows has the following advantages: (i) it enables to establish a seamless link between the computed data, the manuscript and interactively linked the corresponding structures avoiding tedious and error-prone supporting information; (ii) this is particularly attractive for complex databases with massive reaction networks and/or several material compositions, (iii) the workflow reduces the computing time, systematizes the nomenclature and labeling of the different species reducing the chaos an increasing transferability; (iv) the metadata directly embedded in ioChem-BD is made transparent through the Authorea and can improve the design (and self-definition) of the working flows that could potentially include data analysis through coupled machine learning algorithms; (v) the data curation is thus directly enforced by this procedure.

References

Reproducibility in density functional theory calculations of solids. (2016). *Science*, *351*(6280). https://doi.org/10.1126/science.aad3000

Fully Automated Quantum-Chemistry-Based Computation of SpinSpin-Coupled Nuclear Magnetic Resonance Spectra. (2017). *Angewandte Chemie International Edition*, *56*(46), 1476314769. https://doi.org/10.1002/anie.201708266

Machine learning in catalysis. (2018). *Nature Catalysis*, *1*(4), 230232. https://doi.org/10.1038/s41929-018-0056-y

The role of computational results databases in accelerating the discovery of catalysts. (2018). *Nature Catalysis*, *1*(11), 809810. https://doi.org/10.1038/s41929-018-0176-4

The Materials Project: A materials genome approach to accelerating materials innovation. (2013). *APL Materials*, *1*(1), 011002. https://doi.org/10.1063/1.4812323

https://materialsproject.org. https://materialsproject.org

https://nomad-coe.eu. https://nomad-coe.eu

https://www.materialscloud.org/home. https://www.materialscloud.org/home

https://cmr.fysik.dtu.dk. https://cmr.fysik.dtu.dk

The FAIR Guiding Principles for scientific data management and stewardship. (2016). *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. (2019). In *Nat. Comm. (In press)*.

First-principles-based multiscale modelling of heterogeneous catalysis. (2019). *Nature Catalysis*, *2*(8), 659670. https://doi.org/10.1038/s41929-019-0298-3

Machine Learning for Organic Cage Property Prediction. (2018). *Chemistry of Materials*, *31*(3), 714727. https://doi.org/10.1021/acs.chemmater.8b03572

Machine learning for molecular and materials science. (2018). *Nature*, *559*(7715), 547555. https://doi.org/10.1038/s41586-018-0337-2

Read between the Molecules: Computational Insights into Organic Semiconductors. (2018). *Journal of the American Chemical Society*, *140*(48), 1637016386. https://doi.org/10.1021/jacs.8b07985

To address surface reaction network complexity using scaling relations machine learning and DFT calculations. (2017). *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms14621

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. (2018). *ACS Central Science*, *4*(2), 268276. https://doi.org/10.1021/acscentsci.7b00572

Machine Learning for Computational Heterogeneous Catalysis. (2019). *ChemCatChem*, *11*(16), 35813601. https://doi.org/10.1002/cctc.201900595

Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. (2018). *Chemical Science*, *9*(35), 70697077. https://doi.org/10.1039/c8sc01949e

Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation. (2019). *ACS Catalysis*, *9*(9), 82438255. https://doi.org/10.1021/acscatal.9b02165

Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. (2019). *Matter*, *1*(1), 219234. https://doi.org/10.1016/j.matt.2019.03.002

The atomic simulation environmenta Python library for working with atoms. (2017). *Journal of Physics: Condensed Matter*, *29*(27), 273002. http://stacks.iop.org/0953-8984/29/i=27/a=273002

AiiDA: automated interactive infrastructure and database for computational science. (2016). *Computational Materials Science*, *111*, 218230. https://doi.org/https://doi.org/10.1016/j.commatsci.2015.09.013

http://openbabel.org/wiki/Main_Page. http://openbabel.org/wiki/Main_Page

Universal Chemical Synthesis and Discovery with 'The Chemputer'. (2019). *Trends in Chemistry*. https://doi.org/10.1016/j.trechm.2019.07.004

Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. (2019). *The Journal of Physical Chemistry A*, *123*(11), 22812285. https://doi.org/10.1021/acs.jpca.9b00311

Dynamic Workflows for Routine Materials Discovery in Surface Science. (2018). *Journal of Chemical Information and Modeling*, *58*(12), 23922400. https://doi.org/10.1021/acs.jcim.8b00386

A high-throughput framework for determining adsorption energies on solid surfaces. (2017). *Npj Computational Materials*, *3*(1). https://doi.org/10.1038/s41524-017-0017-z

Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO2 Reduction. (2017). *ACS Catalysis*, *7*(10), 66006608. https://doi.org/10.1021/acscatal.7b01648

Accelerating the discovery of materials for clean energy in the era of smart automation. (2018). *Nature Reviews Materials*, *3*(5), 520. https://doi.org/10.1038/s41578-018-0005-z

Materials property prediction using symmetry-labeled graphs as atomic-position independent descriptors. (2019). *ArXiv e-Prints*, arXiv:1905.06048.

SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. (1988). *Journal of Chemical Information and Computer Sciences*, *28*(1), 3136. https://doi.org/10.1021/ci00057a005

FireWorks: a dynamic workflow system designed for high-throughput applications. (2015). *Concurrency and Computation: Practice and Experience*, *27*(17), 50375059. https://doi.org/10.1002/cpe.3505

Managing the computational chemistry big data problem: the ioChem-BD platform. (2014). *Journal of Chemical Information and Modeling*, *55*(1), 95103. https://doi.org/10.1021/ci500593j

https://www.iochem-bd.org. https://www.iochem-bd.org

https://authorea.com/. https://authorea.com/

Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. (1996). *Comput. Mater. Sci.*, *6*(1), 1550. https://doi.org/10.1016/0927-0256(96)00008-0

Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. (1996). *Phys. Rev. B*, *54*(16), 1116911186. https://doi.org/10.1103/PhysRevB.54.11169

Generalized Gradient Approximation Made Simple. (1996). *Phys. Rev. Lett.*, *77*(18), 38653868. https://doi.org/10.1103/PhysRevLett.77.3865

Projector augmented-wave method. (1994). *Phys. Rev. B*, *50*(24), 1795317979. https://doi.org/10.1103/PhysRevB.50.17953

Special points for Brillouin-zone integrations. (1976). *Phys. Rev. B*, *13*(12), 51885192. https://doi.org/10.1103/PhysRevB.13.5188

Efficient methods for finding transition states in chemical reactions: Comparison of improved dimer method and partitioned rational function optimization method. (2005). *The Journal of Chemical Physics*, *123*(22), 224101. https://doi.org/10.1063/1.2104507

Cytoscape: a software environment for integrated models of biomolecular interaction networks. (2003). *Genome Research*, *13*(11), 24982504. https://doi.org/10.1101/gr.1239303

3Dmol.js: molecular visualization with WebGL. (2014). *Bioinformatics*, *31*(8), 13221324. https://doi.org/10.1093/bioinformatics/btu829

Collaborative data science. (2015). Plotly Technologies Inc. https://plot.ly

Ethylene_glycol_reaction_network. (2017). Institute of Chemical Research of Catalonia. https://doi.org/10.19061/iochem-bd-1-37

https://cmr.fysik.dtu.dk/catapp/catapp.html. https://cmr.fysik.dtu.dk/catapp/catapp.html

Building large microkinetic models with first-principles accuracy at reduced computational cost. (2015). *Chemical Engineering Science*, *121*, 190199. https://doi.org/10.1016/j.ces.2014.09.011

Universal fragment descriptors for predicting properties of inorganic crystals. (2017). *Nature Communications*, *8*, 15679. https://doi.org/10.1038/ncomms15679

https://www.dublincore.org. https://www.dublincore.org

Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. (2000). *The Journal of Chemical Physics*, *113*(22), 99789985. https://doi.org/10.1063/1.1323224

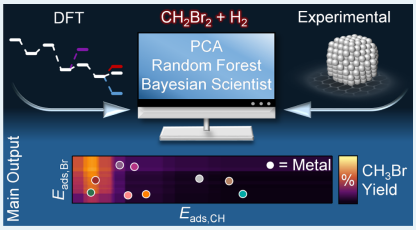
A climbing image nudged elastic band method for finding saddle points and minimum energy paths. (2000). *The Journal of Chemical Physics*, *113*(22), 99019904. https://doi.org/10.1063/1.1329672

This document is the Accepted Manuscript version of a Published Work that appeared in final form in ACS Catalysis, copyright © American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see https://doi.org/10.1021/acscatal.0c00679.

Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning

A. J. Saadun,[§] S. Pablo-García,[§] V. Paunović, Q. Li, A. Sabadell-Rendón, K. Kleemann, F. Krumeich, N. López,* and J. Pérez-Ramírez*

ABSTRACT: The catalyzed semihydrogenation of dibromomethane (CH_2Br_2) to methyl bromide (CH_3Br) is a key step in the bromine-mediated upgradation of methane. This study presents a cutting-edge strategy combining density functional theory (DFT), catalytic tests complemented with the extensive characterization of a wide range of metal catalysts (Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir, and Pt), and statistical tools for a computer-assisted investigation of this reaction. The steady-state catalytic tests identified four classes of materials comprising (i) poorly active (<8%) Fe/SiO_2, Co/SiO_2, Cu/SiO_2, and Ag/SiO_2; (ii) Rh/SiO_2 and Ni/SiO_2, which exhibit intermediate CH_3Br selectivity (<60%); (iii) Ir/SiO_2 and Pt/SiO_2, which display great propensity to CH_4 (>50%); and (iv) Ru/SiO_2, which exhibits the highest selectivity to CH_3Br (up to 96%). In-depth characterization of representative catalysts in fresh and used forms was done by X-ray diffraction, inductively coupled plasma optical emission spectroscopy, N_2 sorption, temperature-programmed reduction, Raman spectroscopy, electron microscopy, and X-ray photoelectron spectroscopy. The dimensionality reduction performed on the 272 DFT intermediate adsorption energies using principal component analysis identified two descriptors that, when employed together with the experimental data in a random forest regressor, enabled the understanding of activity and selectivity trends by connecting them to the energy intervals of the descriptors. In addition, a representative analytic model was found using the Bayesian inference. These findings illustrate the exciting opportunities presented by integrated experimental/computational screening and set the fundamental basis for the accelerated discovery of superior hydrodebromination catalysts and beyond.



KEYWORDS: *methane activation, dibromomethane, hydrodebromination, principal component analysis, random forest classifier, statistical analysis, density functional theory*

1. INTRODUCTION

The development of innovative approaches enabling the efficient and economical on-site valorization of natural gas into fuels and chemicals has become a strategic research area.[1−4] Among the various technologies, halogen-mediated processes have emerged as viable routes for the transformation of methane, the main constituent of natural gas, into transportable liquids.[5,6] In this regard, bromine is the preferred halogen over chlorine as it provides higher selectivities to the desired bromomethane (CH_3Br).[7,8] In addition, the weaker C−Br bond (2.95 eV) compared to the C−Cl bond (3.51 eV) allows facile HBr elimination, vital for halogen recycling within the process.[9] Nonetheless, the formation of significant amounts of dibromomethane (CH_2Br_2, selectivity up to 32%) in the gas-phase bromination of CH_4 hinders this technology to prosper at industrial scale.[10] Polyhalogenated byproducts shorten the lifetime of zeolites, main catalytic systems for the downstream halomethane coupling step, due to the increased coke formation.[11] The elimination of CH_2Br_2 by CH_4 bromination over heterogeneous catalysts has a limited scope, since the noncatalytic gas-phase radical pathways cannot be fully suppressed,[12] while reproportionation of CH_2Br_2 with CH_4 into CH_3Br requires long residence times (up to 60 s) and is thermodynamically constrained.[13]

In contrast, the selective reforming of polyhalomethanes *via* catalytic hydrodehalogenation, a class of semihydrogenation reactions, presents a practicable approach.[14,15] This reaction

has mainly been studied by Ding *et al.*, reporting the performance of noble-metal-based catalysts (Ru, Rh, Pd, Ag, Pt, and Au) supported on silica.[15] Therein, Pd/SiO_2 was shown to produce oligomers, revealing its full transformation into Pd_6C/SiO_2 under reaction conditions. Selective CH_2Br_2 hydrodebromination to CH_3Br was reported over Ru/SiO_2 (<96% selectivity) and Rh/SiO_2 (<83% selectivity), whereas Pt/SiO_2 produced mainly CH_4 (>47% selectivity to CH_4). Ag was found to rapidly oxidize to AgBr under reaction conditions, whereas Au was inactive. Still, the amount of literature on hydrodebromination chemistry on catalytic surfaces is very limited. In particular, other known hydrogenation catalysts such as Fe, Co, Ni, Cu, and Ir were never studied in this reaction and the stability performance of systems that selectively hydrodebrominate CH_2Br_2 to CH_3Br is not reported.[16,17]

The search for new catalysts typically starts with high-throughput experimental screening and by property similarity within the periodic table. New techniques based on statistical learning (SL) have been tentatively applied to guide this quest.[18] Some degree of success has been achieved when activity and, to a lesser extent, selectivity were taken as response functions.[19−27] The variables introduced as potential descriptors often correspond to preparation variables, for instance, the catalytic composition for optimization of dopant concentrations *via* artificial neural networks.[22−27] The main issues preventing the extensive application of such strategies are the lack of consistency between the databases originating from previous works and the fact that open literature contains only successful experiments, whereas SL techniques require the full scenario (*i.e.*, including conclusive but unsuccessful results) to make robust predictions. Activity descriptors have been reported from linear-scaling relationships (LSRs);[28] however, the identification of these descriptors can benefit SL techniques.

In this work, we present a systematic catalytic preparation and testing protocol coupled to mechanistic studies based on density functional theory (DFT) that can be employed as a complete database for the use of statistical learning inference of trends for a highly active and selective CH_2Br_2 hydrodebromination catalyst. Herein, the toolbox of statistical learning methodologies applied contains dimensionality reduction *via* principal component (PC) analysis (PCA) clustering, and classification techniques.[29−31] The findings reported in this work are first attempts directed at elucidating hydrodebromination performance patterns to lay the foundations for future catalyst design and to pave the way for the wider application of machine learning techniques to, for instance, multimetallic systems.

2. MATERIALS AND METHODS

2.1. Catalyst Preparation. Commercial SiO_2 (Evonik, AEROPERL 300/30, S_{BET} = 257 $m^2\ g^{-1}$, V_{pore} = 0.95 $cm^3\ g^{-1}$, >99.0%) was calcined at 973 K for 5 h in static air (heating rate 5 K min^{-1}) prior to its use as a support in the synthesis protocol. The metal precursors, $Fe(NO_3)_3{\cdot}9H_2O$ (Acros Organics, 99%), $Co(NO_3)_2{\cdot}6NH_3$ (abcr, 99%), $Ni(NO_3)_2{\cdot}6H_2O$ (Strem Chemicals, 99.9%), $Cu(NO_3)_2{\cdot}xH_2O$ (Sigma-Aldrich, 99.999%), $RhCl_3{\cdot}H_2O$ (Acros Organics, 99%), $AgNO_3$ (Sigma-Aldrich, 99.8%), $IrCl_4{\cdot}xH_2O$ (abcr, 99.9%), $RuCl_3{\cdot}xH_2O$ (abcr, 99.9%), and $Pt(NH_3)_4Cl_2{\cdot}xH_2O$ (Sigma-Aldrich, 99%), were dispersed on the support *via* incipient wetness impregnation. Appropriate amounts of the precursors required to obtain a metal loading of 1 wt % in the final catalyst were dissolved in a volume of deionized water equal to the pore volume of the carrier. The precursor solution was added dropwise to the support, and the mixture was magnetically stirred (500 rpm) for 30 min at room temperature. The resulting solids were dried at 373 K for 12 h and calcined in static air at 623 K (heating rate 5 K min^{-1}) to obtain the SiO_2-supported metal oxides, followed by their reduction under 20 vol % H_2 (PanGas, purity 5.0) in He (PanGas, purity 5.0) flow at elevated temperatures (573−968 K) for 3 h in the catalytic reactor with a heating rate of 10 K min^{-1} prior to their use in catalytic tests. The catalysts were referred to as M/SiO_2, where M denotes the metal (*i.e.*, Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir, or Pt). The specific catalyst obtained by direct reduction was denoted M/SiO_2-NC, where NC stands for "noncalcined."

2.2. Catalyst Characterization. Powder X-ray diffraction (XRD) was carried out in a PANalytical X'Pert PRO-MPD diffractometer with Bragg−Brentano geometry by applying Ni-filtered Cu Kα radiation (λ = 1.54060 Å). The data were recorded in the 10−70° 2θ range with an angular step size of 0.017° and a counting time of 0.51 s per step. The metal loading in the solids was determined by inductively coupled plasma optical emission spectroscopy (ICP-OES) using a Horiba Ultima 2 instrument equipped with photomultiplier tube detection. N_2 sorption at 77 K was measured in a Micromeritics TriStar II analyzer. Samples (*ca.* 0.1 g) were degassed to 50 mbar at 573 K for 12 h prior to the measurement. The Brunauer−Emmett−Teller (BET) method was applied to calculate the total surface area, S_{BET}. The pore volume, V_{pore}, was determined from the amount of N_2 adsorbed at a relative pressure of p/p_0 = 0.98. Temperature-programmed reduction with hydrogen (H_2-TPR) was conducted in a Micromeritics AutoChem II 2920 unit equipped with a thermal conductivity detector. The sample (*ca.* 0.1 g) was loaded in a U-shaped quartz reactor between two plugs of quartz wool and pretreated in He (20 $cm^3\ min^{-1}$) at 473 K for 10 min. The analysis was performed in 5 vol % H_2 in N_2 (20 $cm^3\ min^{-1}$) by heating up the catalyst in the range of 323−1100 K at 10 K min^{-1}. Raman spectroscopy was carried out on a WITec CRM200 confocal system using a 532 nm laser with 20 mW power, a 100× objective lens with numerical aperture (NA) = 0.9 (Nikon Plan), and a fiber-coupled grating spectrometer (2400 lines mm^{-1}), giving a spectral sampling resolution of 0.7 cm^{-1}. High-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) was conducted on an aberration-corrected HD2700CS microscope (Hitachi) at 200 kV. All samples were dispersed in ethanol, and some droplets were deposited onto lacey carbon-coated copper grids. The particle size distribution of the catalysts was obtained by examining more than 100 nanoparticles. X-ray photoelectron spectroscopy (XPS) measurements were performed on a Physical Electronics Quantum 2000 X-ray photoelectron spectrometer using monochromatic Al Kα radiation, generated from an electron beam operated at 15 kV, and equipped with a hemispherical capacitor electron-energy analyzer. The solids were analyzed at an electron takeoff angle of 45° and a pass energy of 46.95 eV. The samples were mounted onto the sample holder by pressing the powders onto an aluminum foil. The spectrometer was calibrated for the 4 Au $4f_{7/2}$ signal to be at 84.0 ± 0.1 eV with a resolution step width of 0.2 eV. The envelopes were fitted by mixed Gaussian−Lorentzian component profiles after Shirley background subtraction. The selected peak positions of the

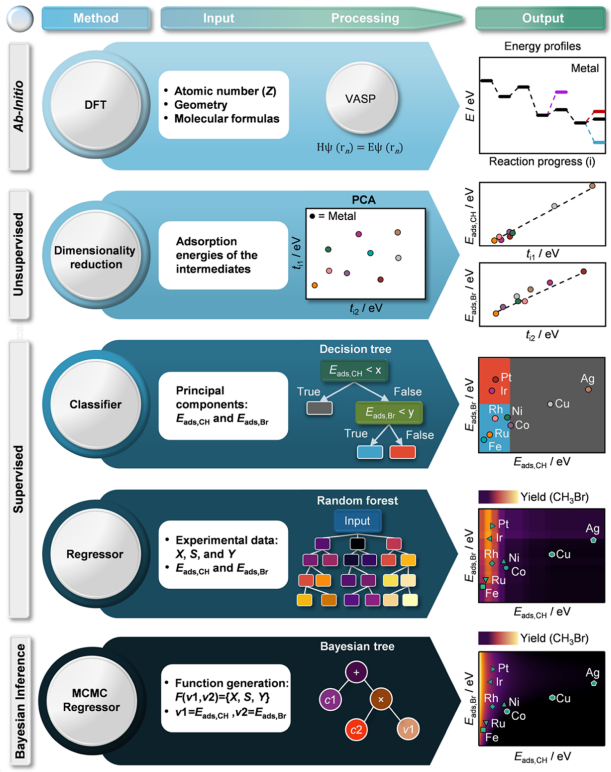


Figure 1. Overview of the multitechnique strategy combining experimentally obtained data, DFT results, and statistical tools to analyze the activity and selectivity of metal-catalyzed CH_2Br_2 hydrodebromination.

different species are based on literature-reported data and fixed with an error of ±0.3 eV.

2.3. Catalyst Testing. The hydrodebromination of dibromomethane was performed at ambient pressure in a home-made continuous-flow fixed-bed reactor setup. H_2 (PanGas, purity 5.0), He (Carrier gas, PanGas, purity 5.0), and Ar (internal standard, PanGas, purity 5.0) were dosed by a set of digital mass flow controllers (Bronkhorst), and liquid CH_2Br_2 (Acros Organics, 99%) was supplied by a syringe pump (Fusion 100, Chemyx) equipped with a water-cooled syringe to a vaporizer unit operated at 393 K. The quartz reactor (internal diameter, d_i = 12 mm) containing the reduced catalyst (catalyst weight, W_{cat} = 0.1−1 g, particle size, d_p = 0.4−0.6 mm) was heated to the desired temperature (T = 423−623 K) in an electric oven under He flow. The catalyst bed was allowed to stabilize for at least 10 min at the desired temperature before the reaction mixture was fed at a total volumetric flow (F_T) of 20−150 cm^3 STP min^{-1} and the desired feed composition of $CH_2Br_2/H_2/Ar/He$ = 6:24:4.5:65.5 (mol %), unless otherwise stated. Downstream linings were heated at 393 K to prevent the condensation of unconverted reactants and/or products. Carbon-containing compounds (CH_2Br_2, CH_3Br, and CH_4) and Ar were quantified online *via* a gas chromatograph equipped with a GS-Carbon PLOT column coupled to a mass spectrometer (GC−MS, Agilent GC 6890, Agilent MSD 5973N). The effluent gas stream was then passed through two impinging bottles in series, containing a 1 M NaOH aqueous solution, for neutralization prior to its release in the ventilation system. After the catalytic tests, the reactor was quenched to room temperature in He flow, and the catalyst was retrieved for characterization studies.

The conversion of dibromomethane in CH_2Br_2 hydrodebromination, $X(CH_2Br_2)$, was calculated using eq 1

$$X(CH_2Br_2) = \frac{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out}}{n(CH_2Br_2)_{in}} \times 100, \ \% \quad (1)$$

where $n(CH_2Br_2)_{in}$ and $n(CH_2Br_2)_{out}$ are the molar flows of the reactant at the reactor inlet and outlet, respectively. The selectivity, $S(j)$, to product j (j: CH_3Br, CH_4) was calculated according to eq 2

$$S(j) = \frac{n(j)_{out}}{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out}} \times 100, \ \% \quad (2)$$

where $n(j)_{out}$ is the molar flow of product j at the reactor outlet. The turnover frequency, TOF, and reaction rate based on CH_2Br_2 consumption, r, were calculated using eqs 3 and 4, respectively

$$\mathrm{TOF} = \frac{n(CH_2Br_2)_{in} \times X(CH_2Br_2)}{W_{cat} \times \omega_M \times D_M}, \ h^{-1} \quad (3)$$

$$r = \frac{n(CH_2Br_2)_{in} \times X(CH_2Br_2)}{W_{cat}}, \ mol_{CH_2Br_2}\, s^{-1}\, g_{cat}^{-1} \quad (4)$$

where ω_M is the metal loading determined by ICP-OES and D_M is the metallic dispersion and is expressed as

$$D_M = \frac{6 \times \phi_M}{\bar{d} \times \sigma_M} \quad (5)$$

where the area occupied by one surface metal atom is σ_M and the volume occupied by an atom in the metallic state is ϕ_M. The error of the carbon balance, ε_C, used to specify the selectivity to coke, was determined using eq 6

$$\varepsilon_C = \frac{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out} - n(j)_{out}}{n(CH_2Br_2)_{in}} \times 100, \ \% \quad (6)$$

Evaluation of the dimensionless moduli based on the criteria of Carberry, Mears, and Weisz–Prater[32,33] indicated that the catalytic tests were performed in the absence of mass and heat transfer limitations. In addition, CH_2Br_2 hydrodebromination tests over selected catalysts performed at variable flow rates and constant F_T/W_{cat} as well as using catalyst particles of different sizes at constant F_T/W_{cat} verified the absence of extra- and intraparticle mass-transfer limitations (Figure S1), respectively.

2.4. Density Functional Theory. Density functional theory on slab models representing the different metals was employed as implemented in the Vienna Ab initio Simulation Package (VASP 5.4.4).[34,35] Generalized gradient approximation with the Perdew–Burke–Ernzerhof (GGA-PBE) functional was used to obtain the exchange–correlation energies.[36] Projector-augmented wave (PAW) method was chosen to represent the inner electrons, and the valence monoelectronic states were represented by plane waves with a cutoff energy of 450 eV. The Γ-centered k-point mesh was generated through the Monkhorst–Pack method.[37−39] Van der Waals interactions were described *via* the Grimme's DFT-D2 with the reparametrized C_6 values for metals by our group.[40,41] Gas-phase molecules were optimized in a box of 15 × 15 × 15 Å^3. The optimized bulk lattice parameters were 2.9175 Å for Ag, 2.5009 Å (c/a 1.6052) for Co, 2.5609 Å for Cu, 2.8317 Å for Fe, 2.7146 Å for Ir, 2.4752 Å for Ni, 2.7959 Å for Pt, 2.6997 Å for Rh, and 2.7058 Å (c/a 1.5824) for Ru. All metals were modeled by a four-layer $p(3 \times 3)$-(111) face-centered cubic (fcc) slab, with the exception of using a four-layer $p(3 \times 3)$-(110) for Fe and a $p(3 \times 3)$-(0001) for Co and Ru. The top two layers were allowed to relax, while the bottom two were fixed to the bulk lattice in all slabs, which were interspaced along the z-direction by a vacuum space of 15 Å, and the arising dipole was corrected.[42] The thresholds were 10^{-5} eV and 0.03 eV Å^{-1} for electronic and ionic relaxations, respectively. Climbing image nudged elastic band (CI-NEB) method,[43,44] improved dimer method,[45,46] and quasi-Newton algorithms were employed to locate the transition states (TSs) in the reaction profiles, where the TSs were further verified by their single imaginary frequency character.[45] This data constitutes the first step of our computational analysis as shown in Figure 1. All of the structures have been uploaded to the ioChem-BD database.[47−49]

2.5. Statistical Learning Toolbox. Two main techniques, principal component analysis (PCA) and random forest (RF) regressor,[50,51] have been employed in the statistical treatment of the experimental data and DFT results (Figure 1). One of the major issues of DFT simulations for complex reaction networks is the large number of elementary steps (Figure 1, first row) and that reaction profiles cannot be directly used to map activity and selectivity. Descriptors have been traditionally found by a combination of linear-scaling relationships (LSRs, linking the thermodynamics of adsorption of some intermediates to others) and heuristics based on simple chemical concepts (number of bonds and valences).[52] However, this choice is somewhat nonunivocal as it is not based on a rigorous mathematic algorithm (for instance, it does not ensure orthogonality of the different descriptors).[53] This explains why there are dependencies in multidimensional descriptors, particularly in metals. Alternatively, statistical learning techniques provide a mathematically sound framework to identify descriptors. Thus, the dimensionality reduction of the adsorption energies of all of the intermediates involved in the CH_2Br_2 reaction was done *via* an unsupervised statistical learning method to retrieve the principal components from the data set containing 272 DFT-computed intermediate adsorption energies (Figure 1, second row).[29] The outcomes of the analysis are the principal components (linearly uncorrelated variables), in our case, two different energy terms. Thus, the procedure maps the mathematical descriptors to two energies that represent the covalent and redox contributions. More importantly, PCA allows us to cleanse the data to avoid dependencies and therefore focuses the search between the catalytic response and the descriptors.

The next step constitutes the catalytic performance analysis. Ideally, a full microkinetic model based on the DFT-computed parameters would be the response function to be fitted by the descriptors that were identified *via* the principal components. However, the accuracy and robustness of these methodologies prevent us from performing a full *in silico* analysis.[54] Therefore, the experiments are taken as input and understood with the descriptors obtained computationally. The simplest approach would be to utilize a classifier employing, for instance, K-means, which is popular in data mining, or clustering techniques. However, these methods fail when applied to understand selectivity problems since very small variations in the energy scale of intermediates can lead to a complete switch of the selectivity.[28,29] Alternatively, a simple classifier such as the decision tree (DT) can be employed by answering a list of simple energy-related questions for the two descriptors (vide

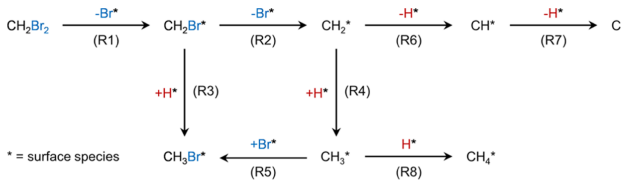


Figure 2. Reaction network of CH_2Br_2 hydrodebromination showing the pathways leading to CH_3Br, CH_4, and C that accounts for coke formation. Arrow labels indicate species involved in the reaction, while the labels in parentheses indicate thermodynamic and kinetic parameters that are detailed in Tables S1−S7.

infra) as shown in the third step in Figure 1. However, these models tend to overfit and alternatives have been proposed in the literature to mitigate this drawback.[50] Therefore, we opted for an ensemble learning method that allows simultaneous classification and regression, thereby overcoming the previously mentioned issues (Figure 1, bottom row). The RF[51] technique operates by constructing an ensemble of decision trees with different seeds, thereby limiting overfitting. Hereafter, the average of the forest is taken as the outcome of the statistical analysis, although a few disclaimers are needed. Typically, random forests are applied to larger data sets and are benchmarked *via* cross-validation. Even though the sets derived from nanoparticles of metals of comparable size are limited, the filtering introduced with the PCA ensures that the results are qualitatively sound. Finally, the analytical functions for the experimental activity, selectivity, and yield have been identified through the Bayesian Machine Scientist (BMS).[55] The algorithm searches for candidate functions that describe the behavior of a given data set using the Markov chain Monte Carlo (MCMC) method. At every Monte Carlo step, the algorithm evaluates the quality of the current function using a complexity parameter, based on the entanglement of the mathematical operators, and an error parameter that depends on the sum of the square estimate of the errors (SSEs). Filtering these values makes it is possible to obtain simple and accurate equations that properly describe the nature of the data set. In our case, the data set contains both the experimental results of activity, selectivity, and yield and the PCA that provides the descriptors from the DFT part. Expansion of the data to variable nanoparticle sizes, speciation, alloys, and intermetallics is beyond the scope of this study but would constitute a natural extension to this work.

3. RESULTS AND DISCUSSION

The state-of-the-art catalytic research work normally encompasses the experimental testing of a few materials, typically reporting only the best hits in terms of activity and selectivity and the study of the reaction network presenting the list of elementary steps and the reaction profiles provided by DFT. Advanced methodologies would include the use of LSR to make a dimensionality reduction that provides one or two descriptors (in some cases, the selection is nonunivocal). A kinetic model based on the particular mechanism found for a single catalyst (typically, a metal) is then simplified *via* a rate-determining-step concept while considering the experimental conditions. Applying the LSR, the rate is simplified to be a function of a single descriptor parameter, leading to what are known as volcano plots.[28] Screening for different materials is then done by computing energies of the descriptor over a family of materials. However, this approach presents a few pitfalls: (i) the selection of the descriptor is nonunivocal; (ii) the probability that there are hidden dependencies is high when there is more than one descriptor; as a consequence, multidimensional analysis (considering more than one descriptor) typically ends up with the activity of all metals falling along the same line; (iii) selectivity is difficult to track due to the small energy differences involved in the selectivity switches (cliffs); and (iv) the studies of stability are introduced in a separate step as a filter to the overall results.

In this work, we have taken an alternative route to combine both experimental and theoretical results to avoid some of these bottlenecks by the generation of an extensive database of catalytic materials, using their activity and selectivity, and employing DFT data as an independent source for the performance descriptors through the inference *via* statistical learning techniques. To this end, the dimensionality reduction of the energies for intermediates is performed *via* a PCA and the random tree and random forest methodologies are employed as a classifier and regressor, respectively, to understand and determine the key binding energies (BEs) for the descriptors that limit the areas with higher activity and selectivity. Formulae for the different response functions, activity, selectivity, and yield, are found using the Bayesian machine scientist employing the computed descriptors.

3.1. Mechanistic Studies and Dimensionality Reduction. The DFT calculations revealed the reaction profiles of CH_2Br_2 hydrodebromination over the nine metals (Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir, and Pt; see Figure S2). The energies for the kinetic and thermodynamic parameters of the reactions shown in Figure 2 are presented in Tables S1−S7 and in the ioChem-BD database, where they can be downloaded as a csv file.[48,49] Three main reasons led to the use of pure metal surfaces as a starting point in the reactivity studies: (i) they allow a systematic investigation while keeping the nature of the materials constant (and thus equivalent electronic structures); (ii) they can provide the fundamentals for the phase transition during the induction process and explain the sources of catalyst instability simultaneously; and (iii) they constitute the simplest type of DFT calculations. The starting point for obtaining the reaction profiles over all nine metals was the mapping of the full path over the Ru surface (Figure 3). The potential transition-state structures on the other metal surfaces (Figure S2) were inherited from that of Ru by applying a previously reported algorithm that uses the potential seed TSs as input for an improved dimer method or a CI-NEB refinement, after which the final structure is confirmed through vibrational analysis.[56] This procedure allowed us to reduce the computational time as the intermediates along the reaction do not

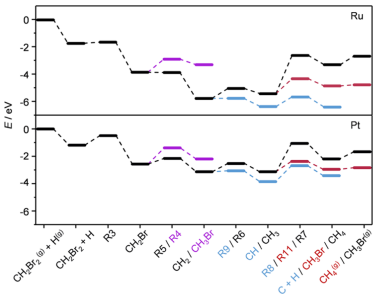


Figure 3. Energy profiles of CH_2Br_2 hydrodebromination on ruthenium and platinum surfaces.

require large optimizations and even the transition states could be directly located from the Ru seed.

The reaction profiles show the sequential elimination of either Br and H atoms and the recombination of the carbonaceous fragments with the main products (CH_4 and CH_3Br) and surface coke illustrated by a carbon precursor. This sequential mechanism exists in all cases except for Fe (Figure S2), on which the CH_2Br intermediate cannot be located. The large reactivity of Fe with respect to the reactants is responsible for the direct decomposition of CH_2Br_2, losing its both Br atoms on the surface. To identify the activity descriptors, a PCA was performed by taking all of the 272 intermediate energies of CH_xBr_y and the isolated atoms H, C, and Br. For all species, several potential adsorption sites were investigated.[29] In the case of CH_2Br_2, physisorption and dissociative adsorption of the C−Br bond occurs on the metal surface. Ultimately, the values of the principal components are 92.8, 5.2, and 1.2%. Therefore, two principal components account for 98% of the adsorption energies and are thus taken as relevant. Then, to unravel the descriptors, we adopted the strategy of mapping these principal component terms to the energies of some of the smallest fragments in the reaction network. We have followed the same mathematical procedure as in our previous work,[29] where the energies of the fragments are correlated with the PC taking the ones showing the lowest error in the prediction. In a second step, we map the PC to the intermediate energies that are better represented by a single PC. When doing so, the first and second PCs were determined as the CH and Br species on the hcp sites, whereas their adsorption energies were considered the descriptors for the covalent (CH) and redox (Br) terms (Figures S3 and S4), respectively, due to their exclusive contribution to these principal components. The role of coverage effects on these parameters for the most active catalyst is shown in Figure S5.

3.2. Catalytic Performance in CH_2Br_2 Hydrodebromination. Gas-phase CH_2Br_2 hydrodebromination was investigated over Fe-, Co-, Ni-, Cu-, Ru-, Rh-, Ag-, Ir-, and Pt-based catalysts supported on SiO_2 (1.0 wt % metal basis), which were chosen based on previously reported CH_2Br_2 hydrogenation studies.[15] The inertness of SiO_2 in hydrodehalogenation reactions and its minimal interaction with the active phase allow studying the intrinsic catalytic performance of each transition metal. Synthesis of the catalysts was done by incipient wetness impregnation, involving a calcination step in air followed by reduction in H_2 prior to their exposure to reaction conditions. The reduction temperature was based on H_2-TPR analysis and was chosen as such to ensure the formation of the metallic phase (Figure S6). Characterization data of the catalysts revealed the close similarity of the specific surface areas (S_{BET}, 244−258 $m^2\ g^{-1}$) and pore volumes (V_{pore}, 0.71−0.83 $cm^3\ g^{-1}$) (Table 1). XRD analysis of the catalysts

Table 1. Characterization Data of the Catalysts

		S_{BET}[b] ($m^2\ g^{-1}$)		V_{pore}[c] ($cm^3\ g^{-1}$)	
catalyst	metal loading[a] (wt %)	fresh	1 h (10 h)	fresh	1 h (10 h)
Fe/SiO_2	1.1	254	254	0.77	0.76
Co/SiO_2	0.9	258	250	0.79	0.78
Ni/SiO_2	1.0	254	252 (240)	0.75	0.80 (0.71)
Ni/SiO_2-NC	1.0	251	244	0.77	0.79
Cu/SiO_2	0.9	255	251	0.78	0.77
Ru/SiO_2	1.0	247	254 (246)	0.71	0.73 (0.70)
Ru/SiO_2-NC	1.0	249	240	0.83	0.80
Rh/SiO_2	1.0	250	220 (228)	0.77	0.72 (0.70)
Ag/SiO_2	0.9	250	249	0.76	0.72
Ir/SiO_2	1.0	254	253 (246)	0.76	0.73 (0.72)
Pt/SiO_2	1.0	244	242 (245)	0.77	0.75 (0.73)

[a]ICP-OES. [b]BET model. [c]Volume of N_2 adsorbed at $p/p_0 = 0.98$.

showed diffraction peaks compatible with the metallic phases, whereas no evidence of the oxide phases was observed (Figure 4a). Further confirmation of the elemental composition was provided by ICP-OES, which showed that the metal content in the materials was approximately the targeted 1.0 wt %. The structural information gained from the utilization of XRD was corroborated by Raman spectroscopy (Figure 4b). All catalysts displayed low-frequency bands centered at 490 and 603 cm^{-1}, which are characteristic of the SiO_2 support.[57,58] In addition, the Raman spectrum of Ir/SiO_2 evidenced bands at 548 and 716 cm^{-1}, which could be ascribed to a minor IrO_2 phase,[59,60] and the spectrum of Cu/SiO_2 displayed a band at 972 cm^{-1}, suggesting that a Cu_xSiO_y phase is likely present in the respective supported Ir and Cu catalysts.[61,62] The presence of Cu_xSiO_y could explain the reduction peak at *ca.* 500 K observed in H_2-TPR (Figure S6).[62]

The catalysts were evaluated in CH_2Br_2 hydrodebromination at different reaction temperatures (423−623 K). Assessment of the hydrodebromination activity, expressed as CH_2Br_2 conversion, at 523 K allowed derivation of the following order for the respective supported catalyst: Fe ≈ Co ≈ Cu ≈ Ag (4−7%) < Ni (11%) < Ru (19%) < Rh (32%) < Ir ≈ Pt (50−52%) (Figures 5a and S7). Evaluation of the product selectivities at 523 K and *ca.* 20% CH_2Br_2 conversion achieved by adjustment of the space velocity is shown in Figure 5b. The performance of Fe, Co, Cu, and Ag showed consistency with the generally reported inferior hydrogenation activity of these elements compared to the platinum group metals. From a cost perspective, it is interesting that Ni displays a selectivity pattern comparable to that of Rh.

3.3. Classification of Metal Catalysts Using the Random Forest Regressor. Four clusters are classified based on the experimentally determined product distribution of the SiO_2-supported metal catalysts: (i) poor hydrodebromination activity over Fe, Co, Cu, and Ag, with coke

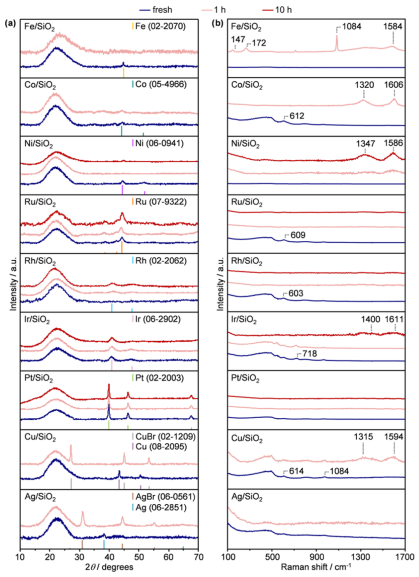


Figure 4. (a) XRD patterns and (b) Raman spectra of the catalysts in a fresh form and after CH_2Br_2 hydrodebromination. Reference diffraction patterns are shown as vertical lines below the measured diffractograms and are identified with their ICDD-PDF numbers. Reaction conditions: $CH_2Br_2/H_2/Ar/He$ = 6:24:4.5:65.5, F_T/W_{cat} = 25−150 $cm^3\ min^{-1}\ g_{cat}^{-1}$, T = 523 K, P = 1 bar, and time on stream (tos) = 1 or 10 h.

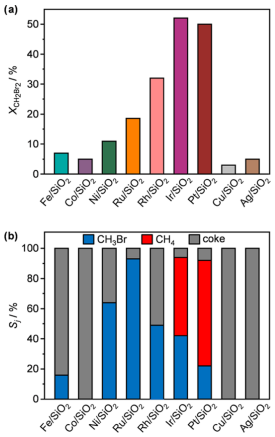


Figure 5. (a) Conversion of CH_2Br_2 and (b) product selectivity of the catalysts in CH_2Br_2 hydrodebromination. In (a), the conversion was assessed at a constant space velocity of F_T/W_{cat} = 40 $cm^3\ min^{-1}\ g_{cat}^{-1}$, while product selectivities in (b) were determined at *ca.* 20% CH_2Br_2 conversion achieved by adjusting the space velocity in the range of F_T/W_{cat} = 20−150 $cm^3\ min^{-1}\ g_{cat}^{-1}$. Other reaction conditions: $CH_2Br_2/H_2/Ar/He$ = 6:24:4.5:65.5, T = 523 K, P = 1 bar, and tos = 0.25 h.

as the main product. (ii) Intermediate activity and selectivity to CH_3Br (<55 and <58%, respectively) over Ni and Rh, coupled with pronounced coke formation. Worth mentioning is the absence of CH_4 generation over these two groups of catalysts at any reaction temperature applied (Figure S7). (iii) Great propensity to CH_4 (>48 and >61%, respectively) besides producing CH_3Br (<28 and <39%, respectively) over Ir and Pt with minor coking. (iv) The highest selectivity to CH_3Br (<96%) over Ru with coke (<23%) as the byproduct (notably higher than any other metal). The qualitative analysis of the main product with a single decision tree shows that only two decisions are needed to classify all of the observations (Figure 1): the main decision corresponds to the CH binding energy, while the second is the binding energy of Br, in agreement with the PCA.

Then, experimental performance can be analyzed in terms of the PC through the random forest technique. Therein, the experimentally determined CH_2Br_2 conversion, CH_3Br yield, CH_3Br selectivity, and the two descriptors obtained by PCA were taken as inputs (Figure 1). A random forest algorithm composed of 128 trees was trained to sample the phase space spanned by the descriptors.[50] Tests with different tree lengths and number of trees are presented in the SI (Figure S8). The accuracy of the training set is 0.91 for the conversion, 0.79 for the selectivity, and 0.87 for the yield. This allows the partition of the performance of the different catalysts just by taking the properties of the naked metal surfaces without considering potential phase transformations (Figure 6).

For the highest conversion, the optimal CH adsorption energy ranges from −6.50 to −6.75 eV (with respect to the $CH^{(g)}$ fragment), while the ideal Br adsorption energy requires a weak M−Br bond (Figure S2). This explains the inactive behavior of Co, Cu, and Ag, since they hold either far-too-strong or far-too-weak CH adsorption regions. The selectivity frontiers are different from the activity ones, as previously described for general reactivity models.[28] In particular, selectivity toward CH_3Br requires a weaker CH and stronger Br adsorption energy than the one for activity, with ranges from −6.30 to −6.75 eV and from −2.2 to −2.3 eV, respectively. It should be emphasized that there is a narrow selectivity window for CH_3Br, which spans only 0.1 eV along the Br binding energy direction. Ru, with the best catalytic behavior, is close to the sweet spot, and Ir and Pt are in the weak binding region (Figure 6). Finally, the performance of Fe sets up a new region due to its different packing way that leads to the variations of surface metal positions from the standard fcc metals. As a result, the descriptors correspond to very exothermic energies and the area, likely to be shared with other early transition metals, would correspond to a poorly active region.

3.4. Identification of the Functional Forms for the Catalytic Activity of the Metal Catalysts Using the Bayesian Machine Scientist (BMS). The RF methods present three main drawbacks: (i) their black-box nature masks the physical interpretability of the trained model (a

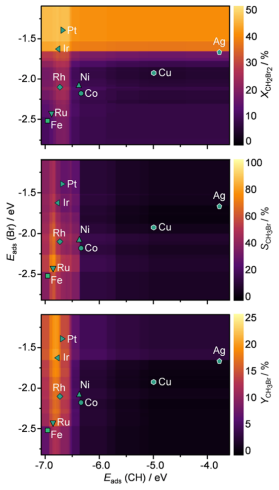


Figure 6. CH_2Br_2 conversion, CH_3Br selectivity, and CH_3Br yield for different values of E_{ads}(CH) and E_{ads}(Br) obtained with a random forest regressor algorithm containing 128 trees.

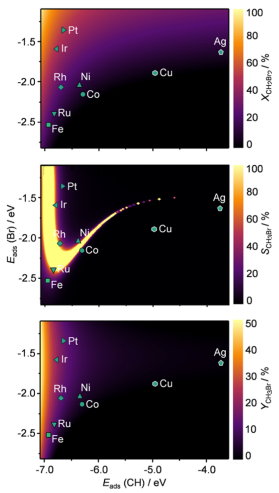


Figure 7. CH_2Br_2 conversion, CH_3Br selectivity, and CH_3Br yield for different values of E_{ads}(CH) and E_{ads}(Br) obtained with the Bayesian machine scientist using the functions: $X = -(E_{ads}(Br_{hcp}) - c_{c_1})\cdot c_{c_2} + \frac{c_{c_3}}{E_{ads}(CH) + c_{c_4}}$ for the conversion, $S = -((E_{ads}(Br) - c_{s_1})\cdot(E_{ads}(CH) + c_{s_3})^{E_{ads}(CH)+c_{s_3}+1} + (E_{ads}(CH) + c_{s_3}) + c_{s_2})^{-2}$ for the selectivity, and $Y = -(E_{ads}(Br) + c_{y_1})^2\cdot c_{y_2} + \frac{c_{y_2}}{(E_{ads}(CH) + c_{y_3})^2}$ for the yield. Fe was not included during the functional form search.

numerical value with an error is retrieved as a response), (ii) the range and accuracy are closely related to the number of samples used in the training phase, and (iii) extracting a functional form is difficult. To overcome these issues, BMS[55] has been employed to search for general functional forms (Figure S9 and Tables S9–S11) that describe the response functions (experimental conversion, selectivity, and yield) for CH_2Br_2 hydrodebromination taking DFT-PCA descriptors as variables. To homogenize our data set, Fe has been excluded from the input data set due to its different atom packing.

Figure 7 shows the best functions obtained for each experimental parameter and their graphical representation. Conversion has the simplest equation, displaying a direct dependence on the adsorption energy of Br and an inverse dependence on the adsorption energy of CH. These dependencies point to a possible surface Br contamination in metals with the most exothermic Br–metal bonds. Contrarily, conversion improves at lower values of CH adsorption energy. Significantly, the equation for the yield is the squared of the conversion. Therefore, it describes a volcano-like area with the highest response values at the most exothermic CH adsorption and intermediate values of Br adsorption energies. Finally, the function obtained to represent the selectivity is rather complex and hard to interpret. However, its graphic representation shows a cliff cutting the area between the metals that coke and those that do not, correctly reproducing the opposite behavior of Co and Ni. Table S12 contains the fitting parameters of the found functions. Additionally, Figure S10 shows an additional simple functional form found for selectivity.

Figure 8 shows the obtained SSE comparing the experimental and predicted values using the two regressors. BMS is the method that achieves the best accuracy for all of the experimental values. However, the dependency of RF on the number of samples and the complexity of the selectivity equation obtained with BMS leads us to think that for rough areas such as selectivity, RF will perform a better prediction with denser samples. Contrarily, BMS can provide simple models with a limited number of observations, as can be seen for conversion and yield. Additional information about prediction errors can be found in Table S8.

3.5. Stability and Characterization of the Used Catalysts. Establishing typical product distributions for each metal naturally raises questions as to how the materials behave upon exposure to reaction conditions for longer times. Therefore, product-based trends identified were complemented with short-term (10 h) CH_2Br_2 hydrodebromination tests over the active SiO_2-supported catalysts (Ni, Rh, Ir, Pt, and Ru), evidencing deactivation of all catalytic systems with *ca.* 25–80% loss of initial activity (Figure S11). The following order of decreasing stability over 10 h was found: Rh ≈ Ru ≈ Ni < Ir ≈ Pt. Therein, Ni falls out of the trends by showing a significant loss of CH_2Br_2 conversion in the first 2 h on stream and displaying relatively stable performance for the remainder of the reaction time. On the other hand, Ru follows a similar deactivation profile as Rh, gradually losing activity over time. A moderate decrease in performance was achieved over Ir and Pt, showing *ca.* 1.5–3.5 times more activity compared to the rest of the metal catalysts. Depletion of activity was appended by moderate changes in the product distribution. With time on

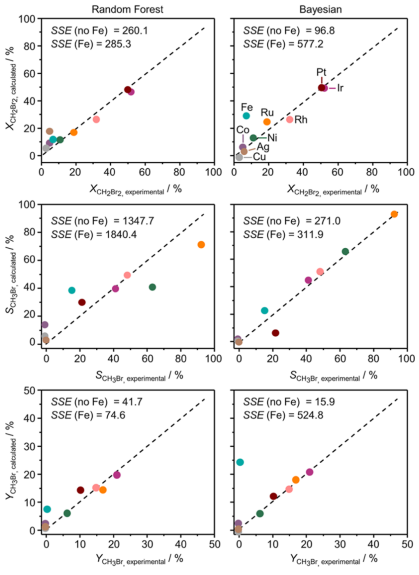


Figure 8. Sum of squared estimate of errors (SSEs) comparing the experimentally obtained CH_2Br_2 conversion, CH_3Br selectivity, and CH_3Br yield with their predicted values using the Bayesian machine scientist and the trained random forest regressor.

stream, the selectivity to CH_3Br progressively decreased over Rh, Ir, and Pt, whereas a slight increase was observed over Ni and Ru. The propensity to form CH_4 was enhanced over Ir and Pt, while changes in coke production were mainly observed over Ni and Rh. These findings emphasize that catalyst robustness remains a challenge, for which understanding the origin of deactivation is essential to develop superior CH_2Br_2 hydrodebromination catalysts.

SiO_2-supported metals that exhibit poor activity (Fe, Co, Cu, and Ag) were characterized after 1 h on stream, whereas analysis of the catalytically active systems (Ni, Ru, Rh, Ir, and Pt) was performed after 1 and 10 h use in CH_2Br_2 hydrodebromination. In particular, N_2-sorption, ICP-OES, XRD, and Raman spectroscopy were applied to study the development of the materials during exposure to the reaction environment. Examination of the textural properties by N_2-sorption showed minimal differences in the specific surface areas (S_{BET}) and pore volumes (V_{pore}) of the used (1 and/or 10 h) and fresh samples (Table 1). Furthermore, quantification of the metal content in used systems by ICP-OES analysis pointed toward the preservation of the metal content (1.0 wt %) in all catalysts, indicating the absence of active phase leaching or volatilization. Further investigation revealed three main deactivation mechanisms: (i) bromination, (ii) fouling by coking, and (iii) active phase restructuring.

Generally, the presence of a halogen introduces structural stability challenges for many catalysts.[63] Analysis of the Ag/SiO_2 and Cu/SiO_2 samples confirmed the severe effects of bromination on activity. In line with the literature, the poor performance of Ag could be explained by its rapid oxidization to AgBr as observed in XRD (Figure 4a).[15] A similar behavior was observed for Cu, which was promptly restructured to CuBr. The XRD reflections of used Co/SiO_2 resembled those of the fresh material, suggesting the absence of active phase sintering or extensive bromination. In contrast, the disappearance of the reflection assigned to metallic iron in Fe/SiO_2 points toward restructuring after exposure to reaction conditions. Complemented with Raman analysis (Figure 4b), the spectra of used Fe/SiO_2 evidenced minor bands at 147 and 272 cm^{-1}, which could be attributed to Fe−Br and Fe−C stretching modes, respectively, with the latter most likely due to the presence of Fe_xC_y phases.[64−67] Moreover, a strong band was detected at 1084 cm^{-1}, which relates to the asymmetric stretching mode of Fe−O−Si bonds in tetrahedrally coordinated Fe species.[68] Altogether, Raman analysis suggests the prompt restructuring of metallic Fe to $FeBr_x$, Fe_xC_y, and FeO_xSi_y under hydrodebromination conditions, rendering the catalyst very complex (in agreement with the difficulties found in the simulations) and mostly inactive. In addition, the Raman spectra of used Fe/SiO_2, Cu/SiO_2, and Ni/SiO_2 display the well-documented D and G bands at *ca.* 1320 and 1585 cm^{-1}, which are ascribed to graphitic species and normal vibrations in graphene, respectively.[69,70] In addition to the D band, Co/SiO_2 and Ir/SiO_2 show a significant peak at *ca.* 1606 cm^{-1}, commonly denoted the D′ band, signifying the presence of imperfect graphite or disordered carbon.[71] This indicates that coke formation, the main deactivation mechanism over Co/SiO_2 and Ir/SiO_2, follows a different pathway compared to Fe and Cu.

In addition to the previously mentioned characterization techniques, XPS and HAADF-STEM microscopy were adopted to further study the catalytically active systems that were exposed to 10 h hydrodebromination conditions (Figure S11). Therein, HAADF-STEM observations disclosed the dispersion of Ni, Ru, Rh, Ir, and Pt nanoparticles on SiO_2. Rh- and Pt-based materials displayed a narrow size distribution, whereas a wide size distribution was found for Ni, Ru, and Ir-based systems (Figure 9). Although the micrographs indicate that Pt/SiO_2 exhibits an average particle size distribution of 2.2 nm, a few particles in the range of 18−28 nm were distinguished, too. Moreover, large Pt crystallites of up to *ca.* 100 nm could be appraised quantitatively from XRD analysis, providing this numerical estimation based on the Scherrer equation (Figure 4a). Compared to their fresh analogues, used Ni-, Ir-, and Pt-based systems showed small changes in the average metal nanoparticle size. In contrast, pronounced active phase sintering in Ru/SiO_2 and Rh/SiO_2 strongly reduced the fraction of surface metal atoms after 10 h on stream by *ca.* 30 and 40%, respectively, demonstrating that this deactivation mechanism occurs on various nanoparticle-based catalysts regardless of the halogen source.[14] The systems were further analyzed by XPS to assess the presence of brominated metal species on the surface of the used materials, revealing that bromination plays a limited role in catalyst deactivation (Figure 10). The Ru 3d, Rh 3d, Ir 4f, and Pt 4f core-level spectra of the fresh and used systems displayed pronounced peaks at binding energies (BEs) of *ca.* 279.1, 306.9, 60.4, and 70.7 eV, respectively, corresponding to the metallic phase.[72−74] Careful fitting of the spectra revealed contributions at BEs of 72.2 and 74.0 eV in Pt/SiO_2, at 61.9 eV in Ir/SiO_2, and at

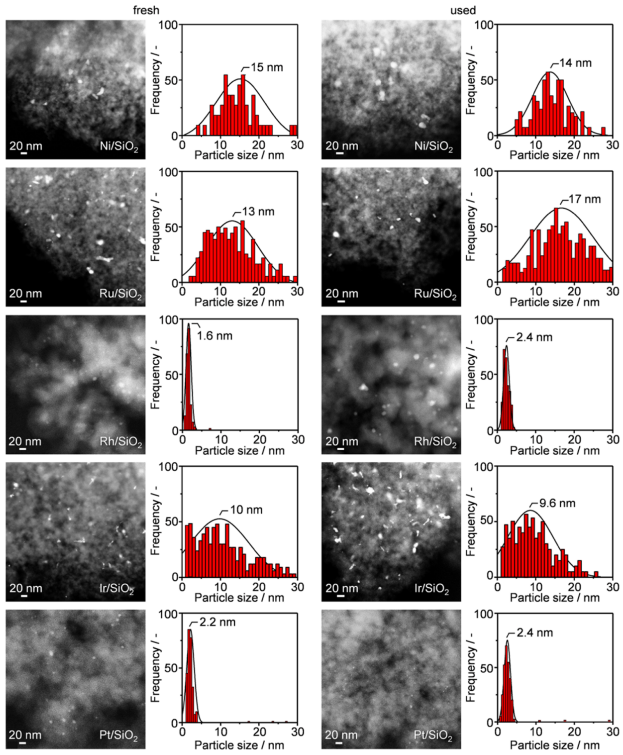


Figure 9. HAADF-STEM micrographs and derived particle size distributions of selected catalysts in a fresh form and after 10 h in CH_2Br_2 hydrodebromination. The conditions specified in Figure 4 apply here.

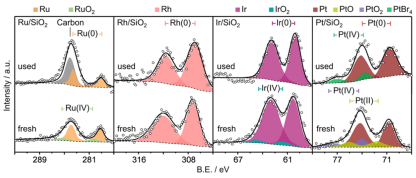


Figure 10. Ru 3d, Rh 3d, Ir 4f, and Pt 4f XPS core-level spectra of selected catalysts in a fresh form and after 10 h in CH_2Br_2 hydrodebromination. The solid lines and the open circles represent the overall fit and the raw data, respectively, while the colored areas beneath them indicate the different contributions. The conditions specified in Figure 4 apply here.

280.6 eV in Ru/SiO_2, all designated to oxidized species,[72,73,75] likely due to exposure of the sample to air. Minor active phase bromination (<15%) was observed over used Pt/SiO_2, as indicated by the peak at a BE of 73.8 eV, which is ascribed to $PtBr_4$ species.[76] The presence of carbon species (*ca.* 30%) was detected over used Ru/SiO_2, as evidenced by the strong contribution at a BE of 284.2 eV. The Ni-based system suffered from the formation of an oxidized Ni layer formed by contact of the sample with air, a process well-studied in literature.[77] Sputtering prior to XPS analysis would remove potential Br species present on the surface, preventing any realistic comparison with other catalysts. The use of operando spectroscopic techniques could provide more insights into the deactivation phenomena. However, the application of such methods requires the design of adequate cells resistant to the corrosive nature of this reaction, which will be the subject of future investigations.

3.6. Kinetic Analysis. To further compare and benchmark the performance of the catalytically active systems, TOF values of selected materials were determined (Figure 11a). Ni/SiO_2, Ru/SiO_2, and Pt/SiO_2 were chosen as representative systems of their respective performance group. The dispersion was

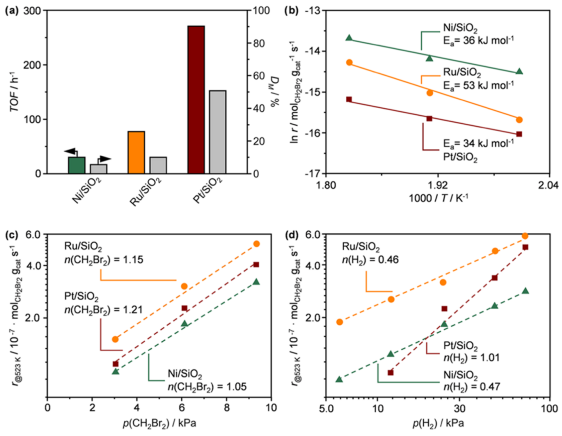


Figure 11. (a) Turnover frequencies and rates of CH_2Br_2 hydrodebromination over selected catalysts as a function of (b) temperature and inlet partial pressures of (c) CH_2Br_2 and (d) H_2. Each catalytic data point was gathered using materials in a fresh form to exclude the possible influence of catalyst deactivation. Reaction conditions: (a) $CH_2Br_2/H_2/Ar/He$ = 6:24:4.5:65.5, F_T/W_{cat} = 40 $cm^3\ min^{-1}\ g_{cat}^{-1}$, and T = 523 K; (b) $CH_2Br_2/H_2/Ar/He$ = 6:24:4.5:65.5, F_T/W_{cat} = 40−200 $cm^3\ min^{-1}\ g_{cat}^{-1}$, and T = 498−548 K; (c) $CH_2Br_2/H_2/Ar/He$ = 3−9:24:4.5:62.5−68.5, F_T/W_{cat} = 40−200 $cm^3\ min^{-1}\ g_{cat}^{-1}$, and T = 523 K; and (d) $CH_2Br_2/H_2/Ar/He$ = 6:6−72:4.5:17.5−83.5, F_T/W_{cat} = 40−200 $cm^3\ min^{-1}\ g_{cat}^{-1}$, and T = 523 K. All tests were conducted at P = 1 bar and tos = 0.25 h.

calculated based on the average metal nanoparticle size, assuming a hemispherical geometry of the metal site as observed in the micrographs (Figure 9). The activity of the catalysts decreased in the following order: $Pt/SiO_2 \gg Ru/SiO_2 > Ni/SiO_2$, with dispersions of 51, 10, and 7%, respectively, confirming the previously determined trends (Figure 5a). As shown in Figure 11a, Pt/SiO_2 displays TOF values *ca.* 1 order of magnitude higher than those observed over the other systems. The dispersion values of Ni/SiO_2 and Ru/SiO_2 differ significantly from that of Pt/SiO_2. To eliminate a possible influence of this parameter, reference catalysts with comparable active phase dispersion were synthesized by omitting the calcination step during the synthesis to obtain Ni/SiO_2-NC and Ru/SiO_2-NC. XRD analysis and HAADF-STEM microscopy confirmed the attainment of systems lacking large crystallites and with an average nanoparticle size of 2 nm for both Ni and Ru (Figures S12 and S13), resulting in a metal dispersion of 48 and 56%, respectively (Figure S14). Notably, for the Ru- and Ni-based systems, selectivity patterns as well as CH_2Br_2 conversion remained unchanged (Figure S14). Consequently, the TOF values of Ni/SiO_2-NC and Ru/SiO_2-NC were lower than those of Ni/SiO_2 and Ru/SiO_2, thereby stressing the outstanding hydrodebromination activity attained over Pt/SiO_2 (Figure 11a). Moreover, these results suggest that activity over Ni- and Ru-based systems is structure dependent. The impact of active phase nanostructuring on CH_2Br_2 hydrodebromination performance deserves attention in future dedicated studies.

Further insights were gained by conducting kinetic analysis over the three systems, showing differences in the apparent activation energies with values of 53 kJ mol^{-1} (Ru/SiO_2), 36 kJ mol^{-1} (Ni/SiO_2), and 34 kJ mol^{-1} (Pt/SiO_2) (Figure 11b). On the other hand, similarities in the derived partial orders with respect to CH_2Br_2 were found, with values ranging between 1.05 and 1.21 (Figure 11c). Particularly interesting are the relatively low reaction orders in H_2 for Ni/SiO_2 and Ru/SiO_2 of *ca.* 0.47, deviating from that of a system that mainly produces CH_4, which exhibited a partial order of 1.01 (Figure 11d). These kinetic fingerprints are a direct consequence of the observed patterns described in the reaction profiles (Figures 3 and S2). For the reactant, the reaction order is around 1, which is in line with the stoichiometric term in the general equation. The same applies to the H_2 dependencies, where the production of CH_4 results in a partial order of *ca.* 1 if the reaction of the second H-atom is considered as rate determining (Figure 2). On the other hand, the production of CH_3Br requires a single H-atom and therefore shows the observed partial order in H_2 of *ca.* 0.5. Comparable results were found over nanostructured catalysts for selective CH_2Cl_2 hydrodechlorination, where a kinetic model could give an account of the possible origin of the observed selectivity differences.[14] The effects of inlet partial pressures of CH_2Br_2 and H_2 were further studied to determine the effect of these compounds on product distribution (Figure 12). Therein, adjusting the CH_2Br_2 inlet partial pressure showed a little influence on the selectivity patterns. On the other hand, changing the partial pressure of H_2 from 6 to 72 kPa had a significant effect on product distributions and activity. With increasing H_2 concentration in the feed, a higher selectivity to CH_3Br was obtained over Ru/SiO_2 and Ni/SiO_2. In contrast, a slight decrease of CH_3Br selectivity at the expense of CH_4 formation was observed over Rh/SiO_2 and Pt/SiO_2. Over all systems, coke formation was curbed at the expense of the production of either CH_3Br or CH_4 with increasing H_2 partial pressures.

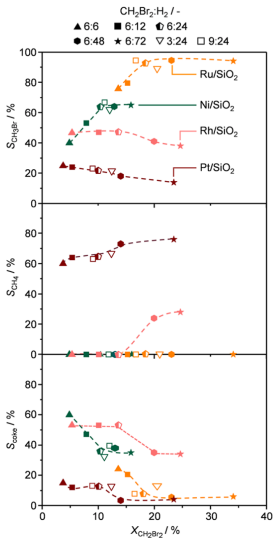


Figure 12. Selectivity to CH_3Br, CH_4, and coke in CH_2Br_2 hydrodebromination as a function of CH_2Br_2 conversion over selected catalysts. The shape and interior of the symbols provide information on the feed composition. Each catalytic data point was gathered using materials in a fresh form to exclude the possible influence of catalyst deactivation. Reaction conditions: $CH_2Br_2/H_2/Ar/He$ = 3−9:6−72:4.5:17.5−83.5, F_T/W_{cat} = 40 200 $cm^3\ min^{-1}\ g_{cat}^{-1}$, T = 523 K, P = 1 bar, and tos = 0.25 h.

4. CONCLUSIONS

A strategy combining computational techniques and catalyst testing was devised with the aim to understand the performance of selected metal catalysts in the hydrodebromination of CH_2Br_2 to CH_3Br. Here, we have derived a consistent catalytic data set comprising experimentally obtained and DFT data for the application of statistical techniques. The steady-state catalytic tests of metals supported on SiO_2 revealed four performance groups comprising (i) poorly active Fe, Co, Cu, and Ag; (ii) Rh and Ni, which show intermediate selectivity to CH_3Br (<60%) but do not generate CH_4; (iii) Ir and Pt, which mainly produce CH_4 (>50%); and (iv) Ru, which exhibits the highest selectivity to CH_3Br (>96%). DFT was applied to retrieve the energy profiles over the metals, after which the binding energies of the 272 intermediates were subjected to dimensionality reduction *via* principal component analysis, a robust mathematical construct. The two descriptors obtained from this unsupervised method were, together with the experimental data, employed in the random forest regressor and the Bayesian machine scientist, ultimately connecting the descriptor energy intervals with catalytic activity or selectivity and obtaining the functional forms for the identification of performance trends in terms of CH_3Br yield. This work addresses important aspects in machine-learning-aided research, mainly (i) the use of integrated and complementary experimental and computational first-principles results, (ii) the identification of hot activity/selectivity spots through robust mathematically and nonbiased methodologies, and (iii) extraction of physically meaningful mathematical expressions to describe the performance of the catalytic systems. Ideally, these methodologies shall be able to identify new candidates and verify them experimentally. In practice, synthetic methods for different metals can end up producing different species. Therefore, the particular active site speciation, nanoparticle size, coordination, and existence of nonremoved ligands could affect the final performance and would require a much more dedicated analysis and a better understanding of the synthetic protocols. However, our approach lays the foundations for future studies targeting the full ab initio prediction of catalytic performance.

■ ASSOCIATED CONTENT

Assessment of the absence of extraparticle and intraparticle mass-transfer limitations; elementary steps of catalyzed CH_2Br_2 hydrodebromination; reaction barriers, reaction energies, and adsorption energies; imaginary vibrational frequencies; reaction energies for the $C_{hcp} \rightarrow C_{subsurface}$ reaction; energy profiles for CH_2Br_2 hydrodebromination; comparison between the two main components obtained using PCA analysis and the DFT adsorption energies of Br_{hcp} and CH_{hcp}; ratio and difference between the DFT adsorption energies of Br_{hcp} and CH_{hcp}; H_2-TPR profiles of the SiO_2-supported metal oxides; conversion of CH_2Br_2 as a function of temperature and product selectivity as a function of CH_2Br_2 conversion; accuracy with the training set for various random forest regressors containing a different number of trees; schematic representation of a BMS tree; sum of squared errors (SSE), mean squared error (MSE), root MSE (RMSE), fitting functions, and fitting constants; comparison between the calculated and experimental data using the CH_3Br selectivity fitting function; stability tests; XRD patterns, HAADF-STEM micrographs, and derived particle size distributions of the noncalcined catalysts; CH_2Br_2 conversion, product selectivity, and turnover frequency over the noncalcined catalysts (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

N. López − *Institute of Chemical Research of Catalonia, ICIQ, The Barcelona Institute of Science and Technology, 43007 Tarragona, Spain;* orcid.org/0000-0001-9150-5941; Email: nlopez@iciq.es

J. Pérez-Ramírez − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland;* orcid.org/0000-0002-5805-7355; Email: jpr@chem.ethz.ch

Authors

A. J. Saadun − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland*

S. Pablo-García − *Institute of Chemical Research of Catalonia, ICIQ, The Barcelona Institute of Science and Technology, 43007 Tarragona, Spain*
V. Paunović − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland;* orcid.org/0000-0001-6630-1672
Q. Li − *Institute of Chemical Research of Catalonia, ICIQ, The Barcelona Institute of Science and Technology, 43007 Tarragona, Spain;* orcid.org/0000-0001-5568-2334
A. Sabadell-Rendón − *Institute of Chemical Research of Catalonia, ICIQ, The Barcelona Institute of Science and Technology, 43007 Tarragona, Spain*
K. Kleemann − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland*
F. Krumeich − *Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland;* orcid.org/0000-0001-5625-1536

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.0c00679

Author Contributions

[§]A.J.S. and S.P.-G. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the ETH Research Grant ETH-43 181 and Ministerio de Ciencia, Innovación y Universidades (No. RTI2018-101394-B100). The authors thank BSC-RES for generously providing computational resources. The authors thank the Scientific Center for Optical and Electron Microscopy, (ScopeM) and Prof. R. Spolenak of ETH Zurich for the use of their facilities and Raman spectroscopy, respectively. Dr. R. Hauert is acknowledged for assistance with XPS measurements. The authors thank M. J. Saadun, Profs. R. Guimerà, and M. Sales for discussions on statistical learning.

■ REFERENCES

(1) McFarland, E. Unconventional Chemistry for Unconventional Natural Gas. *Science* **2012**, *338*, 340−342.
(2) Horn, R.; Schlögl, R. Methane Activation by Heterogeneous Catalysis. *Catal. Lett.* **2015**, *145*, 23−39.
(3) Lunsford, J. H. Catalytic Conversion of Methane to more useful Chemicals and Fuels: a Challenge for the 21st Century. *Catal. Today* **2000**, *63*, 165−174.
(4) Tang, P.; Zhu, Q.; Wu, Z.; Ma, D. Methane Activation: the Past and Future. *Energy Environ. Sci.* **2014**, *7*, 2580−2591.
(5) Lin, R.; Amrute, A. P.; Pérez-Ramírez, J. Halogen-Mediated Conversion of Hydrocarbons to Commodities. *Chem. Rev.* **2017**, *117*, 4182−4247.
(6) Lange, J.-P.; Tijm, P. Processes for Converting Methane to Liquid Fuels: Economic Screening through Energy Management. *Chem. Eng. Sci.* **1996**, *51*, 2379−2387.
(7) Zichittella, G.; Paunovic′, V.; Amrute, A. P.; Pérez-Ramírez, J. Catalytic Oxychlorination versus Oxybromination for Methane Functionalization. *ACS Catal.* **2017**, *7*, 1805−1817.
(8) Paunović, V.; Zichittella, G.; Moser, M.; Amrute, A. P.; Pérez-Ramírez, J. Catalyst design for Natural-Gas Upgrading through Oxybromination Chemistry. *Nat. Chem.* **2016**, *8*, 803−809.
(9) Paunović, V.; Lin, R.; Scharfe, M.; Amrute, A. P.; Mitchell, S.; Hauert, R.; Pérez-Ramírez, J. Europium Oxybromide Catalysts for efficient Bromine Looping in Natural Gas Valorization. *Angew. Chem., Int. Ed.* **2017**, *56*, 9791−9795.
(10) Paunović, V.; Hemberger, P.; Bodi, A.; López, N.; Pérez-Ramírez, J. Evidence of Radical Chemistry in Catalytic Methane Oxybromination. *Nat. Catal.* **2018**, *1*, 363−370.
(11) Nilsen, M. H.; Svelle, S.; Aravinthan, S.; Olsbye, U. The Conversion of Chloromethane to Light Olefins over SAPO-34: The Influence of Dichloromethane Addition. *Appl. Catal., A* **2009**, *367*, 23−31.
(12) Paunović, V.; Pérez-Ramírez, J. Catalytic Halogenation of Methane: a Dream Reaction with Practical Scope? *Catal. Sci. Technol.* **2019**, *9*, 4515−4530.
(13) Lorkovic, I. M.; Sun, S.; Gadewar, S.; Breed, A.; Macala, G. S.; Sardar, A.; Cross, S. E.; Sherman, J. H.; Stucky, G. D.; Ford, P. C. Alkane Bromination Revisited: "Reproportionation" in Gas-Phase Methane Bromination Leads to Higher Selectivity for CH_3Br at Moderate Temperatures. *J. Phys. Chem. A* **2006**, *110*, 8695−8700.
(14) Saadun, A. J.; Zichittella, G.; Paunović, V.; Markaide-Aiastui, B. A.; Mitchell, S.; Pérez-Ramírez, J. Epitaxially Directed Iridium Nanostructures on Titanium Dioxide for the Selective Hydrodechlorination of Dichloromethane. *ACS Catal.* **2020**, *10*, 528−542.
(15) Ding, K.; Derk, A. R.; Zhang, A.; Hu, Z.; Stoimenov, P.; Stucky, G. D.; Metiu, H.; McFarland, E. W. Hydrodebromination and Oligomerization of Dibromomethane. *ACS Catal.* **2012**, *2*, 479−486.
(16) Vilé, G.; Albani, D.; Almora-Barrios, N.; López, N.; Pérez-Ramírez, J. Advances in the Design of Nanostructured Catalysts for Selective Hydrogenation. *ChemCatChem* **2016**, *8*, 21−33.
(17) Mäki-Arvela, P.; Hájek, J.; Salmi, T.; Murzin, D. Y. Chemoselective Hydrogenation of Carbonyl Compounds over Heterogeneous Catalysts. *Appl. Catal., A* **2005**, *292*, 1−49.
(18) Williams, T.; McCullough, K.; Lauterbach, J. A. Enabling Catalyst Discovery through Machine Learning and High-Throughput Experimentation. *Chem. Mater.* **2020**, *32*, 157−165.
(19) Serra, J. M.; Chica, A.; Corma, A. Development of a Low Temperature Light Paraffin Isomerization Catalysts with Improved Resistance to Water and Sulphur by Combinatorial Methods. *Appl. Catal., A* **2003**, *239*, 35−42.
(20) Corma, A.; Serra, J. M.; Serna, P.; Valero, S.; Argente, E.; Botti, V. Optimisation of Olefin Epoxidation Catalysts with the Application of High-Throughput and Genetic Algorithms Assisted by Artificial Neural Networks (Softcomputing Techniques). *J. Catal.* **2005**, *229*, 513−524.
(21) Holena, M.; Cukic, T.; Rodemerck, U.; Linke, D. Optimization of Catalysts using Specific, Description-Based Genetic Algorithms. *J. Chem. Inf. Model.* **2008**, *48*, 274−282.
(22) Holeña, M.; Baerns, M. Feedforward Neural Networks in Catalysis: a Tool for the Approximation of the Dependency of Yield on Catalyst Composition, and for Knowledge Extraction. *Catal. Today* **2003**, *81*, 485−494.
(23) Rodemerck, U.; Baerns, M.; Holena, M.; Wolf, D. Application of a Genetic Algorithm and a Neural Network for the Discovery and Optimization of New Solid Catalytic Materials. *Appl. Surf. Sci.* **2004**, *223*, 168−174.
(24) Corma, A.; Serra, J. M.; Argente, E.; Botti, V.; Valero, S. Application of Artificial Neural Networks to Combinatorial Catalysis: Modeling and Predicting ODHE Catalysts. *ChemPhysChem* **2002**, *3*, 939−945.
(25) Huang, K.; Chen, F.-Q.; Lü, D.-W. Artificial Neural Network-Aided Design of a Multi-Component Catalyst for Methane Oxidative Coupling. *Appl. Catal., A* **2001**, *219*, 61−68.
(26) Hou, Z.-Y.; Dai, Q.; Wu, X.-Q.; Chen, G.-T. Artificial Neural Network Aided Design of Catalyst for Propane Ammoxidation. *Appl. Catal., A* **1997**, *161*, 183−190.
(27) Zavyalova, U.; Holena, M.; Schlögl, R.; Baerns, M. Statistical Analysis of Past Catalytic Data on Oxidative Methane Coupling for New Insights into the Composition of High-Performance Catalysts. *ChemCatChem* **2011**, *3*, 1935−1947.

(28) Pérez-Ramírez, J.; López, N. Strategies to Break Linear Scaling Relationships. *Nat. Catal.* **2019**, *2*, 971−976.
(29) García-Muelas, R.; López, N. Statistical Learning goes Beyond the *d*-Band Model providing the Thermochemistry of Adsorbates on Transition Metals. *Nat. Commun.* **2019**, *10*, No. 4687.
(30) Lakuntza, O.; Besora, M.; Maseras, F. Searching for Hidden Descriptors in the Metal-Ligand Bond through Statistical Analysis of Density Functional Theory (DFT) Results. *Inorg. Chem.* **2018**, *57*, 14660−14670.
(31) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752−2759.
(32) Carberry, J. J. In *Catalysis: Science and Technology*; Anderson, J. R.; Boudart, M., Eds.; Springer: Berlin, Heidelberg, 1987; Vol. *8*, pp 1−262.
(33) Mears, D. E. Diagnostic Criteria for Heat Transport Limitations in Fixed Bed Reactors. *J. Catal.* **1971**, *20*, 127−131.
(34) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(35) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169−11186.
(36) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(37) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(38) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758−1775.
(39) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953−17979.
(40) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(41) Almora-Barrios, N.; Carchini, G.; Błoński, P.; López, N. Costless Derivation of Dispersion Coefficients for Metal Surfaces. *J. Chem. Theory Comput.* **2014**, *10*, 5002−5009.
(42) Makov, G.; Payne, M. C. Periodic Boundary Conditions in Ab Initio Calculations. *Phys. Rev. B* **1995**, *51*, 4014−4022.
(43) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978−9985.
(44) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901−9904.
(45) Heyden, A.; Bell, A. T.; Keil, F. J. Efficient Methods for Finding Transition States in Chemical Reactions: Comparison of Improved Dimer Method and Partitioned Rational Function Optimization Method. *J. Chem. Phys.* **2005**, *123*, No. 224101.
(46) Henkelman, G.; Jónsson, H. A Dimer Method for Finding Saddle Points on High Dimensional Potential Surfaces Using only First Derivatives. *J. Chem. Phys.* **1999**, *111*, 7010−7022.
(47) Álvarez-Moreno, M.; de Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55*, 95−103.
(48) Pablo-García, S. CH_2Br_2 Dataset. ioChem-BD Computational Chemistry Datasets, 2020.
(49) Pablo-García, S. CH_2Br_2 Additional Dataset. ioChem-BD Computational Chemistry Datasets, 2020.
(50) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media, 2009; pp 587−602.
(51) Tin Kam, H. In *Random Decision Forests*, Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995; Vol. *1* pp 278−282.
(52) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99*, No. 016105.
(53) Falsig, H.; Hvolbæk, B.; Kristensen, I. S.; Jiang, T.; Bligaard, T.; Christensen, C. H.; Nørskov, J. K. Trends in the Catalytic CO Oxidation Activity of Nanoparticles. *Angew. Chem., Int. Ed.* **2008**, *47*, 4835−4839.
(54) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-Principles-Based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2*, 659−670.
(55) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. A Bayesian Machine Scientist To Aid In The Solution Of Challenging Scientific Problems. *Sci. Adv.* **2020**, *6*, No. eaav6971.
(56) Pablo-García, S.; Álvarez, M.; López, N. Turning Chemistry into Information for Heterogeneous Catalysis. *Int. J. Quantum Chem.*, in press, 2020.
(57) Galeener, F. L.; Mikkelsen, J. C.; Geils, R. H.; Mosby, W. J. The Relative Raman Cross Sections of Vitreous SiO_2, GeO_2, B_2O_3, and P_2O_5. *Appl. Phys. Lett.* **1978**, *32*, 34−36.
(58) Hardcastle, F. D.; Wachs, I. E. Raman Spectroscopy of Chromium Oxide Supported on Al_2O_3, TiO_2 and SiO_2: a Comparative Study. *J. Mol. Catal.* **1988**, *46*, 173−186.
(59) Korotcov, A. V.; Huang, Y.-S.; Tiong, K.-K.; Tsai, D.-S. Raman Ccattering Characterization of Well-Aligned RuO_2 and IrO_2 Nanocrystals. *J. Raman Spectrosc.* **2007**, *38*, 737−749.
(60) Amada, Y.; Watanabe, H.; Tamura, M.; Nakagawa, Y.; Okumura, K.; Tomishige, K. Structure of ReO_x Clusters Attached on the Ir Metal Surface in Ir−ReO_x/SiO_2 for the Hydrogenolysis Reaction. *J. Phys. Chem. C* **2012**, *116*, 23503−23514.
(61) Pérez-Robles, F.; García-Rodríguez, F. J.; Jiménez-Sandoval, S.; González-Hernández, J. Raman Study of Copper and Iron Oxide Particles Embedded in an SiO_2 Matrix. *J. Raman Spectrosc.* **1999**, *30*, 1099−1104.
(62) Owens, L.; Tillotson, T. M.; Hair, L. M. Characterization of Vanadium/Silica and Copper/Silica Aerogel Catalysts. *J. Non-Cryst. Solids* **1995**, *186*, 177−183.
(63) Lin, R.; Amrute, A. P.; Pérez-Ramírez, J. Halogen-Mediated Conversion of Hydrocarbons to Commodities. *Chem. Rev.* **2017**, *117*, 4182−4247.
(64) Wang, X.; Zhang, P.; Gao, J.; Chen, X.; Yang, H. Facile Synthesis and Magnetic Properties of Fe_3C/C Nanoparticles via a Sol−Gel Process. *Dyes Pigm.* **2015**, *112*, 305−310.
(65) Wang, X.; Zhang, S.; Li, J.; Xu, J.; Wang, X. Fabrication of Fe/Fe_3C@porous Carbon Sheets from Biomass and their Application for Simultaneous Reduction and Adsorption of Uranium(VI) from Solution. *Inorg. Chem. Front.* **2014**, *1*, 641−648.
(66) Clausen, C. A.; Good, M. L. Moessbauer and Far-Infrared Studies of Tetrahaloferrate Anions of the Type $FeCl_4$-nBr_n. *Inorg. Chem.* **1970**, *9*, 220−223.
(67) Anderson, A.; Lo, Y. W. Raman and Infrared Spectra of Crystals with the Cadmium Iodide Structure. *Spectrosc. Lett.* **1981**, *14*, 603−615.
(68) Li, Y.; Feng, Z.; Xin, H.; Fan, F.; Zhang, J.; Magusin, P. C. M. M.; Hensen, E. J. M.; van Santen, R. A.; Yang, Q.; Li, C. Effect of Aluminum on the Nature of the Iron Species in Fe-SBA-15. *J. Phys. Chem. B* **2006**, *110*, 26114−26121.
(69) Sinha, K.; Menéndez, J. First- and Second-Order Resonant Raman Scattering in Graphite. *Phys. Rev. B* **1990**, *41*, 10845−10847.
(70) Darmstadt, H.; Sümmchen, L.; Ting, J. M.; Roland, U.; Kaliaguine, S.; Roy, C. Effects of Surface Treatment on the Bulk Chemistry and Structure of Vapor Grown Carbon Fibers. *Carbon* **1997**, *35*, 1581−1585.
(71) Lázaro, M. J.; Echegoyen, Y.; Suelves, I.; Palacios, J. M.; Moliner, R. Decomposition of Methane over Ni-SiO_2 and Ni-Cu-SiO_2 Catalysts: Effect of Catalyst Preparation Method. *Appl. Catal., A* **2007**, *329*, 22−29.

(72) Freakley, S. J.; Ruiz-Esquius, J.; Morgan, D. J. The X-Ray Photoelectron Spectra of Ir, IrO_2 and $IrCl_3$ Revisited. *Surf. Interface Anal.* **2017**, *49*, 794−799.
(73) Morgan, D. J. Resolving ruthenium: XPS Studies of Common Ruthenium Materials. *Surf. Interface Anal.* **2015**, *47*, 1072−1079.
(74) Muhler, M.; Paál, Z.; Schlögl, R. XPS of Platinum in Pt/SiO_2 (Europt-1): Possibilities and Limitations of the Method. *Appl. Surf. Sci.* **1991**, *47*, 281−285.
(75) Zhu, Z.; Tao, F.; Zheng, F.; Chang, R.; Li, Y.; Heinke, L.; Liu, Z.; Salmeron, M.; Somorjai, G. A. Formation of Nanometer-Sized Surface Platinum Oxide Clusters on a Stepped Pt(557) Single Crystal Surface Induced by Oxygen: A High-Pressure STM and Ambient-Pressure XPS Study. *Nano Lett.* **2012**, *12*, 1491−1497.
(76) Katrib, A.; Stanislaus, A.; Yousef, R. M. XPS Investigations of Metal—Support Interactions in Pt/SiO_2, Ir/SiO_2 and Ir/Al_2O_3 Systems. *J. Mol. Struct.* **1985**, *129*, 151−163.
(77) Wang, C.-M.; Baer, D. R.; Bruemmer, S. M.; Engelhard, M. H.; Bowden, M. E.; Sundararajan, J. A.; Qiang, Y. Microstructure of the Native Oxide Layer on Ni and Cr-Doped Ni Nanoparticles. *J. Nanosci. Nanotechnol.* **2011**, *11*, 8488−8497.

Dimensionality Reduction of Complex Reaction Networks in Heterogeneous Catalysis: From Linear-Scaling Relationships to Statistical Learning Techniques

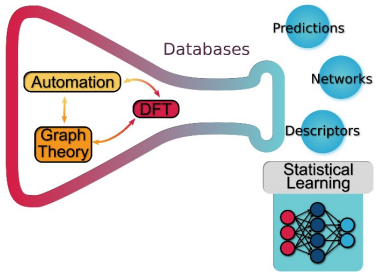
Sergio Pablo-García[1], Rodrigo García-Muelas[1], Albert Sabadell-Rendón[1], Núria López[1]

[1]Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology. Av. Països Catalans 16, 43007 Tarragona. Spain.

Abstract

The mechanistic analysis in heterogeneous catalysis is based on listing all elementary steps and evaluating explicitly their energies. To this end, computational models based on Density Functional Theory have become a standard to estimate the information needed in mechanistic studies. Typically, either the minimum energy paths or those with the smaller span are summarized in reaction profiles. Such simplifications gather a lot of information, although further dimensionality reduction is required to obtain the most relevant descriptors of catalytic activity to generate the so-called volcano plots. The selection of descriptors has been traditionally based on simple intermediates, such as central atoms in small molecules (as C in CH_4), which have good thermodynamic correlations to other fragments containing them. Yet, in emerging processes (recent studies), the number of intermediates involved increase, configurational effects and lateral interactions become significant, and complex materials with low symmetry are employed, thus the simple rules encapsulated in linear scaling relationships lose their predictive power due to error accumulation. At the same time large datasets generated for the intermediates call for statistical analysis and thus these techniques are being leveraged to chemical systems, particularly to reduce their dimensionality.

Graphical/Visual Abstract and Caption



TOC CAPTION: Summary of the new approaches needed in chemical network exploration to successfully process Density Functional Theory with statistical methods.

1. INTRODUCTION

Many industrially relevant chemical processes rely on the activation of a few atoms with a set of common intermediates characterized by a few main atoms, like the Haber-Bosch process for generating ammonia [1,2]. As raw materials taken as reactants become more complex [3], selectivity issues start to be central to attain performance [4–6] since separations are ecologically and economically costly. In addition, stability is the ultimate challenge that limits catalytic implementation and, although widely academically neglected, it constitutes a must in the design of new processes [7–9]. Finally, catalytic architectures at the industrial level are characterized by complexity including the active material, carrier [10], molecular modifiers [11] or dopants [12] plus the binders [13,14] while computational models barely go beyond pure crystals and simple orientations [15]. Therefore, complexity is an intrinsic parameter to catalysis and arises from three sources: the catalyst's structure, the reaction network, and environmental factors: (i) The catalyst's complexity appears through in homogeneous structures with multiple potential active sites; (ii) The reaction network leading from reactants to desired and unwanted products may have hundreds or thousands of intermediates and transition states; (iii) Environmental effects can be caused by pressure, local concentrations being different from the bulk of the fluid phase, external forces [16]. Their combination masks our understanding of catalytic processes.

Linear-scaling relationships, LSR, generate constraints in the available chemical space that define volcanoes for activity and selectivity. Thus, catalyst optimization relies in our ability to break such limitations [17,18]. Although this can be done effectively in labs, the search for a wider phase-space and the relatively small knowledge of complex materials has severely limited our ability to predict their catalytic properties. Heuristics have been at the core of catalytic development and large experimental databases were even generated in the early days of catalysis such as the Haber-Bosch process. This heuristic knowledge has been mainly stored in companies while academia has discouraged the publication of negative results. In the later years, there has been a change in paradigm [19,20]. Besides, computational results can now be obtained cheaply and systematically for simple but widely used families of materials, including metals and alloys. DFT can be employed to extract energies of phases that might not exist under reaction conditions due to phase changes or poisoning, and this widens our ability to interrogate the phase space more comprehensively. Spanning a larger range of, for instance, adsorption energies, simplifies the evaluation of the properties and descriptor

identification. Finally, negative results can be generated with minor cost. This has paved the way for integrating statistical learning (machine learning, ML) techniques into computational heterogeneous catalysis. ML can simplify complexity through different perspectives [21]. The leverage of these techniques to chemical systems has only been possible after many years due to the following: (i) ability to generate wide set of data with significant similar error independent of the code [22]; (ii) emergence of databases with computational data (or where to introduce this computational data) [19,23–27]; (iii) the awareness to move towards open science models [20]. In this context is interesting that such approaches have been already proved successful in biology of proteins where large databases for crystal structures have been available for the last 50 years allowing the very recent successful implementation of Artificial Intelligence (AI) algorithms [28]. In heterogeneous catalysis, like in many other communities, there is a need for AI algorithms that ensure their interpretability. This would allow to provide new synthetic routes while retaining the scientific understanding on the process. In the following we present the most acute issues to address complexity and how dimensionality reduction tools coming from statistical learning techniques can allow a more robust approach.

2. LINEAR SCALING RELATIONSHIPS

Sabatier identified that the rate of some reactions depends on a single parameter, namely oxide formation, which is the first descriptor for activity [29]. These so-called volcano plots were introduced by Balandin [30] However, descriptors have been rather elusive and mostly their use has only been employed after Density Functional Theory has been extensively employed. The reason for that is that in many cases phases change for the terminal parts of a given descriptor (*i.e.* for too exothermic O adsorption energies the system is likely to evolve to the oxide phase). And thus, even if DFT systems were reporting an ideal scenario where a phase that does not exist under reaction conditions can indeed be computed, thus decoupling phase stability and reactivity in an effective manner.

Descriptors à-la-antique

There are two main families of linear scaling relationships as described below. In the first one, structural or topological features define the thermochemistry (*i.e.* describing the adsorption energy of intermediates). In the second one, the thermochemistry defines the kinetics.

Early thermochemical models were based on group additivity rules and applied widely on molecular systems [31]. On heterogeneous catalysis, their application started on reactions involving hydrogenation of rather inert species, like the Haber-Bosch process and methanation. These could be extensively computed by the 1990 and ended up with simplified descriptors, the corresponding central atoms (as N in NH_3 or C in CH_x) [32–37]. Thus, for the activity in NH_3 formation, N_2 is physisorbed and the NH_x are found to depend on the energy of the Nitrogen atom. Although the adsorption energy of H is sometimes used as descriptor [38], it also correlates to that of N or C on transition metals [35,39] thus leaving N as the unique descriptor. The origin for the dependence of NH_x, x=1–3 with N can be related to valence considerations and thus the relative energies ended up wrapping up to a single term [32]. These dependencies on the properties of metals where further extended to relationships between central atoms [34] and to structural effects [40,41]. The electronic structure of molecular systems can be much more convoluted and thus LSR have not been derived till very recently [42]. Extensions to oxides, nitrides and other compounds were also proposed [43] but, due to the semiconductor nature of some of these materials further refinements are needed [44,45].

In oxidation processes the equations turned out not to be so straightforward and thus multivariable approaches were proposed [46]. For CO oxidation, it was found that O binding energy as a single descriptor was insufficient, and thus heuristically at least a second contribution was needed, namely CO. More recently, it was found that the two descriptors come from the decomposition of the metal-adsorbate chemical bond into covalent and electrostatic terms, and the chemical space of adsorbates on pure metals is 2D [39,47]. Two-dimensional plots have also been employed in metal oxides, particularly in the Oxygen Evolution Reaction [48,49].

Kinetics as a function of thermodynamic parameters

Besides purely structural descriptors defining thermochemistry, kinetic parameters can also be approximated from the later ones following the so-called Bell-, or Brønsted-Evans-Polanyi relations [50–54]. There, the activation energy of a given family of elementary step (e.g., hydrogenations), E_a is put as a linear function of the reaction energy, ΔE. The factor multiplying ΔE is exactly 0.5 for symmetric reactions, such in SN_2, and approaches either 0.0 or 1.0 if the transition state structure resembles more the initial or final states respectively. These scaling relationships were later on extended to put the energy of the transition state E_{TS} as a function of either the initial or final states, E_{IS} or E_{FS} [4,55–57]. and the factors multiplying E_{IS} or E_{FS} were shown to be necessarily 1.0 in order to

be universal; this is, independent on the chosen energy references [57]. Older qualitative approaches, non-linear methods such as Shustorovich Bond Order Conservation theory [58] and the UBI-QEP method [59] have been trying to rationalize the origin of such structure, thermochemistry, and kinetics correlations in terms of bonds.

3. MULTI-SCALE MODELLING

To get observables directly comparable to experiments, the energy profiles obtained *via* DFT have often been used as input to multi-scale models [6,60–63]. In those models, the LSR devised in the previous section have been employed to approximate the input parameters of the Ordinary Differential Equations (ODE) that define the chemical kinetics of a particular mechanism, and thus obtain the volcano plots. Here we describe the most popular among these methods: Microkinetics (MK), Kinetic Monte Carlo (KMC), and Computational Fluid Dynamics (CFD) simulations. MK is a mean-field method that solves the entire coupled ODE system defined by each one of the different balances present in a reaction network [63], and the boundary conditions of the experimental settings. MK can provide the composition as a function of time and can take the kinetic parameters from LSR. Thus, volcano plots can be generated by coupling simple DFT and MK *via* LSR. MK models can provide valuable insights in both homogeneous and heterogeneous catalytic systems. For instance, precise reproduction of experimental kinetic data has been achieved for the condensation of n-butylamine and benzaldehyde [64] and Rh-based hydroformylation [65]. In single atom catalysis, MK has been applied to investigate the Oxygen Reduction Reaction (ORR) with metal doped graphene [66]. In heterogeneous catalysis, several examples exists, such as alcohol reforming [57] or ethylene epoxidation on Ag [67]. The limitations in the use of DFT or LSR-derived energies became also evident in MK modelling [5]. For instance, in formic acid decomposition on Au/SiC and Pt/C, Mavrikakis *et al*. [68] tuned iteratively the DFT parameters by introducing coverage effects and improving the active site model until they reach the experimental values. Kinetic parameters can also be estimated employing Bayesian Statistics [62]. Those errors can be even bigger if employing LSR to account for the dependences between different adsorption energies or barriers. This points out towards the needs of much more accurate LSR and descriptors through ML techniques.

For highly anisotropic systems the spatial information is crucial and thus Kinetic Monte Carlo is needed [69,70]. Instead of solving the ODE system defined in pure MK, KMC calculates the

probabilities to transform the lattice current state to all possible future states. These are related to the kinetic constant, generally *via* Eyring equation, where the thermodynamic parameters are obtained by DFT. Next, the following state is selected, the lattice is updated, and the simulation time is advanced [69–71]. KMC has been implemented in several different codes, such as SPARKS [72], ZACROS [73], and kmos [74]. In transition metals, for example, KMC has been used to study the water shift reaction on Pd(100) [75]. In alloys, the mobility of atoms on metal surfaces [76] in CO oxidation on RuO_2(110) [77,78] or to find Cu and Fe percolation with amine arylation activity on graphitic carbon nitride [79]. The main drawback of this approach is the number of configurations for which DFT simulations are needed. The number of barriers to be evaluated is explosive, limiting the use of these codes. The energies for intermediates can be found from cluster expansions, while LSR can be employed to estimate the barrier without performing the DFT transition state search itself. This explains why descriptors of simple systems are found even if DFT evaluations are done in the low coverage regime.

MK [80], or, in some cases, KMC solvers [81] can be coupled to CFD codes that solve numerically all the possible balances present on the system, eg, mass, momentum, or energy. The governing equations are a set of Partial Differential Equations (PDE) and ODE [82–84]. Applications include CH_4 partial oxidation on Rh [85], CO oxidation on RuO_2(110) and Pd(100) [81], and ethylene oxidation on silver [67]. In summary, as the number of intermediates and reactions rises, the system becomes more difficult (or virtually impossible) to tackle with DFT coupled to MK, KMC or CFD alone. Likely ML techniques with adequate uncertainty estimation can ensure a seamless input to multi-scale modelling [86,87].

4. AUTOMATED NETWORK GENERATION AND ANALYSIS

First, reaction networks present strong dependencies in the nature of the intermediates adsorbed on the surface [5,6], this can be seen as an intrinsically symmetry due to the locality of the chemical bond [4,32,54,88] and can be employed in the dimensionality reduction. As this holds for many intermediates the structures linking them are also subjected to these dependencies and thus Linear Scaling Relationships, LSR, appear both for the energies of intermediates and transition states, **Figure 1**. However, as reaction processes are larger the number of related structures increases and this coupled to the intrinsic error associated to the linear fittings in LSR make the predictions,

particularly of selectivity, more uncertain. The uncertainties in the activation energies get magnified when producing activity and selectivity, as they are introduced into an exponential term (Arrhenius equation); as a result, typical errors are approximately of 2-3 orders of magnitude [5,6].

Labelling and generation

The simplest set of reactions correspond to the addition or removal of an atom or moiety. Given the active role of surfaces participating in these events they are the easiest. In comparison, concerted steps are much less common catalyzed by surfaces. Ideally, more advanced reaction search needs to implement a labelling technique to easily classify the intermediates and the reactions involved in a reaction network. For example, SMILES labelling allows the codification of an entire molecule using simple text string [89]. Another important step during the generation of a network is the definition of the connectivity between the different elements of the network. Graph theory becomes a powerful tool when connectivity plays an important role in the system, allowing to model the moieties that compose the network as nodes that are connected between edges. However, the use of graph theory is not restricted to the network, molecules can also be converted to graphs, defining their atoms as nodes connected between their chemical bonds (edges). This approach simplifies representation of molecules, easing their generation and analysis [90]. Much progress has been made in this direction, materializing into tools to generate and model complex reaction networks, as for example NetGen [91], RING [92], RMGcat [93], among others [3,94]. The use of these tools can be extended to other branches of chemistry and provide a complete set of analysis tools to extract information from a reaction network. For a discussion on the scope and limitations of all these methods the readers are directed to the recent review [95]. However, these graphs are codified to be optimal in machine language and their exploration is rather difficult, thus, requiring image techniques to be understandable by humans. This is where graphical programming languages such as DOT [96] and graph serialization tools [97] excel, generating illustrations of the network that are highly interpretable for the naked eye. Graph networks are incredibly flexible, they allow not only to focus on specific parts of the network *via* the generation of subgraphs but also to manipulate the size of the network and coupled to automation techniques.

Automation frameworks

As the reaction network grows in complexity, the amount of DFT calculations to fully describe an entire system becomes a challenge. During these years, the need to create frameworks simplifying

input generation process for massive DFT studies has emerged. Initiatives such ASE [98] or open Babel Project [99] simplify the generation of input files, drastically reducing the time needed to prepare input files *via* scripting. The combination of these automation frameworks with graph theory becomes an extremely powerful tool. Graph-modelled networks can generate the connectivity of the molecules inside the network and link them *via* elementary steps, while automation frameworks can be used to calculate these intermediates. The generated DFT data can be then integrated inside the network and used to further expand the network, generating a positive feedback loop. As the preparation and classification of DFT data is not straightforward, workflows become imperative during the automatic data generation for large systems (more than 102-103 DFT points) [100–103]. Workflows allow the application of recipes that effectively automate the energy calculations *via* taking the control of the decision making and error handling processes. Fireworks [104] and AiiDA [105,106] are only two examples of the available frameworks that implement these workflows in production environments. Due to the smart assemble of graphs, the information contained in those graph-modelled networks is already sorted and classified, being prone to be stored into databases.

Feature extraction and databasing

As the flow of data increases, it also does the need to sort, classify and store this data to be available worldwide and understandable for everyone. Online databases and chemical repositories provide an essential service to supply this need [107]. An increasing number of initiatives started emerging last decade, some of them aiming to store general DFT data such as Nomad [24] and ioChem-BD [23] while others aiming to be more specific such as Materials Project Initiative [108], Materials Cloud [25], Computational Materials Repository [26] and Catalysis Hub [109]. Some of these databases, such as ioChem-BD [23], provide an automatic refinement process, where the data provided is analysed and preprocessed before becoming public, extracting the most valuable features and explicitly exposing them to the final users. However, the true potential of these databases lies in their role as nexus between the data generation and the data exploration/analysis. Statistical learning techniques provide a powerful tool to predict and inspect the behaviour of chemical systems [19,27]. Nonetheless, the accuracy of these techniques relies on the amount and the quality of the data available to describe the system. Thus, the classification and storage provided by online databases plays a centre role during the discovery process.

Missing pieces in network coding

Networks in organic modelling like in prebiotic chemistry are more advanced [110] and such approaches need to be introduced for the reactivity on surfaces. The complexity of the graph-modelled networks resides in the connectivity between the fragments that compose the network. When the complexity of the intermediates that belong to the network increases, the entire process falls apart. Many efforts have been made to globally describe convoluted reaction networks [111] and the interactions between all the components of a chemical system [112]. Although a full description of a chemical system is virtually the most accurate approach to solve these problems, complex molecules tend to be computationally expensive and performing a full DFT analysis for a network involving a non-negligible number of complex molecules and their interactions with the rest of the system is impractical. Smart chemical space sampling [113] and thermodynamic prediction *via* machine learning [114] are some of the solutions proposed to reduce the DFT weight of these systems. However, the biggest challenge of analysing complex reaction networks is to predict of experimental rates and selectivities. Here, microkinetics is the most suitable tool to predict activities, selectivities, reaction orders, preferred reaction paths, and most-abundant reaction intermediates. Graph-modelled networks excel generating the full set of microkinetic equations [93]. However, the usage of automation lead to a combinatory explosion of intermediate and transition state energies that need to be considered in the microkinetic model. Graph analysis can be used to prune the microkinetic model through the extraction of specific subgraphs of the network and/or applying a kinetic criterion to the transition states. However, these models are limited, and cannot accurately predict reaction rates by themselves [5,6]. To overcome these issues, many efforts have been made to apply statistical learning techniques and to include other phenomena traditionally neglected in DFT and microkinetic models, as described next.

5. SIMPLIFICATIONS IMPLICIT TO AUTOMATIC MODELLING

In the automated multiscale modelling of heterogeneous catalytic processes complexity arises from three sources, namely (i) the catalyst's structure, (ii) the molecular complexity of the reactant, and (iii) environmental effects, **Figure 2**.

Catalyst complexity

Computational models generated automatically normally assume simple crystalline systems. In contrast, real catalysts may have non-regular shapes and have different ensembles [115], dopants

can change reaction paths [12] and the intrinsic nature of active sites with clear speciation can be elusive [116,117]. More generally, a catalytic material is composed by the active phase and a support intended to be and inactive carrier and meant for the dilution of the expensive catalytic metal phase. Depending on the carrier, the "inert" simplification can be disputed. The carrier may either affect the electronic and geometric structure of the active phase in the so-called strong metal-support interaction, or actively interact with the reactant as a co-catalyst. Depending on the particular effect, their activity would need to be taken into account in a separately from the main active site.

In addition, LSR are very well-developed for metals and alloys where the deviations from the d-band model are relatively small and the surface orientation and coordination can be approximated *via* metal coordination scaling [32,41]. Oxides and many other material materials have complex electronic structures that need to be considered accurately [118]. However, the LSR relations have lower accuracy when applied in oxides, sulphides, nitrides and in general multicomponent phases where two of the elements have marked differences in electronegativity [43,48,119]. Seven pillars controlling the reactivity of oxides were identified by Grasselli in 2002, namely the host structure, the strength of metal-oxygen bonds, lattice oxygens, redox properties, multifunctionality, active site isolations, and phase cooperation [119]. As these effects are overlapping, the minimum descriptors needed were found to be the vacancy formation energies as well as the basicity and acidity of the reaction centres but still much research is needed in this area [120].

Reactant complexity

The reaction network leading from reactants to desired and unwanted products may have hundreds or thousands of intermediates and transition states, depending on the size of reactants and products. LSR in heterogeneous catalysis were developed for very simplified molecular systems, rarely containing more than two central atoms [32,33,88]. However, deviations occur when increasing the size and chemical complexity of the molecule [31,121]. Firstly, when several alcohols and amines functional groups are present, the number of conformations grows exponentially [122] and the number of hydrogen bonds need to be maximized. Besides, functional groups may repeal each other mediated by the surface, thus breaking the additivity of thermochemical rules. Long hydrocarbon chains tend to maximize their interaction with the surface [122] and within themselves [121]. Molecules containing rings also have rigidity [31], which constrains which parts of them may effectively interact with the catalyst [123]. The deviations are exacerbated for conjugated and

aromatic molecules [31,124], as the energies to form intramolecular double bonds shall be compared with the molecule-surface bonds [39]. Besides, such molecules may exist in and intermediate between two states depending on the metal [39] or have stable counterintuitive radical forms [125]. Moreover, chiral centres can be formed or controlled either in multifunctional catalysts or through bifunctional strategies [126,127]. Finally, future algorithms would need to be detect if well-known organic-chemistry-textbook reactions could take place independently on the catalyst-mediated reactions. For instance, in aqueous phase, a large part of the adsorbates do interact with the solvent and their acid/base properties, thus the steps containing the formation of tautomers and zwitterions should be included in the reaction network. Also, some intermediates may fully desorb and react in the solvent without the mediation of the catalysts [128]. These complementary reaction sets are also needed with a balanced estimation error.

External factors

Besides the catalyst and the reactant, the reaction environment is another source of complexity. The more pervasive of them are the solvent interactions, which can affect the potential energy of all intermediates in a given reaction network, and may fully switch the selectivity [57,129,130]. The presence of solvent may also modify diffusion of key reactants and products, thus affecting their local concentration around the surface and the local pH. Diffusion may in turn affect the coverage. Reaction profiles are typically described in the low-coverage regime for all the intermediates, considering a mean-field approximation where all lateral interactions are neglected. This greatly simplifies the definition of linear-scaling relationships although it might introduce severe deviations, particularly for large fragments. On large molecules, steric effects may also govern selectivity. For instance, for acrolein hydrogenation on Pd, it was found that more space is required to hydrogenate the C=O functional group over the C=C one, and a different final product would be found in low and high coverage regimes [131]. In materials with small pore sizes, confinement dramatically affects the properties of the fluid phase. Finally, for photo- and electrochemical applications, external potentials can excite the electronic structure of both the catalyst and the adsorbate. The combination of all these effects altogether obscures the understanding of catalytic processes and may limit the accuracy of the models unless all relevant effects are considered.

6. DESCRIPTORS FROM MACHINE LEARNING TECHNIQUES

12

Finding appropriate descriptors for complex systems is not evident. To solve this issue, machine learning techniques are being introduced. This has the following benefits: (i) the descriptors are statistically robust and (ii) the descriptor acquisition is easily automated. The statistical techniques for extracting descriptors can be clustered in two main groups: feature selection and classification and dimensionality reduction, as illustrated in **Figure 3**. The first group selects the most representative features without changing the final dimension, while the second one transforms those features into a lower dimension.

Feature selection and classification

Two of the most popular feature selection supervised techniques are the Least Absolute Shrinkage and Selection Operator (LASSO, **Figure 3** panel **b)**), and the Elastic Net Regularization (ENR). LASSO and ENR are supervised ML techniques, in which a Y response variable is needed to perform the feature selection. Those methods are also used for regression. The LASSO method is a feature selection method based in a regression (linear, logistic, or others), in which data is shrunk towards a central point, *i.e.* the mean. LASSO adds a penalty equals to the absolute value of the coefficient (L1-regularization). Thus, regression coefficients that corresponds to correlated variables are near zero. LASSO returns simple models avoiding overfitting but if the number of points is much larger than the number of features, LASSO tends to select all the number of features in an arbitrary way. LASSO has been applied to homogeneous catalysis for predicting regioselectivities of alkenes [132] and the electronic structure of transition metal complexes with different organic ligands [133] and with a L2-regularization method (Kernel Ridge Regression), to select the best catalysts in cross-coupling reactions [134]. In heterogeneous catalysis, LASSO has been used for generating a method for screening new possible catalysts [135] and to investigate properties of single atom catalysis [136,137].

The ENR method is a variation of the LASSO method with quadratic penalty term (L2Regularization). ENR shares the strong points of LASSO, with the extra benefit that solves the issue of the number of features extracted (i) but still has the issue of the correlated variables (ii) with a high computational cost. ENR has been applied in homogeneous catalysis for quantifying steric effects on organic molecules [138] and as benchmark regressor in the screening of (111)-bimetallic alloys in heterogeneous catalysis [139].

Feature selection methods can also be coupled to classification methods to make predictions. Two of the most preferred ML classifiers in catalysis are Artificial Neural Networks (ANN), and Random Forest Classifier (RFC, **Figure 3** panel **c)**), which are supervised ML techniques. The ANN method consists in a set of connected input and output units (neurons), where each connection has an associated weight. During the training process, the network adjusts the weights to obtain the feature classification. ANN performance is particularly good for nonlinear data with large number of features and once is trained, the ANN is a very fast method. However, ANN outputs are not interpretable (black box), and strongly depends on the training data (more than another classification or regression ML methods). Thus, generalization to other datasets is more difficult and overfitting can appear. As example, in homogeneous catalysis ANN has been applied in predicting and analysing $60 \cdot 10^3$ cross-coupling reactions [140] and the prediction of formation enthalpies of hydrocarbons [141], while in heterogeneous catalysis, surface properties [142], and CO_2 reduction [3] on bimetallic surfaces have been explored.

The RFC is based in generating a set of decision trees, extract a prediction for every tree and select the best solution by scoring all the ensemble solutions. The features are selected as a function of the weight that they have in the final decision. The robustness of the method depends on the number of trees used but in general RFC is highly accurate and robust and avoids overfitting (*via* forest averaging). Again, has the drawbacks of a black box model, limiting interpretability and the cost raises with the number of trees used. RFC has been applied olefin oligomerization using Cr as homogeneous catalyst [143] and in heterogeneous catalysis to screen C_2 transformation catalysts [144], or the HER evolution on NiP_2 systems [145]. RFC has been also used together with multi-scale modelling to MK and KMC coupled to CFD for simulating ethylene oxidation on Ag [86] and in CO oxidation on RuO_2(110) [87] respectively.

Other supervised non-linear techniques for classification and regression are convolutional graph and graph embedding neural networks applied on organic molecules with biological interest [146,147], and diffusion maps, applied on proteins [148]. Even if these last methods are not very popular in our field, they show a huge potential for large systems like big organic molecules synthesis in homogenous catalysis, or mapping all the possible interactions of complex surfaces, such as oxides in heterogenous catalysis.

Dimensionality reduction

14

Common dimensionality reduction techniques are t-Stochastic Neighbour Embedding (t-SNE, **Figure 3** panel **d)**) and Principal Component Analysis (PCA, **Figure 3** panel **e)**). t-SNE transforms the distances between observations into conditional probabilities. Then, instead of comparing distances between the point x_i and its neighbour x_j, the method measures the probability for x_j to be selected assuming a Gaussian Probability Density Distribution (PDD) centred in x_i. Nearly points have a high probability, while further observations have almost 0 probabilities. Later, the algorithm generates two analogue observations y_i and y_j in a lower dimension space, and again calculates the conditional probability of the point y_j to be selected. The final output is visually attractive a 2 or 3D map, in which the input data have been grouped in such way that dense clusters are expanded, and sparse set of points are condensed. t-SNE reduces dimensionality in systems in which other techniques fails providing a very visual understanding of the data. However, 3D is the maximum dimension for the new reduced space and extrapolate to new datasets is limited. In homogeneous catalysis these techniques have allowed the construction of data-driven volcanoes [114] while in the heterogeneous catalysis context, it has been employed as visualization technique in water oxidation [149] or to derive new CO_2 electrocatalytic Cu-based bimetallic materials [113].

The PCA reduces the dimensionality by projecting all data points into the directions that capture most of the variability, called principal components. This projection is done *via* diagonalization of the covariance matrix, thus ensuring that the corresponding vectors are orthogonal. The eigenvalues contain the explained variance of each corresponding eigenvector and are taken as criterion for selecting those that explains more. The PCA presents the following advantages: (i) the precision of dimensionality reduction can be controlled by including more principal components, (ii) the results and predictions are generalizable among similar data-sets, (iii) as the master equation is a bilinear model, it is can be tuned to be interpretable, thus describing additive phenomena such as thermochemistry, and (iv) the equation is highly modular and, provided that the physical interpretation is known, it can be extended to include other terms, such as solvation and coverage. The main drawback of PCA is that it cannot capture non-linear correlations, although they are easier to spot in spaces with lower dimensionality. In homogeneous catalysis PCA has been also applied in asymmetric catalysis [150]. An iterative supervised variant, where PCA was recursively applied after fitting the obtained descriptors to the response variable, was applied for spotting selective ligands for the pyrrole synthesis [151]. The interpretable flavour of PCA has been successful applied in the decomposition of

alcohols [39] and hydrodebromination reactions [47] on metals. There it was found that the largest source of variability almost coincides with the covalent part of the metal-adsorbate bond, commonly mapped to the d-band centre on transition metals [32,152]. The second largest source of variability was associated with the relative redox character of the bonds, in line with the classical view of Pauling [153] as well as the recent interpretation of differences on the coupling matrices [154]. Thus, PCA is a highly versatile method, able to reduce the dimensionality while preserving most of the system information. PCA paves the way for a universal methodology to extract robust and physically meaningful descriptors that can be related to experimental observables [39,47]. In addition, it is possible to overcome the linearity issues of PCA by using the kernel-PCA (kPCA) method, which uses different shaped kernels (as example, the corresponding kernel (k) for linear PCA is k($\mathbf{x}$, $\mathbf{y}$) = $\mathbf{x}\cdot\mathbf{y}$). In kPCA, the covariance matrix is never explicitly diagonalized; the kernel function, which is defined as the dot product of the mapping on the new space of the non-linear combinations (k is a matrix), plays the role of the covariance matrix instead. Then, the eigenvalues and the eigenvectors of the kernel are the principal components of the new space. The kPCA has been used to study the enzyme lactate hydrogenase [155]. All these dimensionality reduction techniques have a huge potential in shortening the estimations through Density Functional Theory based on data already available.

16

Figures and Tables

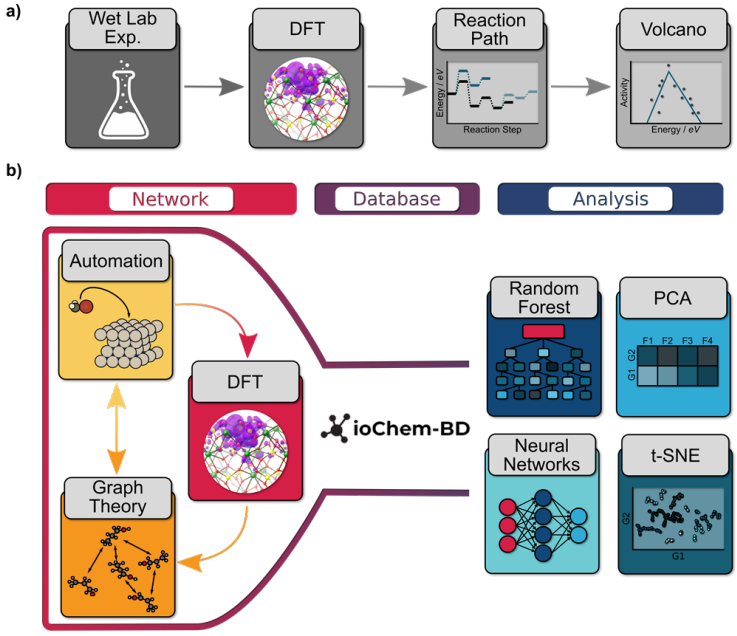


Figure 1: a) Representation of the old workflows *vs*. b) the automated generation of reaction networks coupled with statistical post-processing. In the first approach the experimental data was complemented by Density Functional Energy calculations to obtain the reaction paths, and the linear-scaling relationships between them. The result is the volcano plot. Alternatively, the new procedures automatically set up the calculations that are then stored in a database that can then be analysed with the statistical approaches.

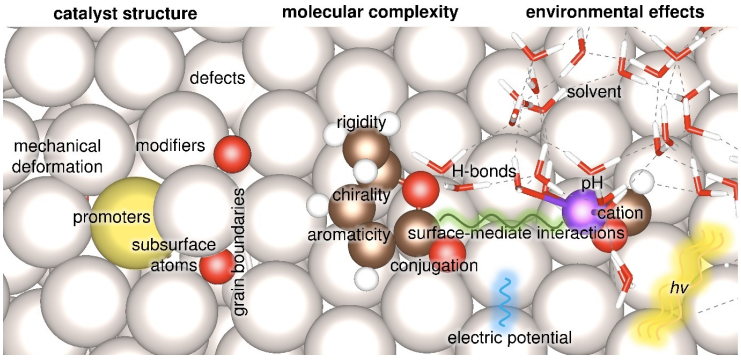


Figure 2: Sources of complexity in heterogeneous catalysis at the levels of the material, the molecule, and the external factors.

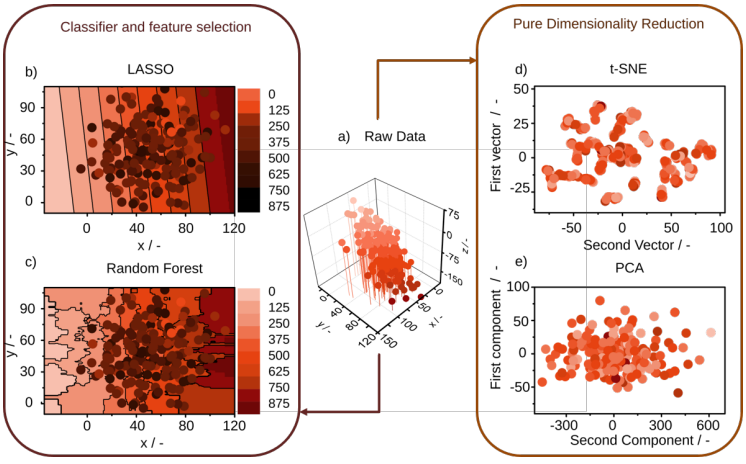


Figure 3: Schematic representation of dimensionality reduction techniques grouped as Pure Dimensional Reduction and Classifiers and Feature Selectors: a) 3 Dimensional plot of $f(x,y,z)$, where x and y are two normal randomly distributed variables, and z is a linear combination of x and y; b) LASSO c) Random Forest Classifier, d) t-SNE, and e) PCA applied on data illustrated in panel a).

Conclusion

Heterogeneous Catalysis has been based on the use of descriptors derived heuristically. However, as complexity arises simplifications based on simple arguments and chemical intuition fade. During the last years, several approaches to identify descriptors through statistical techniques have been put forward. In the present work we have revised the key implementation aspects regarding automatization, structure generation and data extraction and storage, the first bottleneck when working with data-driven approaches. We have identified several areas that require further attention, particularly when increasing the multicomponent phase of materials that serve as catalysts, when larger molecules that cannot be represented by small surrogates, and when external forces such as solvation or electric potentials are imposed to the system. Finally, we present pure dimensionality reduction techniques, like t-SNE or PCA in heterogeneous catalysis, that constitute promising and robust tools for descriptors identification thus paving the way to more advanced models that can account for activity and selectivity in an interpretable manner.

Funding Information

This work was supported by the Spanish Ministries of Science and Innovation, and Universities (RTI2018-101394-B-I00).

Research Resources

The authors thank BSC-RES for generously providing computational resources.

References

1. Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*; John Wiley & Sons, 2017. https://doi.org/10.1002/3527602658.
2. Cui, X.; Tang, C.; Zhang, Q. A Review of Electrocatalytic Reduction of Dinitrogen to Ammonia under Ambient Conditions. *Adv. Energy Mater.* **2018**, *8* (22), 1800369. https://doi.org/10.1002/aenm.201800369.
3. Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8* (1), 1–7. https://doi.org/10.1038/ncomms14621.
4. Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2* (8), 1624–1634. https://doi.org/10.1021/cs3003269.
5. Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016**, *8* (4), 331–337. https://doi.org/10.1038/nchem.2454.
6. Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-Principles-Based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2* (8), 659–670. https://doi.org/10.1038/s41929-019-0298-3.
7. Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I. B.; Nørskov, J. K. Computational High-Throughput Screening of Electrocatalytic Materials for Hydrogen Evolution. *Nat. Mater.* **2006**, *5* (11), 909–913. https://doi.org/10.1038/nmat1752.
8. Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65* (11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4.
9. Singh, A. K.; Mathew, K.; Zhuang, H. L.; Hennig, R. G. Computational Screening of 2D Materials for Photocatalysis. *J. Phys. Chem. Lett.* **2015**, *6*, 1087–1098. https://doi.org/10.1021/jz502646d.
10. Ahmadi, M.; Mistry, H.; Roldan Cuenya, B. Tailoring the Catalytic Properties of Metal Nanoparticles via Support Interactions. *J. Phys. Chem. Lett.* **2016**, *7* (17), 3519–3533. https://doi.org/10.1021/acs.jpclett.6b01198.

11. Almora-Barrios, N.; Novell-Leruth, G.; Whiting, P.; Liz-Marzan, L. M.; López, N. Theoretical Description of the Role of Halides, Silver, and Surfactants on the Structure of Gold Nanorods. *Nano Lett.* **2014**, *14* (2), 871–875. https://doi.org/10.1021/nl404661u.
12. García-Muelas, R.; Dattila, F.; Shinagawa, T.; Martín, A. J.; Pérez-Ramírez, J.; López, N. Origin of the Selective Electroreduction of Carbon Dioxide to Formate by Chalcogen Modified Copper. *J. Phys. Chem. Lett.* **2018**, *9* (24), 7153–7159. https://doi.org/10.1021/acs.jpclett.8b03212.
13. Mitchell, S.; Michels, N.-L.; Pérez-Ramírez, J. From Powder to Technical Body: The Undervalued Science of Catalyst Scale Up. *Chem. Soc. Rev.* **2013**, *42* (14), 6094–6112. https://doi.org/10.1039/c3cs60076a.
14. Hargreaves, J. S. J.; Munnoch, A. L. A Survey of the Influence of Binders in Zeolite Catalysis. *Catal. Sci. Technol.* **2013**, *3* (5), 1165. https://doi.org/10.1039/c3cy20866d.
15. Sholl, D.; Steckel, J. A. *Density Functional Theory: A Practical Introduction*; John Wiley & Sons, 2011.
16. Ardagh, M. A.; Birol, T.; Zhang, Q.; Abdelrahman, O. A.; Dauenhauer, P. J. Catalytic Resonance Theory: SuperVolcanoes, Catalytic Molecular Pumps, and Oscillatory Steady State. *Catal. Sci. Technol.* **2019**, *9* (18), 5058–5076. https://doi.org/10.1039/c9cy01543d.
17. Vojvodic, A.; Nørskov, J. K. New Design Paradigm for Heterogeneous Catalysts. *Natl. Sci. Rev. USA* **2015**, *2* (2), 140–143. https://doi.org/10.1093/nsr/nwv023.
18. Pérez-Ramírez, J.; López, N. Strategies to Break Linear Scaling Relationships. *Nat. Catal* **2019**, 1–6. https://doi.org/10.1038/s41929-019-0376-6.
19. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2.
20. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; t Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S. A.; Schultes, E.;

Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; Van Der Lei, J.; Van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.

21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer New York: New York, NY, 2009; Vol. 99. https://doi.org/10.1007/978-0-387-84858-7.
22. Lejaeghere, K.; Bihlmayer, G.; Bjorkman, T.; Blaha, P.; Blugel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A.; de Gironcoli, S.; Deutsch, T.; Dewhurst, J. K.; Di Marco, I.; Draxl, C.; Du ak, M.; Eriksson, O.; Flores-Livas, J. A.; Garrity, K. F.; Genovese, L.; Giannozzi, P.; Giantomassi, M.; Goedecker, S.; Gonze, X.; Granas, O.; Gross, E. K. U.; Gulans, A.; Gygi, F.; Hamann, D. R.; Hasnip, P. J.; Holzwarth, N. A. W.; Iu an, D.; Jochym, D. B.; Jollet, F.; Jones, D.; Kresse, G.; Koepernik, K.; Kucukbenli, E.; Kvashnin, Y. O.; Locht, I. L. M.; Lubeck, S.; Marsman, M.; Marzari, N.; Nitzsche, U.; Nordstrom, L.; Ozaki, T.; Paulatto, L.; Pickard, C. J.; Poelmans, W.; Probert, M. I. J.; Refson, K.; Richter, M.; Rignanese, G.-M.; Saha, S.; Scheffler, M.; Schlipf, M.; Schwarz, K.; Sharma, S.; Tavazza, F.; Thunstrom, P.; Tkatchenko, A.; Torrent, M.; Vanderbilt, D.; van Setten, M. J.; Van Speybroeck, V.; Wills, J. M.; Yates, J. R.; Zhang, G.-X.; Cottenier, S. Reproducibility in Density Functional Theory Calculations of Solids. *Science* **2016**, *351* (6280), aad3000–aad3000. https://doi.org/10.1126/science.aad3000.
23. Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The IoChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55* (1), 95–103. https://doi.org/10.1021/ci500593j.
24. NoMaD Repository [Internet]. Available from Https://Nomad-Coe.Eu. Accessed on Wed, December 16, 2020.
25. Materials Cloud [Internet]. Available from Https://Www.Materialscloud.Org/Home. Accessed on Wed, December 16, 2020.
26. CatApp Database - COMPUTATIONAL MATERIALS REPOSITORY [Internet]. Available from Https://Cmr.Fysik.Dtu.Dk/. Accessed on Wed, December 16, 2020.
27. Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.;

Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3* (5), 5–20. https://doi.org/10.1038/s41578-018-0005-z.

28. Callaway, E. It Will Change Everything: DeepMind's AI Makes Gigantic Leap in Solving Protein Structures. *Nature* **2020**, *588*, 203–204. https://doi.org/10.1038/d41586-020-03348-4.
29. Sabatier, P. *La Catalyse En Chimie Organique*; Encyclopédie de science chimique appliquée; Ch. Béranger, 1913. https://doi.org/10.14375/NP.9782369430186.
30. Balandin, A. A. Modern State of the Multiplet Theory of Heterogeneous Catalysis. *Adv. Catal.* **1969**, *19*, 1–210. https://doi.org/10.1016/S0360-0564(08)60029-2.
31. Benson, S. W. *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*; Wiley, 1976.
32. Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99* (1), 016105. https://doi.org/10.1103/PhysRevLett.99.016105.
33. Salciccioli, M.; Edie, S. M.; Vlachos, D. G. Adsorption of Acid, Ester, and Ether Functional Groups on Pt: Fast Prediction of Thermochemical Properties of Adsorbed Oxygenates via DFT-Based Group Additivity Methods. *J. Phys. Chem. C* **2012**, *116* (2), 1873–1886. https://doi.org/10.1021/jp2091413.
34. Montemore, M. M.; Medlin, J. W. A Unified Picture of Adsorption on Transition Metals through Different Atoms. *J. Am. Chem. Soc.* **2014**, *136* (26), 9272–9275. https://doi.org/10.1021/ja504193w.
35. Skúlason, E.; Bligaard, T.; Gudmundsdóttir, S.; Studt, F.; Rossmeisl, J.; Abild-Pedersen, F.; Vegge, T.; Jónsson, H.; Nørskov, J. K. A Theoretical Evaluation of Possible Transition Metal Electro-Catalysts for N2 Reduction. *Phys. Chem. Chem. Phys.* **2012**, *14* (3), 1235–1245. https://doi.org/10.1039/c1cp22271f.
36. Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier Principle to a Predictive Theory of Transition-Metal Heterogeneous Catalysis. *J. Catal.* **2015**, *328*, 36–42.

https://doi.org/10.1016/j.jcat.2014.12.033.

37. Montemore, M. M.; Medlin, J. W. Site-Specific Scaling Relations for Hydrocarbon Adsorption on Hexagonal Transition Metal Surfaces. *J. Phys. Chem. C* **2013**, *117* (39), 20078–20088. https://doi.org/10.1021/jp4076405.
38. Bagger, A.; Wenj,]; Varela, A. S.; Strasser, P.; Rossmeisl, J. Electrochemical CO 2 Reduction:AClassification Problem. https://doi.org/10.1002/cphc.201700736.
39. García-Muelas, R.; López, N. Statistical Learning Goes beyond the D-Band Model Providing the Thermochemistry of Adsorbates on Transition Metals. *Nat. Commun.* **2019**, *10* (1), 4687. https://doi.org/10.1038/s41467-019-12709-1.
40. Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Finding Optimal Surface Sites on Heterogeneous Catalysts by Counting Nearest Neighbors. *Science* **2015**, *350* (6257), 185–189. https://doi.org/10.1126/science.aab3501.
41. Batchelor, T. A. A.; Pedersen, J. K.; Winther, S. H.; Castelli, I. E.; Jacobsen, K. W.; Rossmeisl, J. High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule* **2019**, *3* (3), 834–845. https://doi.org/10.1016/j.joule.2018.12.015.
42. Busch, M.; Fabrizio, A.; Luber, S.; Hutter, J.; Corminboeuf, C. Exploring the Limitation of Molecular Water Oxidation Catalysts. *J. Phys. Chem. C* **2018**, *122* (23), 12404–12412. https://doi.org/10.1021/acs.jpcc.8b03935.
43. Fernández, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martínez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. K. Scaling Relationships for Adsorption Energies on Transition Metal Oxide, Sulfide, and Nitride Surfaces. *Angew. Chemie Int. Ed.* **2008**, *47* (25), 4683–4686. https://doi.org/10.1002/anie.200705739.
44. Moser, M.; Czekaj, I.; López, N.; Pérez-Ramírez, J. The Virtue of Defects: Stable Bromine Production by Catalytic Oxidation of Hydrogen Bromide on Titanium Oxide. *Angew. Chem. Intl. Ed.* **2014**, *126* (33), 8772–8777. https://doi.org/10.1002/anie.201483371.
45. Divanis, S.; Kutlusoy, T.; Boye, I. M. I.; Man, I. C.; Rossmeisl, J. Oxygen Evolution Reaction: A Perspective on a Decade of Atomic Scale Simulations. *Chem. Sci.* **2020**, *11*, 2943–2950. https://doi.org/10.1039/C9SC05897D.
46. Falsig, H.; Hvolbæk, B.; Kristensen, I. S.; Jiang, T.; Bligaard, T.; Christensen, C. H.; Nørskov,

J. K. Trends in the Catalytic CO Oxidation Activity of Nanoparticles. *Angew Chem. Intl. Ed.* **2008**, *47*, 4835–4839. https://doi.org/10.1002/anie.200801479.

47. Saadun, A. J.; Pablo-Garcia, S.; Paunovic, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J. Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* **2020**, *10*, 6129–6143. https://doi.org/10.1021/acscatal.0c00679.
48. Vojvodic, A.; Calle-Vallejo, F.; Guo, W.; Wang, S.; Toftelund, A.; Studt, F.; Martínez, J. I.; Shen, J.; Man, I. C.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. K.; Abild-Pedersen, F. On the Behavior of Brønsted-Evans-Polanyi Relations for Transition Metal Oxides. *J. Chem. Phys.* **2011**, *134* (24), 244509. https://doi.org/10.1063/1.3602323.
49. Bajdich, M.; García-Mota, M.; Vojvodic, A.; Nørskov, J. K.; Bell, A. T. Theoretical Investigation of the Activity of Cobalt Oxides for the Electrochemical Oxidation of Water. *J. Am. Chem. Soc.* **2013**, *135* (36), 13521–13530. https://doi.org/10.1021/ja405997s.
50. Bell, R. P. The Theory of Reactions Involving Proton Transfers. *Proc. R. Soc. London. A - Math. Phys. Eng. Sci.* **1936**, *154* (882), 414–429. https://doi.org/10.1098/rspa.1936.0060.
51. Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24. https://doi.org/10.1039/tf9383400011.
52. Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. J. H. Universality in Heterogeneous Catalysis. *J. Catal.* **2002**, *209* (2), 275–278. https://doi.org/10.1006/jcat.2002.3615.
53. Bligaard, T.; Nørskov, J. K.; Dahl, S.; Matthiesen, J.; Christensen, C. H.; Sehested, Jjj. The Brønsted–Evans–Polanyi Relation and the Volcano Curve in Heterogeneous Catalysis. *J. Catal.* **2004**, *224* (1), 206–217. https://doi.org/10.1016/j.jcat.2004.02.034.
54. Zaffran, J.; Michel, C.; Delbecq, F.; Sautet, P. Trade-off between Accuracy and Universality in Linear Energy Relations for Alcohol Dehydrogenation on Transition Metals. *J. Phys. Chem. C* **2015**, *119* (23), 12988–12998. https://doi.org/10.1021/acs.jpcc.5b01703.
55. Loffreda, D.; Delbecq, F.; Vigné, F.; Sautet, P. Fast Prediction of Selectivity in Heterogeneous Catalysis from Extended Brønsted-Evans-Polanyi Relations: A Theoretical Insight. *Angew. Chemie Int. Ed.* **2009**, *48* (47), 8978–8980. https://doi.org/10.1002/anie.200902800.

56. García-Muelas, R.; Li, Q.; López, N. Density Functional Theory Comparison of Methanol Decomposition and Reverse Reactions on Metal Surfaces. *ACS Catal.* **2015**, *5* (2), 1027–1036. https://doi.org/10.1021/cs501698w.

57. Li, Q.; García-Muelas, R.; López, N. Microkinetics of Alcohol Reforming for H2 Production from a FAIR Density Functional Theory Database. *Nat. Commun.* **2018**, *9* (1). https://doi.org/10.1038/s41467-018-02884-y.

58. Shustorovich, E. The Bond-Order Conservation Approach to Chemisorption and Heterogeneous Catalysis: Applications and Implications. *Adv. Catal.* **1990**, *37*, 101–163. https://doi.org/10.1016/S0360-0564(08)60364-8.

59. Shustorovich, E. The UBI-QEP Method: A Practical Theoretical Approach to Understanding Chemistry on Transition Metal Surfaces. *Surf. Sci. Rep.* **1998**, *31* (1–3), 1–119. https://doi.org/10.1016/S0167-5729(97)00016-2.

60. López, N.; Almora-Barrios, N.; Carchini, G.; Błoński, P.; Bellarosa, L.; García-Muelas, R.; Novell-Leruth, G.; García-Mota, M. State-of-the-Art and Challenges in Theoretical Simulations of Heterogeneous Catalysis at the Microscopic Level. *Catal. Sci. Technol.* **2012**, *2* (12), 2405. https://doi.org/10.1039/c2cy20384g.

61. Vlachos, D. G. Multiscale Modeling for Emergent Behavior, Complexity, and Combinatorial Explosion. *AIChE J.* **2012**, *58* (5), 1314–1325. https://doi.org/10.1002/aic.13803.

62. Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catal.* **2019**, *9* (8), 6624–6647. https://doi.org/10.1021/acscatal.9b01234.

63. Motagamwala, A. H.; Dumesic, J. A. Microkinetic Modeling: A Tool for Rational Catalyst Design. *Chem. Rev.* **2021**, *121* (2), 1049–1076. https://doi.org/10.1021/acs.chemrev.0c00394.

64. Pérez-Soto, R.; Besora, M.; Maseras, F. The Challenge of Reproducing with Calculations Raw Experimental Kinetic Data for an Organic Reaction. *Org. Lett.* **2020**, *22* (8), 2873–2877. https://doi.org/10.1021/acs.orglett.0c00367.

65. Brezny, A. C.; Landis, C. R. Development of a Comprehensive Microkinetic Model for Rh(Bis(Diazaphospholane))-Catalyzed Hydroformylation. *ACS Catal.* **2019**, *9* (3), 2501–2513. https://doi.org/10.1021/acscatal.9b00173.

66. Rebarchik, M.; Bhandari, S.; Kropp, T.; Mavrikakis, M. How Noninnocent Spectator Species

Improve the Oxygen Reduction Activity of Single-Atom Catalysts: Microkinetic Models from First-Principles Calculations. *ACS Catal.* **2020**, *10* (16), 9129–9135. https://doi.org/10.1021/acscatal.0c01642.

67. Linic, S.; Barteau, M. A. Construction of a Reaction Coordinate and a Microkinetic Model for Ethylene Epoxidation on Silver from DFT Calculations and Surface Science Experiments. *J. Catal.* **2003**, *214* (2), 200–212. https://doi.org/10.1016/S0021-9517(02)00156-2.
68. Bhandari, S.; Rangarajan, S.; Mavrikakis, M. Combining Computational Modeling with Reaction Kinetics Experiments for Elucidating the In Situ Nature of the Active Site in Catalysis. *Acc. Chem. Res.* **2020**, *53* (9), 1893–1904. https://doi.org/10.1021/acs.accounts.0c00340.
69. Chatterjee, A.; Vlachos, D. G. An Overview of Spatial Microscopic and Accelerated Kinetic Monte Carlo Methods. *J. Comput. Mater. Des.* **2007**, *14* (2), 253–308. https://doi.org/10.1007/s10820-006-9042-9.
70. Stamatakis, M.; Vlachos, D. G. Unraveling the Complexity of Catalytic Reactions via Kinetic Monte Carlo Simulation: Current Status and Frontiers. *ACS Catalysis*. American Chemical Society December 7, 2012, pp 2648–2663. https://doi.org/10.1021/cs3005709.
71. Andersen, M.; Panosetti, C.; Reuter, K. A Practical Guide to Surface Kinetic Monte Carlo Simulations. *Frontiers in Chemistry*. Frontiers Media S.A. April 9, 2019, p 202. https://doi.org/10.3389/fchem.2019.00202.
72. Slepoy, A.; Thompson, A. P.; Plimpton, S. J. A Constant-Time Kinetic Monte Carlo Algorithm for Simulation of Large Biochemical Reaction Networks. *J. Chem. Phys.* **2008**, *128* (20), 205101. https://doi.org/10.1063/1.2919546.
73. Nielsen, J.; D'Avezac, M.; Hetherington, J.; Stamatakis, M. Parallel Kinetic Monte Carlo Simulation Framework Incorporating Accurate Models of Adsorbate Lateral Interactions. *J. Chem. Phys.* **2013**, *139* (22), 224706. https://doi.org/10.1063/1.4840395.
74. Hoffmann, M. J.; Matera, S.; Reuter, K. Kmos: A Lattice Kinetic Monte Carlo Framework. *Comput. Phys. Commun.* **2014**, *185* (7), 2138–2150. https://doi.org/10.1016/j.cpc.2014.04.003.
75. Chutia, A.; Thetford, A.; Stamatakis, M.; Catlow, C. R. A. A DFT and KMC Based Study on the Mechanism of the Water Gas Shift Reaction on the Pd(100) Surface. *Phys. Chem. Chem. Phys.* **2020**, *22* (6), 3620–3632. https://doi.org/10.1039/c9cp05476f.
76. Mahlberg, D.; Groß, A. Vacancy Assisted Diffusion on Single-Atom Surface Alloys.

ChemPhysChem **2021**, *22* (1), 29–39. https://doi.org/10.1002/cphc.202000838.

77. Reuter, K.; Scheffler, M. First-Principles Kinetic Monte Carlo Simulations for Heterogeneous Catalysis: Application to the CO Oxidation at Ru O2 (110). *Phys. Rev. B - Condens. Matter Mater. Phys.* **2006**, *73* (4), 045433. https://doi.org/10.1103/PhysRevB.73.045433.
78. Pogodin, S.; López, N. A More Accurate Kinetic Monte Carlo Approach to a Monodimensional Surface Reaction: The Interaction of Oxygen with the RuO2(110) Surface. *ACS Catal.* **2014**, *4* (7), 2328–2332. https://doi.org/10.1021/cs500414p.
79. Vorobyeva, E.; Gerken, V. C.; Mitchell, S.; Sabadell-Rendón, A.; Hauert, R.; Xi, S.; Borgna, A.; Klose, D.; Collins, S. M.; Midgley, P. A.; Kepaptsoglou, D. M.; Ramasse, Q. M.; Ruiz-Ferrando, A.; Fako, E.; Ortuño, M. A.; López, N.; Carreira, E. M.; Pérez-Ramírez, J. Activation of Copper Species on Carbon Nitride for Enhanced Activity in the Arylation of Amines. *ACS Catal.* **2020**, *10* (19), 11069–11080. https://doi.org/10.1021/acscatal.0c03164.
80. Vandewalle, L. A.; Marin, G. B.; Van Geem, K. M. CatchyFOAM: Euler-Euler CFD Simulations of Fluidized Bed Reactors with Microkinetic Modeling of Gas-Phase and Catalytic Surface Chemistry. *Energy and Fuels* **2020**. https://doi.org/10.1021/acs.energyfuels.0c02824.
81. Matera, S.; Maestri, M.; Cuoci, A.; Reuter, K. Predictive-Quality Surface Reaction Chemistry in Real Reactor Models: Integrating First-Principles Kinetic Monte Carlo Simulations into Computational Fluid Dynamics. *ACS Catal.* **2014**, *4* (11), 4081–4092. https://doi.org/10.1021/cs501154e.
82. Maestri, M.; Cuoci, A. Coupling CFD with Detailed Microkinetic Modeling in Heterogeneous Catalysis. *Chem. Eng. Sci.* **2013**, *96*, 106–117. https://doi.org/10.1016/j.ces.2013.03.048.
83. Maffei, T.; Gentile, G.; Rebughini, S.; Bracconi, M.; Manelli, F.; Lipp, S.; Cuoci, A.; Maestri, M. A Multiregion Operator-Splitting CFD Approach for Coupling Microkinetic Modeling with Internal Porous Transport in Heterogeneous Catalytic Reactors. *Chem. Eng. J.* **2016**, *283*, 1392–1404. https://doi.org/10.1016/j.cej.2015.08.080.
84. Cuoci, A.; Frassoldati, A.; Faravelli, T.; Ranzi, E. A Computational Tool for the Detailed Kinetic Modeling of Laminar Flames: Application to C2H4/CH4 Coflow Flames. *Combust. Flame* **2013**, *160* (5), 870–886. https://doi.org/10.1016/j.combustflame.2013.01.011.
85. Donazzi, A.; Maestri, M.; Michael, B. C.; Beretta, A.; Forzatti, P.; Groppi, G.; Tronconi, E.; Schmidt, L. D.; Vlachos, D. G. Microkinetic Modeling of Spatially Resolved Autothermal CH4

Catalytic Partial Oxidation Experiments over Rh-Coated Foams. *J. Catal.* **2010**, *275* (2), 270–279. https://doi.org/10.1016/j.jcat.2010.08.007.

86. Partopour, B.; Paffenroth, R. C.; Dixon, A. G. Random Forests for Mapping and Analysis of Microkinetics Models. *Comput. Chem. Eng.* **2018**, *115*, 286–294. https://doi.org/10.1016/j.compchemeng.2018.04.019.
87. Bracconi, M.; Maestri, M. Training Set Design for Machine Learning Techniques Applied to the Approximation of Computationally Intensive First-Principles Kinetic Models. *Chem. Eng. J.* **2020**, *400*, 125469. https://doi.org/10.1016/j.cej.2020.125469.
88. Salciccioli, M.; Chen, Y.; Vlachos, D. G. Density Functional Theory-Derived Group Additivity and Linear Scaling Methods for Prediction of Oxygenate Stability on Metal Catalysts: Adsorption of Open-Ring Alcohol and Polyol Dehydrogenation Intermediates on Pt-Based Metals. *J. Phys. Chem. C* **2010**, *114* (47), 20155–20166. https://doi.org/10.1021/jp107836a.
89. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inform. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.
90. Wen, M.; Blau, S. M.; Spotte-Smith, E. W. C.; Dwaraknath, S.; Persson, K. A. BonDNet: A Graph Neural Network for the Prediction of Bond Dissociation Energies for Charged Molecules. *Chem. Sci.* **2021**, *12* (5), 1858–1868. https://doi.org/10.1039/D0SC05251E.
91. Pfaendtner, J.; Broadbelt, L. J. Mechanistic Modeling of Lubricant Degradation. 2. The Autoxidation of Decane and Octane. *Ind. Eng. Chem. Res.* **2008**, *47* (9), 2897–2904. https://doi.org/10.1021/ie071481z.
92. Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Description of RING. *Comput. Chem. Eng.* **2012**, *45*, 114–123. https://doi.org/10.1016/j.compchemeng.2012.06.008.
93. Goldsmith, C. F.; West, R. H. Automatic Generation of Microkinetic Mechanisms for Heterogeneous Catalysis. *J. Phys. Chem. C* **2017**, *121* (18), 9970–9981. https://doi.org/10.1021/acs.jpcc.7b02133.
94. Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient Prediction of Reaction Paths through Molecular Graph and Reaction Network Analysis. *Chem. Sci.* **2018**, *9* (4), 825–835. https://doi.org/10.1039/C7SC03628K.

95. Vernuccio, S.; Broadbelt, L. J. Discerning Complex Reaction Networks Using Automated Generators. *AIChE J.* **2019**, *65* (8). https://doi.org/10.1002/aic.16663.

96. Ellson, J.; Gansner, E. R.; Koutsofios, E.; North, S. C.; Woodhull, G. Graphviz and Dynagraph—Static and Dynamic Graph Drawing Tools. In *Graph drawing software*; Springer-Verlag, 2003; pp 127–148.

97. Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NM (United States)*; 2008.

98. Larsen, H. A.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, L. E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys. Condens. Matter* **2017**, *29* (27), 273002. https://doi.org/10.1088/1361-648X/aa680e.

99. Open Babel: Te Open Source Chemistry Toolbox. Available from: Http://Openbabel.Org/Wiki/Main_Page. Accessed on Wed, December 16, 2020.

100. Montoya, J. H.; Persson, K. A. A High-Throughput Framework for Determining Adsorption Energies on Solid Surfaces. *NPJ Comput. Mater.* **2017**, *3* (1), 1–4. https://doi.org/10.1038/s41524-017-0017-z.

101. Tran, K.; Palizhati, A.; Back, S.; Ulissi, Z. W. Dynamic Workflows for Routine Materials Discovery in Surface Science. *J. Chem. Inf. Model.* **2018**, *58* (12), 2392–2400. https://doi.org/10.1021/acs.jcim.8b00386.

102. Kahle, L.; Marcolongo, A.; Marzari, N. High-Throughput Computational Screening for Solid-State Li-Ion Conductors. *Energy Environ. Sci.* **2020**, *13* (3), 928–948. https://doi.org/10.1039/C9EE02457C.

103. Pablo-García, S.; Álvarez-Moreno, M.; López, N. Turning Chemistry into Information for Heterogeneous Catalysis. *Int. J. Quantum Chem.* **2021**, *121* (1), e26382. https://doi.org/10.1002/qua.26382.

104. Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.;

Rignanese, G.-M.; Hautier, G.; Gunter, D.; Persson, K. A. FireWorks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comput. Pract. Exp.* **2015**, *27* (17), 5037–5059. https://doi.org/10.1002/cpe.3505.

105. Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. AiiDA: Automated Interactive Infrastructure and Database for Computational Science. *Comput. Mater. Sci.* **2016**, *111*, 218–230. https://doi.org/10.1016/j.commatsci.2015.09.013.
106. AiiDA. Available from: Http://Www.Aiida.Net. Accessed on Wed, December 16, 2020.
107. Bo, C.; Maseras, F.; López, N. The Role of Computational Results Databases in Accelerating the Discovery of Catalysts. *Nat. Catal.* **2018**, *1* (11), 809–810. https://doi.org/10.1038/s41929-018-0176-4.
108. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323.
109. Catalysis Hub [Internet]. Available from Https://Www.Catalysis-Hub.Org. Accessed on Mon, March 8, 2021.
110. Wołos, A.; Roszak, R.; Żądło-Dobrowolska, A.; Beker, W.; Mikulak-Klucznik, B.; Spólnik, G.; Dygas, M.; Szymkuć, S.; Grzybowski, B. A. Synthetic Connectivity, Emergence, and Self-Regeneration in the Network of Prebiotic Chemistry. *Science* **2020**, *369* (6511), eaaw1955. https://doi.org/10.1126/science.aaw1955.
111. Eslamibidgoli, M. J.; Eikerling, M. H. Approaching the Self-Consistency Challenge of Electrocatalysis with Theory and Computation. *Curr. Opin. Electrochem.* **2018**, *9*, 189–197. https://doi.org/10.1016/j.coelec.2018.03.038.
112. Vandichel, M.; Busch, M.; Laasonen, K. Oxygen Evolution on Metal-oxy-hydroxides: Beneficial Role of Mixing Fe, Co, Ni Explained via Bifunctional Edge/Acceptor Route. *ChemCatChem* **2020**, *12* (5), 1436–1442. https://doi.org/10.1002/cctc.201901951.
113. Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; Sun, S.; Voznyy, O.; Tan, C.-S.; Askerka, M.; Che, F.; Liu, M.; Seifitokaldani, A.; Pang, Y.; Lo, S.-C.; Ip, A.; Ulissi, Z.; Sargent, E. H. Accelerated Discovery of CO2 Electrocatalysts Using Active Machine Learning. *Nature* **2020**, *581* (7807), 178–183.

31

https://doi.org/10.1038/s41586-020-2242-8.

114. Wodrich, M. D.; Fabrizio, A.; Meyer, B.; Corminboeuf, C. Data-Powered Augmented Volcano Plots for Homogeneous Catalysis. *Chem. Sci.* **2020**, *11* (44), 12070–12080. https://doi.org/10.1039/D0SC04289G.

115. Dattila, F.; García-Muelas, R.; López, N. Active and Selective Ensembles in Oxide-Derived Copper Catalysts for CO 2 Reduction. *ACS Energy Lett.* **2020**, *5* (10), 3176–3184. https://doi.org/10.1021/acsenergylett.0c01777.

116. Frei, M. S.; Mondelli, C.; García-Muelas, R.; Kley, K. S.; Puértolas, B.; López, N.; Safonova, O. V; Stewart, J. A.; Curulla Ferré, D.; Pérez-Ramírez, J. Atomic-Scale Engineering of Indium Oxide Promotion by Palladium for Methanol Production via CO2 Hydrogenation. *Nat. Commun.* **2019**, *10* (1), 3377. https://doi.org/10.1038/s41467-019-11349-9.

117. Frei, M. S.; Mondelli, C.; García-Muelas, R.; Morales-Vidal, J.; Philipp, M.; Safonova, O. V.; López, N.; Stewart, J. A.; Curulla-Ferré, D.; Pérez-Ramírez, J. Nanostructure of Nickel-Promoted Indium Oxide Catalysts Drives Selectivity in CO_2 Hydrogenation. *Nat. Commun.* **2021**, Just accepted.

118. Kauppinen, M. M.; Korpelin, V.; Verma, A. M.; Melander, M. M.; Honkala, K. Escaping Scaling Relationships for Water Dissociation at Interfacial Sites of Zirconia-Supported Rh and Pt Clusters. *J. Chem. Phys.* **2019**, *151* (16), 164302. https://doi.org/10.1063/1.5126261.

119. Grasselli, R. K. Fundamental Principles of Selective Heterogeneous Oxidation Catalysis. *Top. Catal.* **2002**, *21* (1–3), 79–88. https://doi.org/10.1023/A:1020556131984.

120. Capdevila-Cortada, M.; Vilé, G.; Teschner, D.; Pérez-Ramírez, J.; López, N. Reactivity Descriptors for Ceria in Catalysis. *Appl. Catal. B - Environ.* **2016**, *197*, 299–312. https://doi.org/10.1016/j.apcatb.2016.02.035.

121. Wodrich, M. D.; Corminboeuf, C.; Schleyer, P. von R. Systematic Errors in Computed Alkane Energies Using B3LYP and Other Popular DFT Functionals. *Org. Lett.* **2006**, *8* (17), 3631–3634. https://doi.org/10.1021/ol061016i.

122. García-Muelas, R.; López, N. Collective Descriptors for the Adsorption of Sugar Alcohols on Pt and Pd(111). *J. Phys. Chem. C* **2014**, *118* (31), 17531–17537. https://doi.org/10.1021/jp502819s.

123. Li, Q.; López, N. Chirality, Rigidity, and Conjugation: A First-Principles Study of the Key

Molecular Aspects of Lignin Depolymerization on Ni-Based Catalysts. *ACS Catal.* **2018**, *8* (5), 4230–4240. https://doi.org/10.1021/acscatal.8b00067.

124. Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C. Quantification of "Fuzzy" Chemical Concepts: A Computational Perspective. *Chem. Soc. Rev.* **2012**, *41* (13), 4671–4687. https://doi.org/10.1039/c2cs35037h.
125. Gu, J.; Wu, W.; Danovich, D.; Hoffmann, R.; Tsuji, Y.; Shaik, S. Valence Bond Theory Reveals Hidden Delocalized Diradical Character of Polyenes. *J. Am. Chem. Soc.* **2017**, *139* (27), 9302–9316. https://doi.org/10.1021/jacs.7b04410.
126. Han, X.; Xia, Q.; Huang, J.; Liu, Y.; Tan, C.; Cui, Y. Chiral Covalent Organic Frameworks with High Chemical Stability for Heterogeneous Asymmetric Catalysis. *J. Ame. Chem. Soc.* **2017**, *139* (25), 8693–8697. https://doi.org/10.1021/jacs.7b04008.
127. Szőllősi, G. Asymmetric One-Pot Reactions Using Heterogeneous Chemical Catalysis: Recent Steps towards Sustainable Processes. *Catal. Sci. Technol.* **2018**, *8* (2), 389–422. https://doi.org/10.1039/C7CY01671A.
128. Ting, L. R. L.; García-Muelas, R.; Martín, A. J.; Veenstra, F. L. P.; Chen, S. T.; Peng, Y.; Per, E. Y. X.; Pablo-García, S.; López, N.; Pérez-Ramírez, J.; Yeo, B. S. Electrochemical Reduction of Carbon Dioxide to 1-Butanol on Oxide-Derived Copper. *Angew. Chemie Int. Ed.* **2020**, *59* (47), 21072–21079. https://doi.org/10.1002/anie.202008289.
129. Garcia-Ratés, M.; López, N. Multigrid-Based Methodology for Implicit Solvation Models in Periodic DFT. *J. Chem. Theory Comput.* **2016**, *12* (3), 1331–1341. https://doi.org/10.1021/acs.jctc.5b00949.
130. Garcia-Ratés, M.; García-Muelas, R.; López, N. Solvation Effects on Methanol Decomposition on Pd (111), Pt (111), and Ru (0001). *J. Phys. Chem. C* **2017**, *121* (25), 13803–13809. https://doi.org/10.1021/acs.jpcc.7b05545.
131. Tuokko, S.; Pihko, P. M.; Honkala, K. First Principles Calculations for Hydrogenation of Acrolein on Pd and Pt: Chemoselectivity Depends on Steric Effects on the Surface. *Angew. Chemie Int. Ed.* **2016**, *55* (5), 1670–1674. https://doi.org/10.1002/anie.201507631.
132. Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine Learning for Predicting Product Distributions in Catalytic Regioselective Reactions. *Phys. Chem. Chem. Phys.* **2018**, *20* (27), 18311–18318. https://doi.org/10.1039/C8CP03141J.

133. Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8* (7), 5137–5152. https://doi.org/10.1039/C7SC01247K.

134. Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069–7077. https://doi.org/10.1039/C8SC01949E.

135. Suzuki, K.; Toyao, T.; Maeno, Z.; Takakusagi, S.; Shimizu, K.; Takigawa, I. Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data. *ChemCatChem* **2019**, *11* (18), 4537–4547. https://doi.org/10.1002/cctc.201900971.

136. O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction Trends between Single Metal Atoms and Oxide Supports Identified with Density Functional Theory and Statistical Learning. *Nat. Catal.* **2018**, *1* (7), 531–539. https://doi.org/10.1038/s41929-018-0094-5.

137. Su, Y.-Q.; Zhang, L.; Wang, Y.; Liu, J.-X.; Muravev, V.; Alexopoulos, K.; Filot, I. A. W.; Vlachos, D. G.; Hensen, E. J. M. Stability of Heterogeneous Single-Atom Catalysts: A Scaling Law Mapping Thermodynamics to Kinetics. *npj Comput. Mater.* **2020**, *6* (1), 144. https://doi.org/10.1038/s41524-020-00411-6.

138. Yamaguchi, S.; Nishimura, T.; Hibe, Y.; Nagai, M.; Sato, H.; Johnston, I. Regularized Regression Analysis of Digitized Molecular Structures in Organic Reactions for Quantification of Steric Effects. *J. Comput. Chem.* **2017**, *38* (21), 1825–1833. https://doi.org/10.1002/jcc.24791.

139. Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-Throughput Screening of Bimetallic Catalysts Enabled by Machine Learning. *J. Mater. Chem. A* **2017**, *5* (46), 24131–24138. https://doi.org/10.1039/C7TA01812F.

140. Burello, E.; Farrusseng, D.; Rothenberg, G. Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Adv. Synth. Catal.* **2004**, *346* (13–15), 1844–1853. https://doi.org/10.1002/adsc.200404170.

141. Wodrich, M. D.; Corminboeuf, C. Reaction Enthalpies Using the Neural-Network-Based X1 Approach: The Important Choice of Input Descriptors. *J. Phys. Chem. A* **2009**, *113* (13), 3285–

3290. https://doi.org/10.1021/jp9002005.

142. Palizhati, A.; Zhong, W.; Tran, K.; Back, S.; Ulissi, Z. W. Toward Predicting Intermetallics Surface Properties with High-Throughput DFT and Convolutional Neural Networks. *J. Chem. Inf. Model.* **2019**, *59* (11), 4742–4749. https://doi.org/10.1021/acs.jcim.9b00550.

143. Maley, S. M.; Kwon, D.-H.; Rollins, N.; Stanley, J. C.; Sydora, O. L.; Bischof, S. M.; Ess, D. H. Quantum-Mechanical Transition-State Model Combined with Machine Learning Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. *Chem. Sci.* **2020**, *11* (35), 9665–9674. https://doi.org/10.1039/D0SC03552A.

144. Takahashi, K.; Miyazato, I.; Nishimura, S.; Ohyama, J. Unveiling Hidden Catalysts for the Oxidative Coupling of Methane Based on Combining Machine Learning with Literature Data. *ChemCatChem* **2018**, *10* (15), 3223–3228. https://doi.org/10.1002/cctc.201800310.

145. Wexler, R. B.; Martirez, J. M. P.; Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni 2 P from Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (13), 4678–4683. https://doi.org/10.1021/jacs.8b00947.

146. Wang, X.; Li, Z.; Jiang, M.; Wang, S.; Zhang, S.; Wei, Z. Molecule Property Prediction Based on Spatial Graph Embedding. *J. Chem. Inf. Model.* **2019**, *59* (9), 3817–3828. https://doi.org/10.1021/acs.jcim.9b00410.

147. Harada, S.; Akita, H.; Tsubaki, M.; Baba, Y.; Takigawa, I.; Yamanishi, Y.; Kashima, H. Dual Graph Convolutional Neural Network for Predicting Chemical Networks. *BMC Bioinformatics* **2020**, *21* (S3), 94. https://doi.org/10.1186/s12859-020-3378-0.

148. Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11* (12), 5947–5960. https://doi.org/10.1021/acs.jctc.5b00749.

149. Palkovits, R.; Palkovits, S. Using Artificial Intelligence To Forecast Water Oxidation Catalysts. *ACS Catal.* **2019**, *9* (9), 8383–8387. https://doi.org/10.1021/acscatal.9b01985.

150. Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A. Machine Learning and Molecular Descriptors Enable Rational Solvent Selection in Asymmetric Catalysis. *Chem. Sci.* **2019**, *10* (27), 6697–6706. https://doi.org/10.1039/C9SC01844A.

151. See, X. Y.; Wen, X.; Wheeler, T. A.; Klein, C. K.; Goodpaster, J. D.; Reiner, B. R.; Tonks, I. A.

Iterative Supervised Principal Component Analysis Driven Ligand Design for Regioselective Ti-Catalyzed Pyrrole Synthesis. *ACS Catal.* **2020**, *10* (22), 13504–13517. https://doi.org/10.1021/acscatal.0c03939.

152. Hammer, B.; Nørskov, J. K. Electronic Factors Determining the Reactivity of Metal Surfaces. *Surf. Sci.* **1995**, *343* (3), 211–220. https://doi.org/10.1016/0039-6028(96)80007-0.

153. Pauling, L. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. *J. Ame Chem. Soc.* **1932**, *54* (9), 3570–3582. https://doi.org/10.1021/ja01348a011.

154. Montemore, M. M.; Medlin, J. W. Predicting and Comparing C--M and O--M Bond Strengths for Adsorption on Transition Metal Surfaces. *J. Phys. Chem. C* **2014**, *118* (5), 2666–2672. https://doi.org/10.1021/jp5001418.

155. Antoniou, D.; Schwartz, S. D. Toward Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method. *J. Phys. Chem. B* **2011**, *115* (10), 2465–2469. https://doi.org/10.1021/jp111682x.

Nuclearity and Host Effects of Carbon-Supported Platinum Catalysts for Dibromomethane Hydrodebromination

*Ali J. Saadun, Selina K. Kaiser, Andrea Ruiz-Ferrando, Sergio Pablo-García, Simon Büchele, Edvin Fako, Núria López, and Javier Pérez-Ramírez**

The identification of the active sites and the derivation of structure-performance relationships are central for the development of high-performance heterogeneous catalysts. Here, a platform of platinum nanostructures, ranging from single atoms to nanoparticles of ≈4 nm supported on activated- and N-doped carbon (AC and NC), is employed to systematically assess nuclearity and host effects on the activity, selectivity, and stability in dibromomethane hydrodebromination, a key step in bromine-mediated methane functionalization processes. For this purpose, catalytic evaluation is coupled to in-depth characterization, kinetic analysis, and mechanistic studies based on density functional theory. Remarkably, the single atom catalysts achieve exceptional selectivity toward CH_3Br (up to 98%) when compared to nanoparticles and any previously reported system. Furthermore, the results reveal unparalleled specific activity over 1.3–2.3 nm-sized platinum nanoparticles, which also exhibit the highest stability. Additionally, host effects are found to markedly affect the catalytic performance. Specifically, on NC, the activity and CH_3Br selectivity are enhanced, but significant fouling occurs. On the other hand, AC-supported platinum nanostructures deactivate due to sintering and bromination. Simulations and kinetic fingerprints demonstrate that the observed reactivity patterns are governed by the H_2 dissociation abilities of the catalysts and the availability of surface H-atoms.

1. Introduction

The development of heterogeneous metal-based catalysts that integrate a high density of active, selective, and stable ensembles, granting efficient turnover and prolonged lifetimes, is one of the main targets in catalysis science, often driven by increasing societal demands and environmental challenges.[1] This goal could be approached by engineering both the geometry and the electronic properties of the active phase at the nanoscale.[2,3] However, in contrast to well-defined homogeneous catalysts, establishing structure-performance relations and identifying the active sites in heterogeneous systems is challenging due to the inherent material complexity.[3,4] In this regard, employing single-atom heterogeneous catalysts (SACs), containing isolated atoms in discrete chemical environments is an effective approach to enable fundamental and mechanistic studies.[3–7] Beyond this, SACs often exhibit unique performance in diverse reactions due to their high degree of metal dispersion, tunable electronic properties, and unsaturated coordination environments of the active centers in tailored host materials.[7–14] In this regard, the first step in the development of a catalyst design strategy entails a detailed assessment on the impact of metal nuclearity and host effects, from single atoms with defined environments to size-controlled nanoparticles, on reactivity patterns for a given application.[15–17]

A prominent class of reactions, widely used in numerous industrial processes, are hydrogenations,[18–20] which are commonly carried out over supported nanoparticles of precious metals with high sensitivity to their specific ensemble design.[21–27] An example of high practical relevance is the hydrodebromination of dibromomethane (CH_2Br_2) into bromomethane (CH_3Br), an important transformation for the industrial realization of bromine-mediated natural gas upgrading technologies into chemicals and fuels.[28–30] Therein, limited progress has been made toward selective hydrodebromination, where carbon losses in the form of CH_4 and coke represent a major challenge.[30,31] A recent study evaluated the performance of SiO_2-supported nanoparticle-based (NP-based) metal catalysts (1 wt% of Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir, Pt). Therein, at comparable reaction conditions, iron-, cobalt-, copper-, and silver-based catalysts displayed CH_2Br_2 conversion levels of 4–7%, showing consistency with the reported poor hydrodebromination ability of these elements.[30] Among the platinum group metals, selective CH_2Br_2 hydrodebromination to CH_3Br was reported over Ru/SiO_2 (up to 96%, **Table 1**), whereas the CH_3Br selectivity over Rh/SiO_2 (<48%), Ir/SiO_2 (<40%), and Pt/SiO_2 (23%) was

A. J. Saadun, S. K. Kaiser, S. Büchele, Prof. J. Pérez-Ramírez
Department of Chemistry and Applied Biosciences
Institute for Chemical and Bioengineering
ETH Zurich
Vladimir-Prelog-Weg 1, Zürich 8093, Switzerland
E-mail: jpr@chem.ethz.ch
A. Ruiz-Ferrando, S. Pablo-García, E. Fako, Prof. N. López
Institute of Chemical Research of Catalonia (ICIQ)
The Barcelona Institute of Science and Technology
Av. Països Catalans 16, Tarragona 43007, Spain

lower due to coking and over hydrogenation to CH_4, respectively.[31] Despite the promising initial selectivity to CH_3Br, coking and sintering cause the rapid deactivation of Ru-based catalysts. On the contrary, platinum nanoparticles of ≈2 nm show the highest stability all systems investigated. In addition, Pt/SiO_2 displays a turn over frequency (*TOF*) that is *ca.* 8-fold higher than the benchmark ruthenium catalyst (Table 1), which invites further investigations on this metal from an application viewpoint. However, support effects on activity, selectivity, and stability were not explored and no attempts were made to enhance the performance of Pt-based catalysts by nanostructuring of the active site, an approach that has been proven effective for improving catalytic performance in selective dihalomethane hydrodehalogenation.[32] Furthermore, the formulation of structure–performance relationships is hampered by the limited molecular insights available. To date, the binding strength of the metal with CH and Br fragments and the ability to activate H_2 were postulated as activity and selectivity descriptors,[31] leaving ample room for further fundamental investigations.

This encouraged us to systematically study nuclearity and host effects of platinum catalysts in CH_2Br_2 hydrodebromination on their activity, selectivity, and stability. For this purpose, a platform of activated carbon (AC) and nitrogen-doped carbon (NC) supported platinum nanostructures, ranging from single atoms to nanoparticles of gradually increasing size, was adopted. Controlling the functionalization of the carbon host (e.g., N or O, suitable anchoring sites to platinum) and use of thermal activation allowed tuning of the density and structure of the metal coordination sites. By combining catalyst evaluation with in-depth characterization, kinetic analysis, and density functional theory (DFT), we rationalize the activity and selectivity patterns and unravel distinct deactivation mechanisms, thereby providing guidelines for the design of catalysts with improved activity, selectivity, and stability performance.

2. Results and Discussion

2.1. Catalyst Characterization

To study nuclearity and host effects of Pt-catalyzed CH_2Br_2 hydrodebromination, a platform of platinum nanostructures with fixed metal loading (1 wt%) was derived by extending a previously reported synthesis approach,[17] consisting of the dry impregnation of platinum chloride on activated- (AC) and nitrogen doped carbon (NC) carriers, followed by thermal activation and reduction steps. Specifically, thermal activation was performed at different temperatures (T_{act}, 473–1073 K), resulting in catalysts designated as Pt/AC-T_{act} and Pt/NC-T_{act}, which contain platinum ensemble sizes ranging from single atoms with different coordination environment to nanoparticles of 1.3 nm. The Pt/AC-473 and Pt/NC-473 samples underwent the additional reduction step in a H_2-rich atmosphere at elevated temperatures, T_{red}, of 573 and 873 K, with the aim to derive nanoparticles larger than 1.3 nm through enhanced sintering. Those catalysts were denoted as Pt/AC(T_{red}) and Pt/NC(T_{red}) (**Figure 1**a).

Analysis of the porous properties of the as-prepared materials by N_2-sorption (**Table 2**) revealed that the specific surface area and pore volume (S_{BET} = 916–1093 $m^2\ g^{-1}$, V_{pore} = 0.56–0.68 $cm^3\ g^{-1}$) of the AC-supported catalysts are larger than those of the NC-supported catalysts (S_{BET} = 363–545 $m^2\ g^{-1}$, V_{pore} = 0.32–0.42 $cm^3\ g^{-1}$), as expected from the blank supports. Quantification of the metal content by inductively coupled plasma-optical emission spectrometry (ICP-OES) confirmed that the platinum loading of the catalysts were very close to the targeted value of 1 wt% (Table 2). The powder X-ray diffraction (XRD) patterns of Pt/AC-473 and Pt/NC-473 showed broad diffraction peaks compatible with the carbon phase, whereas platinum reflections were not visible (Figure 1b), suggesting the absence of large nanoparticles. In contrast, the Pt/AC(T_{red}) and Pt/NC(T_{red}) samples showed characteristic diffraction patterns of metallic platinum (Figure 1b; Figure S1, Supporting Information), indicating that the additional treatment in H_2 promotes metal sintering. These observations were corroborated with high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM), highlighting atomically dispersed platinum in the catalysts activated at 473 K, irrespective of the host functionalization (i.e., Pt/NC-473 and Pt/AC-473, Figure 1c). The platinum single atoms remain atomically dispersed at a treatment temperature of 1073 K on NC (Pt/NC-1073), whereas sintering into nanoparticles with an average size of 0.6 and 1.3 nm was observed on AC at 673 and 1073 K (Pt/AC-673 and Pt/AC-1073), respectively (Table 2).[17] The micrographs further revealed the morphological evolution of platinum single atoms to nanoparticles upon thermal treatment in H_2 at 573 and 873 K, progressing to an average size of 2.3 and 3.1 nm in Pt/NC(T_{red}) and 2.9–3.8 nm in Pt/AC(T_{red}) (Figure 1c; Figure S2, Supporting Information). Although nanoparticles are dominant in these materials, the presence of some single atoms cannot be excluded, especially for the catalysts reduced at 573 K. A summary of the catalysts studied herein is presented in Table 2.

The extended X-ray adsorption fine structure (EXAFS) spectra showed no peaks characteristic for the Pt—Pt bonds in the samples treated at 473 K (Figure S3, Supporting Information), which supports the absence of nanoparticles in these samples. Rather, contributions were observed revealing 2.3 ± 0.3 and 3.2 ± 0.3 Cl neighbors in Pt/NC-473 and Pt/AC-473, respectively.[17] The dominating environment of the platinum single atoms in the NC-supported system changes from Cl to N/O-neighbors at a thermal activation temperature of 1073 K. Further analysis by X-ray photoelectron spectroscopy (XPS) was performed to gain a deeper understanding on the chemical state of platinum in the single atoms and nanoparticles

Table 1. Catalysts for CH_2Br_2 hydrodebromination. Data taken from ref. [31].

Catalyst	Reactivity[a]	
	S (CH_3Br [%])	*TOF* $[h^{-1}]$[b]
Ni/SiO_2	64	8
Ru/SiO_2	96	14
Rh/SiO_2	48	32
Pt/SiO_2	23	277

[a]Selectivity to CH_3Br and rate were denoted as *S* and *TOF*, respectively; [b]Average metal particle size in the range of 1.6-2.4 nm. Reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:4.5:65.5, F_T:W_{cat} = 40 $cm^3\ min^{-1}\ g_{cat}^{-1}$, *T* = 523 K, *P* = 1 bar, and *tos* = 15 min.

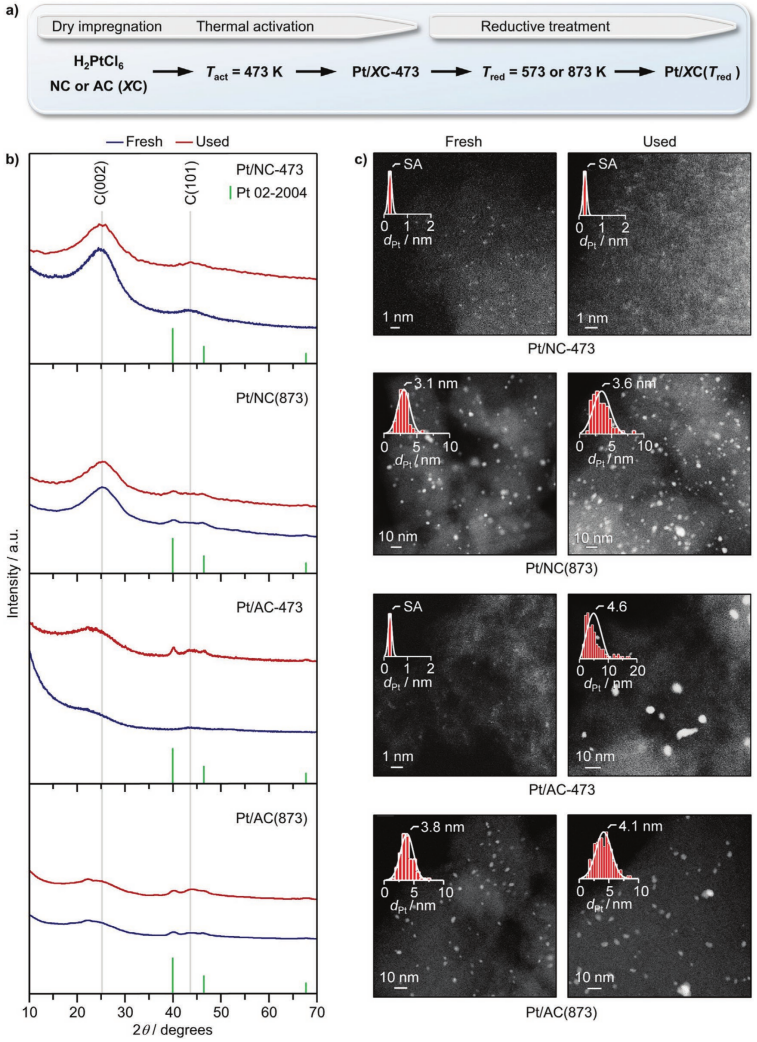


Figure 1. a) Synthetic route for the preparation of carbon-supported platinum nanostructures ranging from single atoms (SA) to nanoparticles. b) XRD patterns and c) HAADF-STEM micrographs with derived particle size distributions of selected catalysts in fresh and used forms. Conditions for used catalysts: CH_2Br_2:H_2:Ar:He = 6:24:5:65, F_T:W_{cat} = 200-500 cm^3 min^{-1} g_{cat}^{-1}, T = 523 K, P = 1 bar, and *tos* = 10 h.

(**Figure 2**; Table S1, Supporting Information). The Pt 4f spectra of the single atom-based (SA-based) systems, Pt/AC-473 and Pt/NC-473, showed two main contributions at a binding energy (BE) of 72.4 ± 0.1 and 73.6 ± 0.1eV, which are assigned to Pt(II) and Pt(IV), respectively.[17] This strongly oxidized character of the single atoms in Pt/AC-473 and Pt/NC-473 is

Table 2. Characterization data of the catalysts.

Catalysts	Pt content[a] [wt%]	Fresh (used)[f]			
		S_{BET}[b] [$m^2\ g^{-1}$]	V_{pore}[c] [$cm^3\ g^{-1}$]	d_{Pt}[d] [nm]	Br content[e] [wt%]
Pt/NC-473	0.97	469 (53)	0.38 (0.10)	SA (SA)	0 (5.5)
Pt/NC-1073	0.94	363 (30)	0.32 (0.10)	SA	–
Pt/NC(573)	1.00	488 (232)	0.40 (0.25)	2.3 (3.0)	–
Pt/NC(873)	1.00	545 (215)	0.43 (0.22)	3.1 (3.6)	0 (6.0)
Pt/AC-473	0.94	922 (952)	0.57 (0.59)	SA (4.6)	0 (1.5)
Pt/AC-673	1.02	916 (860)	0.56 (0.54)	0.6	–
Pt/AC-1073	0.97	1093 (1030)	0.68 (0.65)	1.3	–
Pt/AC(573)	1.01	930 (850)	0.58 (0.52)	2.9 (3.7)	–
Pt/AC(873)	0.98	1050 (920)	0.66 (0.56)	3.8 (4.1)	0 (1.1)

[a]ICP-OES; [b]BET model; [c]Volume of N_2 adsorbed at $p/p_0 = 0.98$; [d]Derived from HAADF-STEM micrographs. SA: single atoms; [e]Quantified by XPS; [f]After 10 h in CH_2Br_2 hydrodebromination. Reaction conditions as specified in the caption of Figure 1.

related to the presence of Cl- and N/O-atoms coordinated to platinum (Figures S3 and S4, Supporting Information), in line with the EXAFS analysis. In contrast, chlorine was not detected in the reduced NP-based systems, Pt/AC(873) and Pt/AC(873), and an additional contribution of metallic platinum at a binding energy of 71.2 ± 0.1 eV was observed (Figure 2). However, these systems still show contributions of peaks compatible with Pt(II) and Pt(IV) assignments, likely related to quantum size effects that are most prominent for small nanoparticles.[33] Notably, Pt/NC(873) shows a strong contribution of Pt(II), which is likely related to an increased metal-support interaction of the NC compared to AC.

For the single atoms, a thorough speciation analysis was conducted via simulations by evaluating the stability of platinum in distinct chemical environments, typically co-existing on N-doped carbon (Figure S5, Supporting Information). Among these, the 3N cavity (denoted as "rol3") was found to yield the most stable Pt/NC configuration and was thus selected as the most suitable representative of the active site. Notice that a single atom reference and a platinum bulk reference are employed to assess the thermodynamic limitations of aggregation and the ability of the cavity to adsorb the platinum single atoms.

2.2. Catalyst Evaluation

The gas-phase CH_2Br_2 hydrodebromination was studied at constant reaction temperature (523 K), feed composition (CH_2Br_2:H_2:Ar:He = 6:24:5:65), and pressure (≈1 bar). Comparison of CH_3Br production rates at a space velocity of 200 $cm^3\ min^{-1}\ g^{-1}$ resulted in the following order of decreasing specific activity: Pt/NC(873) > Pt/NC(573) > Pt/NC-1073 >> Pt/AC-1073 > Pt/AC-673 ≈ Pt/AC(873) ≈ Pt/AC(573) ≈ Pt/NC-473 ≈ Pt/SiO_2 (at a space velocity of 40 $cm^3\ min^{-1}\ g^{-1}$) > Pt/AC-473 (**Figure 3a**). Figure 3a displays these trends as a function of the ensemble size, indicating that NP-based systems show a higher activity compared to their SA-based analogues. More detailed, nanoparticles with an average size of 3.1 and 3.8 nm supported on NC and AC, respectively, exhibit ≈1.5-fold higher activity than their SA-based counterparts. With increasing ensemble size, the AC-supported platinum catalysts achieve the maximum activity at a nanoparticle diameter of 1.3 nm, whereas the NC-supported platinum systems seem to reach a maximum at nanoparticle sizes of 2.3 nm. Notably, the reactivity of the benchmark Pt/SiO_2, with an average nanoparticle

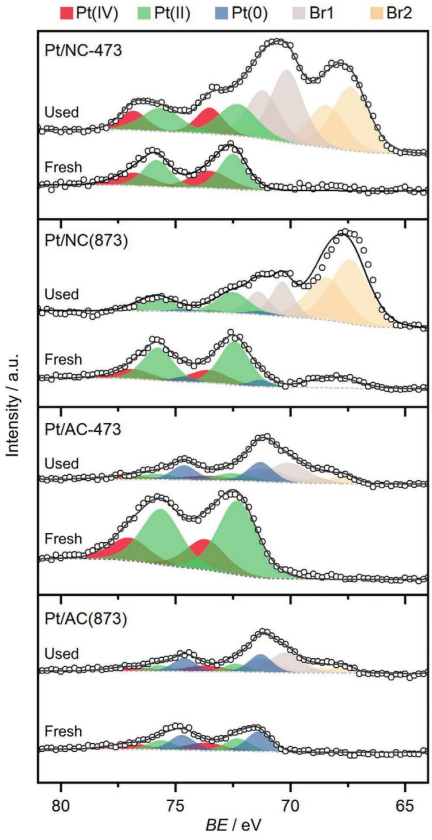


Figure 2. Pt 4f XPS spectra of selected catalysts. The dark gray lines and open circles represent the overall fit and the raw data, respectively, while the colored areas indicate the fit of distinct chemical components. The contribution of the Br 3d signal appearing at a similar binding energy for the used catalysts is separated by fitting two Br species (gray and yellow). Conditions for used catalysts: CH_2Br_2:H_2:Ar:He = 6:24:5:65, F_T:W_{cat} = 200-500 $cm^3\ min^{-1}\ g_{cat}^{-1}$, T = 523 K, P = 1 bar, and tos = 10 h.

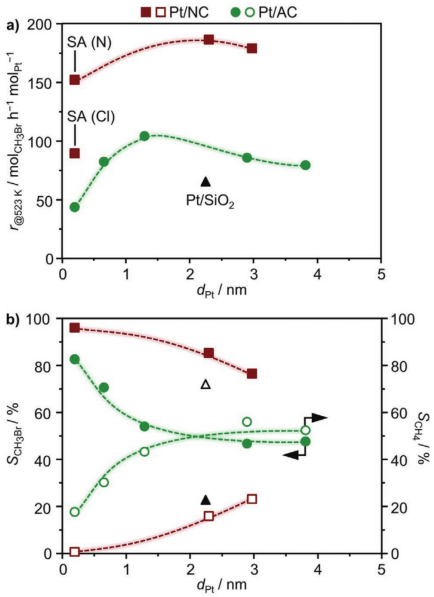


Figure 3. a) Rate of CH_3Br production and b) product selectivity in CH_2Br_2 hydrodebromination over the catalysts as a function of the average platinum particle size (d_{Pt}). In panel (a) the activity was assessed at a constant space velocity of F_T:W_{cat} = 200 cm^3 min^{-1} g_{cat}^{-1}, while product selectivities in (b) were determined at ≈20% CH_2Br_2 conversion achieved by adjusting the space velocity in the range of F_T:W_{cat} = 200-500 cm^3 min^{-1} g_{cat}^{-1}. Other reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:5:65, T = 523 K, P = 1 bar, and *tos* = 15 min. The triangle indicates the performance of Pt/SiO_2, whereas SA stands for single atoms.

size of 2.2 nm, were threefold lower than the NC-supported platinum nanoparticles with comparable sizes. Furthermore, NC-supported single atoms with an N/O-coordination environment displayed ≈1.5-fold higher activity compared to their Cl-coordinated analogues. The higher activity achieved with the N/O-coordination can be explained by the catalytic role of the matrix on the reaction, where, as proposed in the literature and as shown in density functional theory simulations (vide infra), the N-sites can store hydrogen atoms, leaving the metal center free for coordination.[7]

The product distribution over the catalysts was compared at ≈20% CH_2Br_2 conversion, achieved by adjusting the space velocity (Figure 3b). Remarkably, single atoms supported on NC demonstrate unparalleled selectivity to CH_3Br (up to 98%) and low propensity to CH_4 (<20%), regardless of the host coordination environment, whereas platinum single atoms on AC exhibited a CH_3Br selectivity of ≈84%. These results indicate that SA-based platinum catalysts are able to suppress over hydrogenation pathways, which is a major challenge in this reaction. The gradual increase of selectivity to CH_4 at the expense of CH_3Br formation with increasing platinum ensemble size, from single atoms to nanoparticles of ≈4 nm, confirms the superior selectivity performance of single atoms compared to their nanoparticle counterparts (Figure 3b). The CH_3Br selectivity attained over the NC-supported single atoms rivals that of SiO_2-supported ruthenium nanoparticles (up to 96%), the most selective hydrodebromination catalyst reported so far.[31] Moreover, the selectivity performance over the carbon-supported platinum catalysts is dependent on the type of host, whereby NC-supported systems exhibit higher CH_3Br selectivity than their AC-supported counterparts, suggesting that N-functionalities play a central role in the reaction (vide infra).

The established activity and selectivity patterns were complemented with stability tests to gain a complete picture of the performance of each catalytic system (**Figure 4**). These tests were performed at an initial conversion level of ≈20% (Figure S6, Supporting Information). Upon exposure to the reaction environment for 10 h, the SA-based catalysts displayed activity losses up to 70% (compared to the initial performance). The NC-supported platinum single atoms show a significant CH_2Br_2 conversion loss in the first 2 h on stream (≈60%, Figure 4a), whereas a gradual, though substantial depletion of activity (70% drop) was observed over the AC-supported single atom system (Figure 4b). On the other hand, NP-based materials preserved their initial activity better, displaying activity losses in the range of 10–38%. To allow a direct comparison of the stability performance, the deactivation was expressed with the constant k_D (Figure 4c), which indicates the activity losses per hour, and which was derived via linear regression of the data in the time-on-stream (*tos*) range that is indicated in Figure S6, Supporting Information. Displaying the k_D values as a function of the platinum ensemble size emphasized the observed patterns and highlighting host- and platinum ensemble size roles on stability (Figure 4c). Depletion of activity was appended by moderate changes in the product distribution over the catalysts. With *tos*, the NC-supported SA-based Pt/NC-473 and Pt/NC-1073 maintained their high performance with >98% selectivity to CH_3Br, suggesting preservation of the single atom nanostructure. On the other hand, activity losses were accompanied by an increase of CH_4 selectivity (from 18% to 33%, at the expense of CH_3Br) over SA-based Pt/AC-473, likely due to sintering of the active phase (vide infra). NP-based catalysts generally showed no pronounced selectivity changes with *tos*.

2.3. Deactivation Mechanisms

To examine the development of the materials during exposure to the reaction environment, detailed characterization of selected used catalysts by N_2-sorption, XRD, HAADF-STEM, and XPS was performed, revealing three main deactivation mechanisms: i) fouling by coking, ii) active phase sintering, and iii) bromination. N_2-sorption indicated the significant decrease in the specific surface area and pore volume (S_{BET} = 215–30 m^2 g^{-1}, V_{pore} = 0.25–0.10 cm^3 g^{-1}) of the NC-supported systems (Table 2), highlighting the poor stability of the carrier. The decrease of S_{BET} and V_{pore} was more pronounced over the NC-supported SA-based catalysts than their NP-based equivalents, suggesting that the former were more affected by fouling. In contrast,

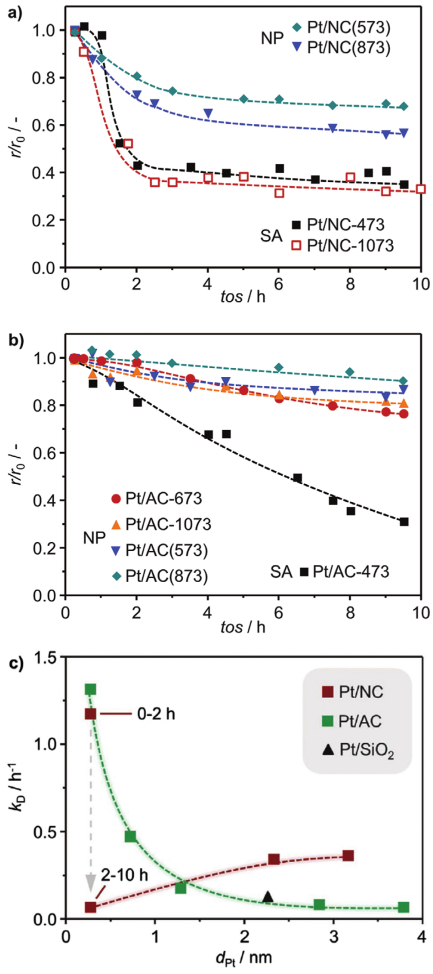


Figure 4. Relative rate of CH_2Br_2 hydrodebromination as a function of time on stream over the a) NC-supported and b) AC-supported platinum catalysts. c) The deactivation constants (k_D) as a function of the average platinum particle size (d_{Pt}). The triangle in (c) indicates the performance of Pt/SiO_2. Reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:5:65, F_T:W_{cat} = 200-500 $cm^3\ min^{-1}\ g_{cat}^{-1}$, T = 523 K, and P = 1 bar. SA and NP stand for single atoms and nanoparticles, respectively.

AC-supported catalysts showed minimal changes in S_{BET} and V_{pore}, likely due to the absence of coking mechanisms. Furthermore, the XRD profiles of Pt/NC-473, Pt/NC(873), and Pt/AC(873) resemble those of the fresh materials (Figure 1b), implying that no severe active phase sintering occurred over these materials. On the other hand, a strong reflection assigned to metallic platinum was detected for used Pt/AC-473, indicating that active phase restructuring occurred, which was confirmed by HAADF–STEM microscopy showing the increase of the average platinum ensemble size from single atoms to nanoparticles of 4.6 nm (Figure 1c). The micrographs further disclosed the absence of nanoparticles on used Pt/NC-473, whereas modest platinum sintering was distinguished over the used NP-based systems, Pt/NC(T_{red}), and Pt/AC(T_{red}) (Table 2). Interestingly, the SA-based Pt/AC-473 displays the highest deactivation rate despite sintering of its active phase into nanoparticles, which were found to be more reactive than single atoms (Figure 3a). Therefore, the observed activity losses likely originate, beside bromination, from the distinct surface sites, which alter the coverage of the surface species and gives rise to a different reactivity compared to other nanoparticles with a similar size. As is recognized in other hydrogenation studies,[24,34] the surface concentration of active corners and edges could significantly alter the hydrogenation performance. This possible influence of the site type in nanoparticles on reactivity broadens the mechanistic diversity of hydrodebromination reactions over platinum surfaces, and deserves to be tackled in future dedicated studies.

XPS analysis was conducted to investigate the chemical state of platinum and the role of surface bromination. The similar binding energies of the Pt 4f and Br 3d signals require a simultaneous fitting approach of distinct chemical components for both elements (Figure 2; Table S1, Supporting Information). The spectra revealed prominent surface bromination for NC-based materials (up to 6 wt% Br), whereas their AC-based counterparts were less affected by Br-poisoning (up to 1.5 wt% Br) (Table 2; Figure S4, Supporting Information). The total Br-content in the catalysts was mainly determined by the carrier functionalization, since NP-based systems displayed comparable Br-content to their SA-based counterparts. For the SA-based materials, no significant change in the chemical state is observed in fresh and used Pt/NC-473, with similar contributions from Pt(II) and Pt(IV), and no peak designated to metallic platinum in Pt/NC-473 (Figure 2), which is well in line with the suggested absence of platinum sintering. In contrast, Pt/AC-473 is reduced with a clear peak of Pt(0) appearing in the used catalyst, showing a comparable chemical state as its NP-based analogue, Pt/NC(873). On the other hand, the NP-based systems display a contribution of Pt(II) (Figure 2), which was less pronounced in the used Pt/AC(873) catalyst, likely due to the bromination of the active phase.

In view of a potential industrial application, the possibility to regenerate Pt/NC-473 was investigated. Whereas the premier cause of activity losses with *tos* is coking, an attempt was made to restore the initial activity (21% CH_2Br_2 conversion) by removal of the carbonaceous residues from the used catalyst, which exhibited a CH_2Br_2 conversion of 7% after 10 h on-stream, via thermal treatment in 20 vol% O_2 in He (F_T = 20 cm^3 STP min^{-1}). Since N-doped carbon remains stable up to 573 K under these conditions,[17] the treatment was conducted at slightly lower temperatures (523 K) for 6 h. The treated catalyst displayed marginal regeneration of its hydrodebromination activity with a CH_2Br_2 conversion level of 11% after 15 min on-stream, suggesting that the structural

changes of the support are not reversible. Moreover, the selectivity to CH_3Br is high (86%), though still lower than the initial performance (98%), possibly due to restructuring of the active phase. The changes introduced in the host and the active sites upon catalyst treatment is beyond the scope of this study and deserves attention in future dedicated studies. These results highlight the fact that catalyst robustness in CH_2Br_2 hydrodebromination remains a major challenge.

2.4. Mechanistic and Kinetic Studies

In nanoparticles, a fraction of the atoms, located in the bulk, are not exposed to the reactants, whereas each of the atomically dispersed single atoms could react with CH_2Br_2 and H_2. Moreover, in contrast to platinum nanoparticles, the single atoms do not have a neighboring platinum atom, which could affect their interaction with the reactant molecules. Expressing the catalytic activity per surface platinum atom will provide first insights on these effects. To this end, the turnover frequency (*TOF*) of selected SA-based systems was compared (at a similar space velocity of 200 $cm^3\ min^{-1}\ g_{cat}^{-1}$) with their NP-based counterparts (**Figure 5**). The dispersion of platinum in the nanoparticle-based catalysts was assessed by CO chemisorption. The activity of the catalysts decreased in the following order: Pt/AC(873) ≈ Pt/NC(873) >> Pt/NC-473 > Pt/AC-473, with dispersion values of 31% and 26% in Pt/NC(873) and Pt/AC(873), respectively. The NP-based systems exhibit a higher *TOF* compared to their SA-based analogues (>6 times), suggesting that neighboring Pt-atoms participate in the reaction, thereby enhancing the activity. The effects of the neighboring atoms on activity is further investigated with density functional theory simulations (vide infra). Furthermore, in line with previous observations (Figure 3a), AC-supported single atoms display lower reactivity than their NC-supported counterparts. In comparison with the benchmark Pt/SiO_2, carbon-supported platinum nanoparticles with a comparable size (≈2 nm) achieve a considerably higher *TOF* (≈3 times), while the *TOF* of the carbon-supported SA-based catalysts are of the same order of magnitude as those observed over Pt/SiO_2 with an average nanoparticle size of 2.2 nm.[31] These results emphasize the impact of the carbon carrier on the catalytic activity of the metal, improving the utilization of the platinum atoms in carbon-supported systems compared to those supported on inert SiO_2, which has minimal interaction with the active phase.

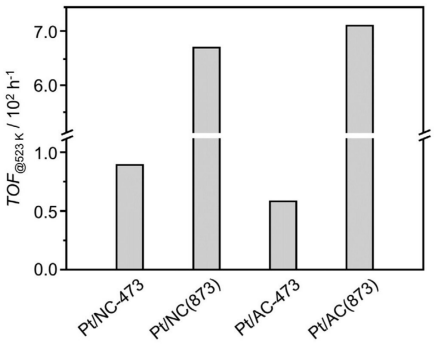


Figure 5. Turnover frequencies over selected catalysts. Each catalytic data point was gathered using materials in fresh form to exclude the possible influence of catalyst deactivation. Reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:5:65, F_T:W_{cat} = 200 $cm^3\ min^{-1}\ g_{cat}^{-1}$, T = 523 K, P = 1 bar, and *tos* = 15 min.

To obtain a molecular-level understanding of the different activity and selectivity patterns, density functional theory (DFT) simulations were performed. The reaction network with the pathways leading to CH_3Br, CH_4, and C (coke) is described (**Figure 6**) and the associated thermodynamic and kinetic parameters were calculated (Table S2, Supporting Information) for the most stable Pt species in the three materials. The Gibbs energies for the reaction profiles over the NC- and AC-supported platinum single atoms and the comparative Pt(111) surface representing the nanoparticles are shown in **Figure 7**. The temperature and pressures in the Gibbs terms were those from experiments. Unfortunately, the direct use of the span model[35] or direct comparison to the *TOF* is not possible due to several reasons including: i) the diversity of potentially active sites in the N-doped material (herein only one of the configurations is presented for conciseness); and ii) the selectivity issue, particularly with the formation of coke, that is difficult to be addressed through the two-state span model.

The network starts by the dissociative adsorption of CH_2Br_2 on the Pt(111) surface, leading to CH_2Br*. However, this intermediate is labile and the second Br is easier to lose than the first one. This occurs because breaking the second C—Br bond is compensated by the formation of a bridge species with two platinum atoms, Pt-CH_2-Pt, thereby fulfilling the valence of the C atom via an ensemble considering the involvement of several metal atoms (CH_2Br sits on top; thus only one atom forms the ensemble). Such cascade decomposition effects have been described in literature and are a direct consequence of the continuous nature of the metal surface and linear-scaling contributions. Typically, they lead to the intermediate that optimally fills the best coordination sites (CH*).[36] Therefore, the main gaseous product for the platinum nanoparticle is the recombination with the hydrogen atoms on the surface to generate CH_4.

In N-doped carbon, several potential coordination sites for platinum atoms co-exist (Figure S5, Supporting Information). While the most widely studied fourfold coordinative pockets cannot trap platinum in an active form, the 3N cavity (denoted as "rol3") is optimally suited to yield an active platinum ensemble which may accommodate intermediates with only one valence left, that is, CH_3 or alike, but is not prone to generate CH_2 species. The formation of the methylene (CH_2*) fragment implies the formation of a double bond with the platinum atom that, due to the geometric constraints of the hybridization, fulfills the valence in a less effective way than an increased number of metal atoms in an ensemble with a continuum platinum surface. This agrees with our results; CH_2Br_2 adsorbs on different platinum sites in the NC- and AC-carrier (Figure 7). The adsorption is quite exothermic in all cases and leads to H_2BrC-Pt-Br fragments. From this intermediate, the following

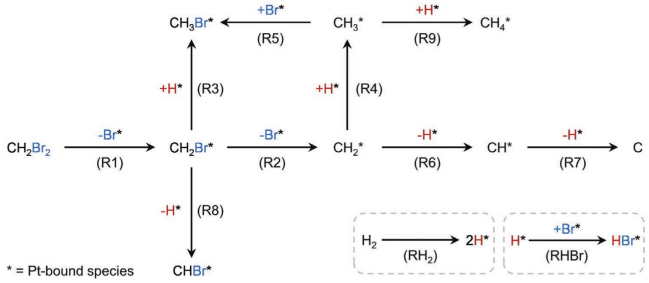


Figure 6. Reaction network of CH_2Br_2 hydrodebromination showing the pathways leading to CH_3Br, CH_4, and C (coke) formation. Dashed boxes show elementary steps for the formation of HBr and the activation of H_2. The labels in parentheses indicate thermodynamic and kinetic parameters that are detailed in Table S2, Supporting Information.

decompositions can in principle happen: i) Br elimination of the scaffold to H_2C-Pt-Br; ii) Br_2 elimination with concomitant formation of H_2C-Pt; or iii) H elimination leading to HBrC-Pt-Br. Out of these structures, reactions (i) and (ii) are very high in energy due to the difficulties of platinum atom to accommodate the methylene group via the ensemble and also since the platinum in the scaffold is already electron rich due to the interaction with the nitrogen atoms in the host. Only path (iii) is reachable, leading to the CHBr intermediate that is less electronically compromised (due to the electron acceptor nature of Br). The Pt-CHBr species is very reactive and can be the origin of the instability found for the NC-supported single atom catalysts. The H_2BrC-Pt-Br system can also dissociatively adsorb H_2, with one H atom ending in the platinum site and the other left in the basic sites of the host. This would be promoted by the basicity of the cavity, in line with the enhanced reactivity observed over NC-supported platinum catalysts compared to the AC-supported analogues. The dissociation is relatively easy, followed by the facile formation of HBr and CH_3Br on the single atom, thereby re-establishing the active site. The full reaction mechanism including all activation steps can be facilitated by the single atom site in both the NC- and AC-supported systems, which includes its immediate coordination environment (N- or O-moieties). Therefore, participation of the host in SA-based catalysts is crucial to enable the adsorption of CH_2Br_2 and H_2 since it traps H atoms that can be later transferred to other moieties. On the models with AC as host (the anchoring points are oxygen defects as shown in Figure S5, Supporting Information), the H_2BrC-Pt-Br dissociation occurs easily, leading to square planar structures which are prone to decomposition and release of HBr. The remaining moieties would trigger polymerization reactions that would lead to undesired products. On the other hand, if H_2 is activated it leads to the weakening of the Pt—AC bond, which likely leads to unstable single atoms, making them mobile and thus resulting in the potential formation of nanoparticles by diffusion and coalescence.

Further insights were gained by conducting kinetic analysis over the selected systems, showing differences in the apparent activation energies with values of 39 (Pt/AC-473), 37 (Pt/NC-473), 27 (Pt/NC(873)), and 24 kJ mol^{-1} (Pt/AC(873)) (**Figure 8**a). The values for the NP-based systems, ≈24 and 27 kJ mol^{-1}, are lower than those attained over platinum nanoparticles supported on SiO_2 (34 kJ mol^{-1}).[31] Rather, these apparent activation energies are in the same range as the values obtained over SA-based catalysts (37 and 39 kJ mol^{-1}), in line with the comparable *TOF* over these materials. From the reaction profiles, the energies required for the most energetic steps are taken as representative for the apparent activation energies. The calculations confirm that the nanoparticles have a slightly lower apparent activation energy than the SA-based systems (≈15 kJ mol^{-1}, CH_2Br_2 dissociation step), while the configuration employed to represent the Pt/NC-473 system shows a larger value (59 kJ mol^{-1}, HBr elimination). The deviations are due to the lack of support and coverage effects on the platinum nanoparticle models and the use of a single configuration to represent the material. The rapid deactivation of the AC-supported SA-based catalyst prevents us from using the computed reaction profiles as side paths dominate the apparent activation energy.

Furthermore, upon increasing the partial H_2 pressure from 12 to 72 kPa, the normalized reaction rate increases significantly over the NP-based catalysts (up to ≈6 times), whereas the SA-based systems were less affected (up to 2.5 times) by these changes in feed composition (Figure 8b). In addition, the increase of partial H_2 pressure showed no influence on the product distribution of SA-based systems, confirming their low sensitivity to increasing inlet H_2 concentrations. The derived partial orders in H_2 are higher over NP-based catalysts, with values of 1.00 and 0.83 for Pt/AC(873) and Pt/NC(873), respectively, while the SA-based systems display partial orders of 0.60 (Pt/AC-473) and 0.41 (Pt/NC-473). The selectivity to CH_3Br attained over these systems correlates inversely with the partial orders in H_2, being higher for lower partial orders in H_2. This agrees with the reaction profiles and the computed main paths obtained by DFT.

The lower reaction order in H_2 of SA-based catalysts and their higher apparent activation energy, compared to their NP-based counterparts, can be rationalized by their ability to activate H_2 and store H-atoms that can react with surface species. The geometry of nanoparticles enables facile H_2 dissociation

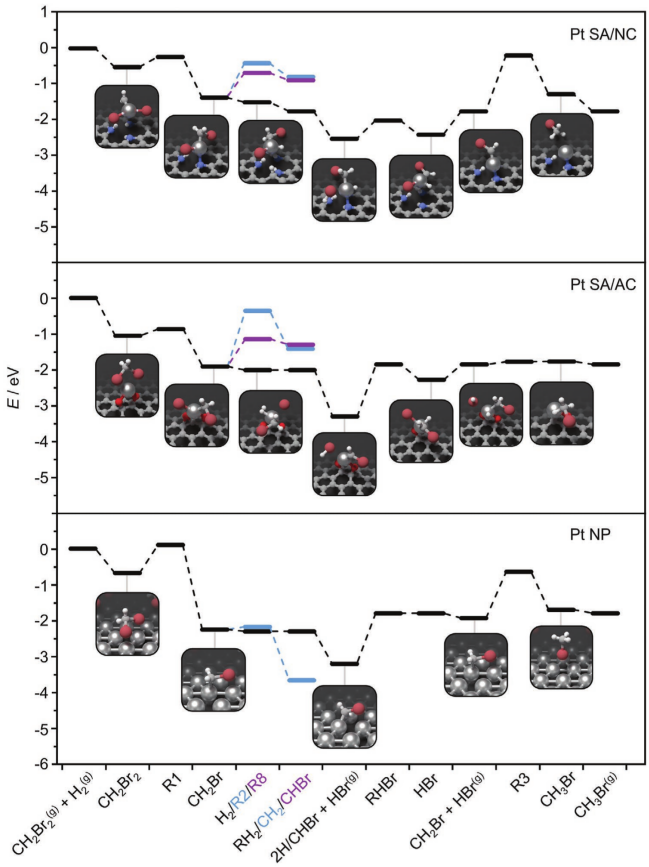


Figure 7. Gibbs energy profiles and schematic depictions of CH_2Br_2 hydrodebromination over platinum single atoms (SA) and a platinum(111) surface representing the nanoparticles (NP). Color code: C, gray; N, blue; O, red; H, white; Pt, metallic gray; Br, brown. The temperatures and pressures are those of the experiments: CH_2Br_2:H_2:Ar:He = 6:24:5:65, T = 523 K, and P = 1 bar.

(via the homolytic process) and allows higher hydrogen coverage, which is limited on single atoms. As for the single atoms in the carbon, the cavities are crucial and the presence of N-functionalities improves ability to heterolytically split H_2 into H-atoms for the SA-based systems.[7] H_2 activation is crucial and explains the higher reactivity of Pt/NC-473 compared to Pt/AC-473. In addition, the production of CH_3Br requires a single H-atom, therefore the partial order in H_2 of selective SA-based systems is ≈0.5, indicating that the reaction of the second H-atom is rate limiting. In contrast, the partial order in H_2 of NP-based systems is *ca.* 1.0, matching their higher propensity to generate CH_4. Moreover, NC-supported systems display lower partial orders in H_2 and higher CH_3Br selectivity than their AC-supported counterparts. This is likely due to the higher halogen uptake on NC-supported systems, a factor which was found to be key in steering the reaction pathway in selective dihalomethane hydrodehalogenations.[32] These kinetic fingerprints suggest that the reaction occurs with different mechanisms over platinum single atoms compared to nanoparticles, and corroborate the crucial role of the host in order to potentially alter the adsorption and/or desorption of the different species.

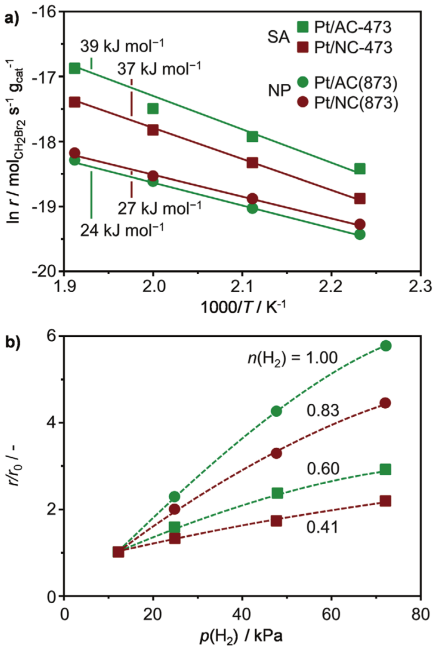


Figure 8. Rate of CH_2Br_2 hydrodebromination of selected catalysts as a function of a) temperature and b) inlet partial pressure of H_2. Each catalytic data point was gathered using materials in fresh form to exclude the possible influence of catalyst deactivation. Reaction conditions: a) CH_2Br_2:H_2:Ar:He = 6:24:5:65, F_T:W_{cat} = 100–500 cm^3 min^{-1} g$_{cat}^{-1}$; T = 448-523 K; b) CH_2Br_2:H_2:Ar:He = 6:6–72:5:17–83, F_T:W_{cat} = 200–800 cm^3 min^{-1} g$_{cat}^{-1}$, T = 523 K. All tests were conducted at P = 1 bar and *tos* = 15 min. SA and NP stand for single atoms and nanoparticles, respectively.

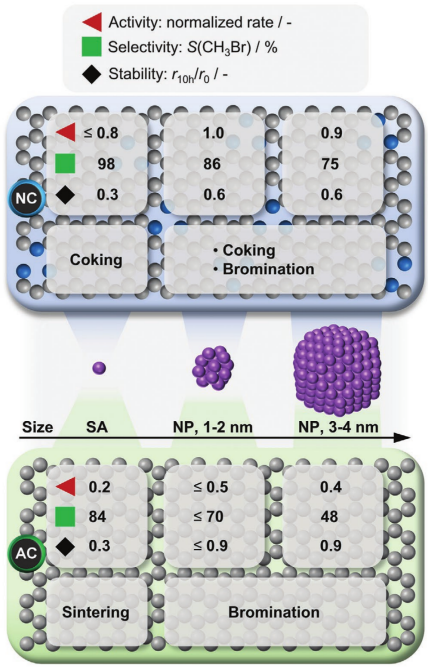


Figure 9. Summary scheme of nuclearity and host effects on the reactivity of carbon-supported platinum catalysts for CH_2Br_2 hydrodebromination, with indication of the main modes of deactivation. SA and NP stand for single atom and nanoparticle, respectively. The conditions specified in the captions of Figures 3 and 4 apply here.

3. Conclusions

In conclusion, a platform of NC- and AC-supported platinum nanostructures, ranging from single atoms to nanoparticles of ≈4 nm was derived to systematically assess nuclearity and host effects on the catalyst activity, selectivity, and stability performance in CH_2Br_2 hydrodebromination, through coupling in-depth characterization, kinetic analysis, and density functional theory (DFT). Catalytic evaluation revealed that i) SA-based catalysts exhibit unparalleled CH_3Br selectivity of up to 98%, whereas the formation of CH_4 (selectivity < 60%) is increasingly favored over larger nanoparticles, ii) the specific activity reaches a maximum over platinum nanoparticles with an average size in the range of 1.3–2.3 nm, thereby achieving up to a threefold higher CH_3Br production rate and *TOF* than Pt/SiO_2, and iii) NC-supported platinum catalysts are more active and selective to CH_3Br compared to their AC-supported analogues. Moreover, NP-based systems displayed improved stability compared to their SA-based analogues, where the functionalization of the carbon host determined the mode of deactivation. Specifically, on AC, sintering of the single atoms was the major deactivation path, while on NC, the single atoms remained structurally stable, but coking prevailed due to the generation of unstable intermediate species during the reaction. A summary scheme of the carbon host and platinum nuclearity effects on catalyst activity, selectivity, and stability in CH_2Br_2 hydrodebromination with the main modes of deactivation is presented in **Figure 9**.

Kinetic and mechanistic studies emphasized the role of N-functionalities in the host, which can store H-atoms thereby enhancing activity. NC-supported materials displayed higher bromine uptake than their AC-supported counterparts, which likely suppressed over-hydrogenation pathways as corroborated by the lower reaction orders with respect to H_2. Furthermore, in contrast to single atoms, the facile dissociation of H_2 and, due to the geometry, higher availability of surface H-atoms

over nanoparticles promotes CH_4 generation. The findings presented in this study highlight the potential of nanostructuring in CH_2Br_2 semi-hydrogenation, opening up options for the development of improved catalytic systems.

4. Experimental Section

Catalyst Preparation: Carbon-supported Pt-catalysts with Pt-nanostructures varying from single atoms to nanoparticles were prepared following the protocols reported by Kaiser et al.[17] Nitrogen-doped carbon (NC) and commercially available activated carbon (AC, Norit ROX 0.8) were ground and sieved into particles of 0.4–0.6 mm prior to their use as supports. The platinum precursor, chloroplatinic acid (H_2PtCl_6, ABCR, 99.9%, 40.0 wt% Pt), was dispersed on the supports via incipient wetness impregnation. Appropriate amounts of the metal precursor required to obtain a platinum content of 1 wt% in the final catalyst were dissolved in deionized water (1.5 $cm^3\,g^{-1}$) and added dropwise to the carbon carriers under continuous magnetic stirring for 2 h. The impregnated solids were dried at 473 K for 16 h in static air (heating rate 5 K min^{-1}) followed by a thermal activation step at higher temperatures, as indicated in the respective sample code (T_{act} = 473–1073 K), yielding two series of Pt catalysts, Pt/NC-T_{act}, and Pt/AC-T_{act}. The Pt/NC-473 and Pt/AC-473 catalysts underwent an additional thermal reductive treatment in 20 vol% H_2/He (PanGas, purity 5.0) flow for 3 h at elevated temperatures (T_{red} = 573 or 873 K, heating rate 10 K min^{-1}) and were denoted as Pt/NC(T_{red}) and Pt/AC(T_{red}). The synthesis steps of the catalysts developed in this study with their respective sample codes is given in Figure 1a.

Catalyst Characterization: Powder X-ray diffraction (XRD) was measured using a PANalytical X'Pert PRO-MPD diffractometer with Cu-K$\alpha\alpha$ radiation (λ = 1.54060 Å). The data was recorded in the 10°–70° 2θ range with an angular step size of 0.017° and a counting time of 0.26 s per step. N_2 sorption at 77 K was measured in a Micromeritics TriStar II analyzer. Samples (≈0.1 g) were evacuated to 50 mbar at 573 K for 12 h prior to the measurement. The Brunauer–Emmett–Teller (BET) method was applied to calculate the total surface area, S_{BET}. The pore volume, V_{pore}, was determined from the amount of N_2 adsorbed at a relative pressure of p / p_0 = 0.98. The platinum content in the catalysts was determined by inductively coupled plasma-optical emission spectrometry (ICP-OES) using a Horiba Ultra 2 instrument equipped with photomultiplier tube detection. The solids were dissolved in a HNO_3:H_2O_2 = 3:1 mixture under sonication until the absence of visible solids. CO pulse chemisorption was performed on a Thermo TPDRO 1100 set-up equipped with a thermal conductivity detector. Prior to the analyses, the NP-based samples (≈0.2 g) were pretreated at 423 K under flowing He (20 cm^3 STP min^{-1}) for 30 min, and reduced at 523 K under flowing 5 vol% H_2/He (20 cm^3 STP min^{-1}) for 30 min. Thereafter, 0.344 cm^3 of 1 vol% CO/He were pulsed over the catalyst bed every 4 min at 308 K. To avoid desorption of CO, the interval between successive pulses was minimized. The platinum dispersion was calculated using an atomic surface density of 1.47 × 10^{19} atoms m^{-2} and an adsorption stoichiometry of Pt/CO = 1. Scanning transmission electron micrographs with a high-angle annular dark-field detector (HAADF–STEM) were acquired on a HD2700CS (Hitachi) microscope operated at 200 kV. All samples were dispersed in ethanol and some droplets were deposited onto lacey carbon coated copper grids and dried in air. The size distribution of the platinum nanostructures was obtained by examining more than 100 particles. X-ray photoelectron spectra (XPS) were acquired on a Physical Electronics Quantum 2000 instrument using monochromatic Al-*Ka* radiation, generated from an electron beam operated at 15 kV, and equipped with a hemispherical capacitor electron-energy analyzer. The samples were analyzed at constant analyzer pass energy of 46.95 eV. The spectrometer was calibrated for the Au $4f_{7/2}$ signal at 84.0 ± 0.1 eV. The envelopes were fitted by mixed Gaussian–Lorentzian component profiles after Shirley background subtraction. The different platinum species were fitted as previously reported (i.e., peak positions fixed ±0.1 eV, full width at half maximum (FWHM) constrained, spin orbit coupling 3.33 eV).[17] A contribution for Pt(0) was only implemented for samples where nanoparticles were detected by HAADF-STEM. The contribution of the Br $3d^{5/2}$ signal in brominated used catalysts was fitted with two chemical states at 70.1±0.2 eV (Br1, attributed to Br—C)[37] and 67.4±0.2 eV (Br_2, attributed to adsorbed Br_2 or Br—Pt).[38,39] The Br 3d(3/2) peaks originating from spin orbit coupling were fixed at a distance of 1.05 eV and constrained to the same FWHM as the corresponding 3d(5/2) peaks. The bromine content was quantified based on the measured C 1s, N 1s, O 1s, Br 3p and the fitted Pt 4f signals using relative sensitivity factors (provided by PHI-MultiPak software), which were corrected for the system transmission function. X-ray absorption fine structure (XAFS) measurements at the Pt L_2 and L_3-edge were carried out at the SuperXAS beamline of the Swiss Light Source. The incident photon beam provided by a 2.9 T superbend magnet was selected by a Si(111) channel-cut Quick-EXAFS monochromator. The rejection of higher harmonics and focusing were achieved with rhodium-coated collimating and toroidal mirrors, respectively, at 2.5 mrad. The beamline was calibrated using Pt foil. The area of sample illuminated by the X-ray beam was 0.5 mm × 0.2 mm. The catalysts (≈0.3 g) were finely ground, mixed homogeneously with five parts of cellulose, and pressed into 13 mm diameter pellets. All spectra were recorded in transmission mode at room temperature. The X-ray absorption near-edge structure (XANES) spectra were calibrated by measuring Pt foil simultaneously with each sample. The extended X-ray absorption fine structure (EXAFS) spectra were acquired with a 1 Hz frequency (0.5 s per spectrum) and then averaged over 10 min. The procedures for analysis and fitting of the EXAFS spectra are reported elsewhere.[17]

Catalyst Evaluation: The hydrodebromination of CH_2Br_2 was carried out at ambient pressure in a continuous-flow fixed-bed reactor set up. H_2 (PanGas, purity 5.0), He (carrier gas, PanGas, purity 5.0), Ar (internal standard, PanGas, purity 5.0) were dosed by a set of digital mass flow controllers (Bronkhorst) and liquid CH_2Br_2 (Acros Organics, 99%) was supplied by a syringe pump (Fusion 100, Chemyx) to a vaporizer unit operated at 393 K. A quartz reactor (internal diameter, d_i = 12 mm) was loaded with the catalyst (catalyst weight, W_{cat} = 0.1–0.25 g, particle size, d_p = 0.4–0.6 mm) and heated to the desired temperature (T = 448–523 K) in an electrical oven under He flow. The catalyst bed was allowed to stabilize for at least 10 min at desired temperature before the reaction mixture was fed at total volumetric flow (F_T) of 20 cm^3 STP min^{-1} and desired feed composition of CH_2Br_2:H_2:Ar:He = 6:24:5:65 (vol%). A composition of CH_2Br_2:H_2:Ar:He = 6:6-72:5:17-83 was applied for the kinetic tests. Downstream linings were heated at 393 K to prevent the condensation of unconverted reactants and/or products. The content of carbon-containing compounds (CH_2Br_2, CH_3Br, and CH_4) and of Ar in the reactor-outlet gas stream was quantified online via a gas chromatograph equipped with a GS-Carbon PLOT column coupled to a mass spectrometer (GC–MS, Agilent GC 6890, Agilent MSD 5973N). After the GC–MS analysis, the gas stream was passed through two impinging bottles in series containing an aqueous solution of NaOH (1 M) for neutralization prior to its release in the ventilation system.

The conversion of dibromomethane, $X(CH_2Br_2)$, was calculated using Equation (1)

$$X(CH_2Br_2) = \frac{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out}}{n(CH_2Br_2)_{in}} \times 100\% \quad (1)$$

where $n(CH_2Br_2)_{in}$ and $n(CH_2Br_2)_{out}$ are the molar flows of CH_2Br_2 at the reactor inlet and outlet, respectively. The selectivity, $S(j)$, to product j (j: CH_3Br, CH_4) was calculated according to Equation (2)

$$S(j) = \frac{n(j)_{out}}{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out}} \times 100\% \quad (2)$$

where $n(j)_{out}$ is the molar flow of product j at the reactor outlet. The reaction rate, r, based on the platinum loading and expressed with

respect to the CH_3Br production or to the CH_2Br_2 consumption, were calculated using Equations (3) and (4), respectively

$$r = \frac{n(CH_2Br_2)_{in} \times X(CH_2Br_2) \times S(CH_3Br)}{100 \times 100 \times W_{cat} \times \omega_{Pt}} \ mol_{CH_3Br}\ h^{-1}\ mol_{Pt}^{-1} \quad (3)$$

$$r = \frac{n(CH_2Br_2)_{in} \times X(CH_2Br_2)}{100 \times W_{cat} \times \omega_{Pt}} \ mol_{CH_3Br_2}\ h^{-1}\ mol_{Pt}^{-1} \quad (4)$$

where W_{cat} is the weight of the catalyst and ω_{Pt} is the platinum loading determined by ICP-OES analysis (Table 2). The turnover frequency, *TOF*, was calculated using Equation (5),

$$TOF = \frac{n(CH_2Br_2)_{in} \times X(CH_2Br_2)}{100 \times W_{cat} \times \omega_{Pt} \times D_{Pt}} \ h^{-1} \quad (5)$$

where D_{Pt} is the platinum dispersion, determined by CO pulse chemisorption. A platinum dispersion of 100% was used for the SA-based catalysts. The error of the carbon balance, ε_C, in all catalytic tests was determined using Equation (6)

$$\varepsilon_C = \frac{n(CH_2Br_2)_{in} - n(CH_2Br_2)_{out} - n(j)_{out}}{n(CH_2Br_2)_{in}} \times 100\% \quad (6)$$

where $n(j)_{out}$ is the molar flow of product j (j: CH_3Br, CH_4) at the reactor outlet.

After the tests, the reactor was quenched to room temperature in He flow and the catalyst was retrieved for further characterization analyses. Evaluation of the dimensionless moduli based on the criteria of Carberry, Mears, and Weisz–Prater indicated that the catalytic tests were performed in the absence of mass and heat transfer limitations. Additionally, CH_2Br_2 hydrodebromination tests over Pt/NC-473 and Pt/AC-473 performed at variable flow rates and constant space velocity (F_T:W_{cat}) as well as by using catalyst particles of different sizes (d_p = 0.15–0.5 mm) at constant space velocity verified the absence of extra- and intraparticle mass transfer limitations, respectively (Figure S7, Supporting Information).

Computational Details: Density functional theory (DFT) on slab models representing the different systems was employed as implemented in the Vienna Ab initio Simulation Package (VASP 5.4.4).[40,41] Generalized gradient approximation with the Perdew–Burke–Ernzerhof (GGA-PBE)[42] functional was used to obtain the exchange-correlation energies with dispersion contributions introduced,[43,44] and spin polarization was allowed when needed. Core electrons were described by projector augmented waves (PAW),[45,46] while valence monoelectronic states were expanded in plane waves with cut-off energy of 450 eV. The Brillouin zone was sampled with a gamma-centered grid of 3 × 3 × 1 k-point grid.

Single atom catalysts were modeled by a one-layer 6 × 6 slab of graphitic carbon separated by 19 Å of vacuum. Defects in the carbon sheet were introduced by replacing some C- by N-atoms and saturating the valence (N-doped carbon) and/or adding oxygen as epoxides (AC), and the Pt atom was then placed on the cavity or coordinating the O atoms. For Pt nanoparticles, the lowest energy surface (111) in a p(3 × 3) supercell was used. The slab had four metal layers built from a bulk with an optimized lattice parameter of 2.6997 Å. Only the adsorbates and the top two layers were allowed to relax. The layers were interspaced along the *z*-direction with a vacuum space of at least 15 Å, applying a correction to the arising dipole. The Brillouin zone was sampled with a gamma centered grid of 5 × 5 × 1. For all the investigated systems the structures were relaxed using convergence criteria of 10^{-4} eV (0.03 eV $Å^{-1}$ for the metallic systems) and 10^{-5} eV for the ionic and electronic steps, respectively.

Transition states were located following the climbing image nudged elastic band procedure (CI–NEB).[47] Some of the transition states were obtained through the improved dimer method,[48,49] using the structures of the transition states found in similar defects as the initial geometry guess.[50] Transition states were confirmed by diagonalizing the numerical Hessian matrix obtained by displacements of ±0.02 Å. All structures presented in this work can be retrieved from the ioChem-BD database.[51,52]

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by ETH research grants ETH-43 18-1 and ETH-40 17-1, and by a predoctoral grant of MINECO La Caixa-Severo Ochoa through the Severo Ochoa Excellence Accreditation 2014–2018 (SEV-2013-0319). The authors thank the Paul Scherrer Institute, PSI, and the Scientific Center for Optical and Electron Microscopy at the ETH Zurich, ScopeM, for access to their facilities. Dr. Frank Krumeich is acknowledged for performing microscopic analyses. The authors thank BSC-RES for providing generous computational resources.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

carbon carriers, hydrodebromination, mechanism, nanoparticles, single atoms, speciation

[1] R. A. Sheldon, *Chem. Soc. Rev.* **2012**, *41*, 1437.
[2] A. T. Bell, *Science* **2003**, *299*, 1688.
[3] L. Liu, A. Corma, *Chem. Rev.* **2018**, *118*, 4981.
[4] X. Cui, W. Li, P. Ryabchuk, K. Junge, M. Beller, *Nat. Catal.* **2018**, *1*, 385.
[5] J. Liu, *ACS Catal.* **2017**, *7*, 34.
[6] G. Kyriakou, M. B. Boucher, A. D. Jewell, E. A. Lewis, T. J. Lawton, A. E. Baber, H. L. Tierney, M. Flytzani-Stephanopoulos, E. C. H. Sykes, *Science* **2012**, *335*, 1209.
[7] S. Mitchell, E. Vorobyeva, J. Pérez-Ramírez, *Angew. Chem., Int. Ed.* **2018**, *57*, 15316.
[8] J. M. Thomas, R. Raja, D. W. Lewis, *Angew. Chem., Int. Ed.* **2005**, *44*, 6456.
[9] J. M. Thomas, *Design and Applications of Single-Site Heterogeneous Catalysts*, World Scientific, London **2012**.
[10] A. Wang, J. Li, T. Zhang, *Nat. Rev. Chem.* **2018**, *2*, 65.
[11] H. Zhang, G. Liu, J. Ye, *Adv. Energy Mater.* **2018**, *8*, 1701343.
[12] C. Zhu, S. Fu, Q. Shi, D. Du, L. Y, *Angew. Chem., Int. Ed.* **2017**, *56*, 13944.
[13] C. Ye, D. Wang, Y. Li, *Chem. Commun.* **2020**, *56*, 7687.
[14] G. Vilé, D. Albani, M. Nachtegaal, Z. Chen, D. Dontsova, M. Antonietti, N. López, J. Pérez-Ramírez, *Angew. Chem., Int. Ed.* **2015**, *54*, 11265.

[15] R. Lin, D. Albani, E. Fako, S. K. Kaiser, O. V. Safonova, N. López, J. Pérez-Ramírez, *Angew. Chem., Int. Ed.* **2019**, *58*, 504.
[16] S. K. Kaiser, R. Lin, S. Mitchell, E. Fako, F. Krumeich, R. Hauert, O. V. Safonova, V. A. Kondratenko, E. V. Kondratenko, S. M. Collins, P. A. Midgley, N. López, J. Pérez-Ramírez, *Chem. Sci.* **2019**, *10*, 359.
[17] S. K. Kaiser, E. Fako, G. Manzocchi, F. Krumeich, R. Hauert, A. H. Clark, O. V. Safonova, N. López, J. Pérez-Ramírez, *Nat. Catal.* **2020**, *3*, 376.
[18] C. H. Bartholomew, R. J. Farrauto, *Fundamentals of Industrial Catalytic Processes*, John Wiley & Sons, New York **2011**.
[19] S. Nishimura, *Handbook of Heterogeneous Catalytic Hydrogenation for Organic Synthesis*, Wiley, New York **2001**.
[20] P. T. Witte, P. H. Berben, S. Boland, E. H. Boymans, D. Vogt, J. W. Geus, J. G. Donkervoort, *Top. Catal.* **2012**, *55*, 505.
[21] D. Astruc, *Nanoparticles and Catalysis*, Wiley-VCH, Weinheim **2008**.
[22] S. Schauermann, N. Nilius, S. Shaikhutdinov, H.-J. Freund, *Acc. Chem. Res.* **2013**, *46*, 1673.
[23] M. Sankar, N. Dimitratos, P. J. Miedziak, P. P. Wells, C. J. Kiely, G. J. Hutchings, *Chem. Soc. Rev.* **2012**, *41*, 8099.
[24] G. Vilé, D. Albani, N. Almora-Barrios, N. López, J. Pérez-Ramírez, *ChemCatChem* **2016**, *8*, 21.
[25] N. Semagina, L. Kiwi-Minsker, *Catal. Rev.: Sci. Eng.* **2009**, *51*,147.
[26] P. Mäki-Arvela, J. Hájek, T. Salmi, D. Y. Murzin, *Appl. Catal., A* **2005**, *292*, 1.
[27] A. J. McCue, J. A. Anderson, *Front. Chem. Sci. Eng.* **2015**, *9*, 142.
[28] E. McFarland, *Science* **2012**, *338*, 340.
[29] R. Lin, A. P. Amrute, J. Pérez-Ramírez, *Chem. Rev.* **2017**, *117*, 4182.
[30] K. Ding, A. R. Derk, A. Zhang, Z. Hu, P. Stoimenov, G. D. Stucky, H. Metiu, E. W. McFarland, *ACS Catal.* **2012**, *2*, 479.
[31] A. J. Saadun, S. Pablo-García, V. Paunovic, Q. Li, A. Sabadell-Rendón, K. Kleemann, F. Krumeich, N. López, J. Pérez-Ramírez, *ACS Catal.* **2020**, *10*, 6129.
[32] A. J. Saadun, G. Zichittella, V. Paunović, B. A. Markaide-Aiastui, S. Mitchell, J. Pérez-Ramírez, *ACS Catal.* **2020**, *10*, 528.
[33] M. Muhler, Z. Paál, R. Schlögl, *Appl. Surf. Sci.* **1991**, *47*, 281.
[34] G. Vilé, D. Baudouin, I. N. Remediakis, C. Copéret, N. López, J. Pérez-Ramírez, *ChemCatChem* **2013**, *5*, 3750.
[35] S. Kozuch, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 795.
[36] F. Abild-Petersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skúlason, T. Bilgaard, J. K. Nørskov, *Phys. Rev. Lett.* **2007**, *99*, 016105.
[37] H. Au, N. Rubio, M. S. Shaffer, *Chem. Sci.* **2018**, *9*, 209.
[38] E. Papirer, R. Lacroix, J.-B. Donnet, G. Nanse, P. Fioux, *Carbon* **1994**, *32*, 1341.
[39] J. Chastain, *Handbook of X-Ray Photoelectron Spectroscopy*, Perkin-Elmer Corporation, Waltham, MA **1992**.
[40] G. Kresse, J. Fürthmuller, *Comput. Mater. Sci.* **1996**, *6*, 15.
[41] G. Kresse, J. Fürthmuller, *Phys. Rev. B* **1996**, *54*, 11169.
[42] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[43] S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.* **2011**, *32*, 1456.
[44] N. Almora-Barrios, G. Carchini, P. Błoński, N. López, *J. Chem. Theory Comput.* **2014**, *10*, 5002.
[45] G. Kresse, D. Joubert, *Phys. Rev. B* **1999**, *59*, 1758.
[46] P. E. Blöchl, *Phys. Rev. B* **1994**, *50*, 17953.
[47] G. Henkelman, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9978.
[48] A. Heyden, *J. Chem. Phys.* **2005**, *123*, 224101.
[49] G. Henkelman, H. Jónsson, *J. Chem. Phys.* **1999**, *111*, 7010.
[50] S. Pablo-García, M. Álvarez-Moreno, N. López, *Int. J. Quantum Chem.* **2021**, *121*, e26382.
[51] A. Ruiz-Ferrando, *Pt Dehydrobromination Dataset*, https://doi.org/10.19061/iochem-bd-1-179 (accessed: November 2020)
[52] S. Pablo-García, CH_2Br_2 Dataset, **2020**, https://doi.org/10.19061/iochem-bd-1-150 (accessed: November 2020).

(13 of 13)

Electrochemical Reduction of Carbon Dioxide to 1-Butanol on Oxide-Derived Copper

*Louisa Rui Lin Ting[+], Rodrigo García-Muelas[+], Antonio J. Martín[+], Florentine L. P. Veenstra, Stuart Tze-Jin Chen, Yujie Peng, Edwin Yu Xuan Per, Sergio Pablo-García, Núria López, Javier Pérez-Ramírez und Boon Siang Yeo**

Abstract: *The electroreduction of carbon dioxide using renewable electricity is an appealing strategy for the sustainable synthesis of chemicals and fuels. Extensive research has focused on the production of ethylene, ethanol and n-propanol, but more complex C_4 molecules have been scarcely reported. Herein, we report the first direct electroreduction of CO_2 to 1-butanol in alkaline electrolyte on Cu gas diffusion electrodes (Faradaic efficiency = 0.056 %, $j_{1\text{-}Butanol} = -0.080\ mA\ cm^{-2}$ at −0.48 V vs. RHE) and elucidate its formation mechanism. Electrolysis of possible molecular intermediates, coupled with density functional theory, led us to propose that CO_2 first electroreduces to acetaldehyde-a key C_2 intermediate to 1-butanol. Acetaldehyde then undergoes a base-catalyzed aldol condensation to give crotonaldehyde via electrochemical promotion by the catalyst surface. Crotonaldehyde is subsequently electroreduced to butanal, and then to 1-butanol. In a broad context, our results point to the relevance of coupling chemical and electrochemical processes for the synthesis of higher molecular weight products from CO_2.*

Introduction

The electrochemical carbon dioxide reduction reaction (CO_2RR) to fuels and chemicals, when powered by renewable electricity, is a potentially sustainable way to alleviate our pressing global energy demands and to avert climate change.[1] Copper-based materials are the only family of catalysts that can reduce CO_2 to multi-carbon molecules with significant Faradaic efficiencies (*FE*) and current densities (*j*).[2,3] Among the multi-carbon products, C_2 molecules such as ethylene and ethanol can be facilely formed (*FE* = 30–50 %).[2,4] The main C_3 product reported is *n*-propanol (*FE* = 10–13 %),[5,6] alongside small quantities of propionaldehyde, allyl alcohol, acetone, propylene and propane (total FE < 3 %).[4,5] Reports on the formation of C_4 molecules, many of which have much higher commercial value, are scarce and have been limited to hydrocarbons showing *FE* < 1 %.[7] The drastic decrease in the selectivity of a product as the number of carbon atoms in it increases suggests that the coupling mechanism to form a multi-carbon product follows a „polymerization“ scheme of *CO that obeys the Flory-Schulz distribution.

Interestingly, the direct production of 1-butanol ($CH_3CH_2CH_2CH_2OH$) from electrochemical CO_2 reduction has not been reported. This oxygenate, which has a high volumetric energy density of 29.2 MJ L^{-1} and is less hygroscopic and corrosive than ethanol, has been suggested for direct use as a fuel or in diesel-blends.[8] Schmid and coworkers have utilized bacteria to convert CO (generated from CO_2 electrolysis) to 1-butanol.[9] More recently, a mechanistic study of CO_2 reduction to *n*-propanol revealed that minor and yet-to-be-quantified amounts of 1-butanol can be co-produced from the electrochemical reduction of acetaldehyde and CO in 0.1 M KOH on oxide-derived Cu electrodes.[10] Still, no strategy to successfully electrosynthesize 1-butanol or any C_4 oxygenates from CO_2 has been conceived. To tackle this challenge, it is crucial to understand and map out the mechanism and kinetics for its formation.

Herein, we report and quantify for the first time the formation of C_4 oxygenates from alkaline electrolysis of CO_2 using CuO-derived Cu gas diffusion electrodes (GDE) in a flow cell. The predominant C_4 product was 1-butanol (*FE* =

[*] L. R. L. Ting,[+] S. T.-J. Chen, Y. Peng, E. Y. X. Per, Prof. Dr. B. S. Yeo
Department of Chemistry, National University of Singapore
3 Science Drive 3, Singapore 117543 (Singapore)
E-Mail: chmyeos@nus.edu.sg
L. R. L. Ting,[+] Prof. Dr. B. S. Yeo
Solar Energy Research Institute of Singapore,
National University of Singapore
7 Engineering Drive 1, Singapore 117574 (Singapore)
Dr. R. García-Muelas,[+] S. Pablo-García, Prof. Dr. N. López
Institute of Chemical Research of Catalonia,
The Barcelona Institute of Science and Technology
Av. Països Catalans 16, 43007 Tarragona (Spain)
Dr. A. J. Martín,[+] F. L. P. Veenstra, Prof. Dr. J. Pérez-Ramírez
Institute for Chemical and Bioengineering,
Department of Chemistry and Applied Biosciences, ETH Zürich
Vladimir-Prelog-Weg 1, 8093 Zürich (Switzerland)
Prof. Dr. J. Pérez-Ramírez
Department of Chemical and Biomolecular Engineering,
National University of Singapore
4 Engineering Drive 4, Singapore 117585 (Singapore)
[+] These authors contributed equally to this work.

0.056 %, $j_{1\text{-Butanol}} = -0.080\ \text{mA cm}^{-2}$) at −0.48 V vs. RHE (Reversible Hydrogen Electrode; from here on, all potentials are referenced to the RHE and all currents are normalized to the exposed geometric surface area of the electrodes, unless otherwise stated). We then elucidate the reaction mechanism by combining analyses of reaction products from electrolyses of possible intermediates and density functional theory (DFT) investigations. The formation of the critical C_4 intermediate, crotonaldehyde ($CH_3CH{=}CHCHO$), was traced to the aldol condensation (C–C bond formation) of two acetaldehyde (CH_3CHO) molecules generated from CO_2 electroreduction (Figure 1 a). The aldol reaction is promoted by both OH^- ions in the electrolyte and the electrocatalyst surface. Crotonaldehyde then undergoes a two-step electroreduction to 1-butanol. We also unveil the critical role of pH at different stages of the reaction mechanism, pointing towards new strategies for increasing the performance of electro-assisted conversion of CO_2 into 1-butanol.

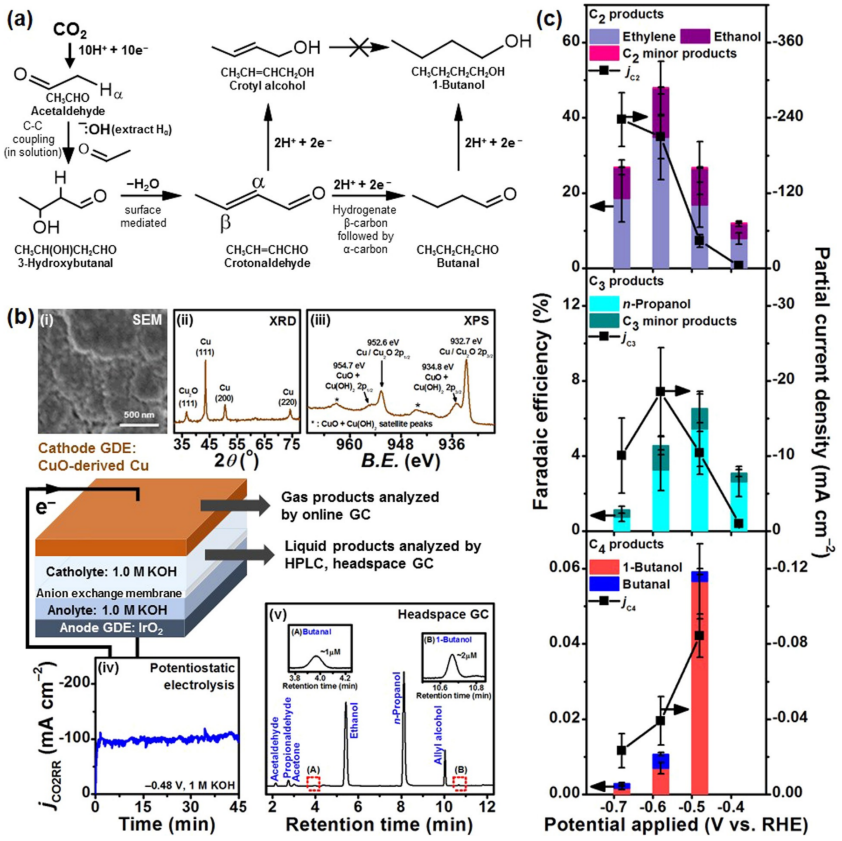


Figure 1. a) Simplified reaction scheme for CO_2 reduction to 1-butanol. Further details are provided in Section S3. b) Diagram of CO_2 flow cell electrolysis set up: characterization of CuO-derived Cu cathodes by (i) SEM, (ii) XRD and (iii) XPS (*B.E.* refers to binding energy). We note that the detection of Cu_2O in the XRD and CuO + $Cu(OH)_2$ signals in the XPS data are likely due to surface oxidation of the electrode from its exposure to air (see Section S2.1 for a more detailed discussion); (iv) large CO_2 reduction current densities (in the order of $-100\ \text{mA cm}^{-2}$) which gave a sufficient rate of product formation to allow detection of minor products; (v) sensitive analytical techniques like headspace GC can quantify minor products down to the μM-scale. c) Faradaic efficiencies (*FE*, colored bars) and partial current densities (j, ■) of C_2, C_3 and C_4 products from electrolysis of CO_2 on CuO-derived Cu GDE in 1.0 M KOH. The major C_2, C_3 and C_4 products are ethylene, *n*-propanol and 1-butanol, respectively. Other detected products are shown in Table S1.

Results and Discussion

We electroreduced CO_2 at various potentials using CuO-derived Cu GDE cathodes in a flow cell (Figure 1b, Sections S1,S2). The catalyst was electrodeposited onto the GDE using a previously-published procedure.[11] Aqueous 1.0 M KOH was used as the electrolyte. The high CO_2RR current densities from the flow cell electrolysis (Figure 1b-(iv)), which circumvents mass transport limitations, combined with the use of highly sensitive headspace gas chromatography (Figure 1b(v)) improves the detection and quantification of liquid products with low *FEs* and current densities. This allows us to detect CO_2 reduction products that have, to-date, never been observed.

The total *FEs* of carbonaceous products were 68-69% at −0.48 and −0.58 V (Figure 1c, Table S1). The major multi-carbon products are C_2 molecules, namely ethylene and ethanol, which are typically formed on oxide-derived copper catalysts.[2,12] The highest FE_{C_2} of 48% was observed at −0.58 V, with a corresponding j_{C_2} of $-210\ mA\,cm^{-2}$. Minor C_2 products ($FE \leq 0.1\%$) such as acetaldehyde and ethane were also detected. In the case of acetaldehyde, the low *FE* is a result of the chemical and/or electrochemical transformations it readily undergoes, as we will discuss below. C_3 species, mainly *n*-propanol, were also detected, with a maximum FE_{C_3} of 6.5% and j_{C_3} of $-18.5\ mA\,cm^{-2}$ obtained at −0.48 and −0.58 V, respectively. Overall, the catalytic activities toward C_2 and C_3 molecules from CO_2 reduction are comparable or higher than values previously reported for Cu catalysts loaded onto carbon GDEs (Table S3). Interestingly, we detected C_4 oxygenates such as 1-butanol and butanal (maximum total $FE_{C_4} = 0.060\%$ at the onset potential of −0.48 V), in contrast to previous studies which only identified hydrocarbons for the C_4 fraction.[7,13] The dominant product was 1-butanol, which is in line with the fact that aldehydes can be easily electro-reduced to their corresponding alcohols, as demonstrated in the cases of formaldehyde and acetaldehyde.[14] No carbon-containing products were found in control experiments performed without applied potentials.

As 1-butanol is the sole C_4 alcohol product, it could not come from C−C coupling of four individual C_1 adsorbates such as *CO in a Flory-Schulz distribution, since 2-butanol was not detected. This led us to postulate that the formation of 1-butanol could occur through a combination of electrochemical and chemical steps. Specifically, the aldol condensation of two C_2 intermediates, such as acetaldehyde, gives rise to the C_4 backbone of crotonaldehyde, which is further reduced to C_4 terminal oxygenates like butanal and 1-butanol. While mechanisms for the formation of major C_2 and C_3 products, including acetaldehyde, have been widely discussed in the literature,[6,10,15–18] pathways for producing C_4 products are rarely mentioned. Herein, we focus on acetaldehyde reactivity as it is much less explored mechanistically. Nonetheless, we also present the two steps preceding acetaldehyde formation, which involve the key ethenyloxy intermediate, CH_2CHO*, and discuss the lateral pathways for CO_2 reduction to C_1-C_3 products in Section S3.

To test our hypothesis for 1-butanol formation via acetaldehyde, we electroreduced 50 mM acetaldehyde in 0.1 M KOH on CuO-derived Cu electrodes (Section S4). The product distribution at an optimized potential of −0.44 V is summarized in Figure 2a (see also Table S5 for product distributions at other applied potentials). The product with the highest selectivity was 1-butanol ($FE = 9.6\%$, $j_{1\text{-Butanol}} = -1.06\ mA\,cm^{-2}$), consistent with expectations from a base-catalyzed aldol condensation C−C coupling step. The remaining electrolysis products were other C_4 oxygenates such as butanal ($CH_3CH_2CH_2CHO$) and crotyl alcohol ($CH_3CH{=}CHCH_2OH$), as well as ethanol. During the electrolysis, we observed some coloration of the anion-exchange membrane due to its exposure to the alkaline acetaldehyde-containing electrolyte, but control experiments excluded its interference with our electrolysis results (Figures S8,9).

Detection of crotonaldehyde (3.3 mM, equivalent to 13.2% acetaldehyde conversion) after electrolysis suggests that the C_4 backbone of 1-butanol could be formed via a base-catalyzed aldol condensation (Figure 2a, inset). We utilized the increase in local pH close to the electrode under electroreduction conditions to test this hypothesis and performed fixed-current electrolyses of 50 mM acetaldehyde in 0.01, 0.1 and 1.0 M potassium phosphate buffer (pH 7, Figure 2b, Section S4.3). Higher buffer concentrations mitigate the local pH increase during electroreduction, thus lowering the overall local pH. Our results reveal that electrolysis in 0.01 M potassium phosphate buffer gave the highest selectivity toward C_4 products ($FE = 0.4\%$, with 1-butanol as the main product) and percentage of acetaldehyde converted to crotonaldehyde (1.5 mM, equivalent to 6.2% acetaldehyde conversion). In contrast, neither C_4 products nor crotonaldehyde were detected from experiments performed in 1.0 M buffer (Figure 2b inset). These observations of an alkaline local pH promoting the production of 1-butanol on copper directly support its formation via the aldol condensation of two acetaldehyde molecules and suggest an enhanced reaction rate close to the catalytic surface under electrolysis conditions.

We further investigated, using DFT, the base-catalyzed aldol condensation mechanism to form the C_4 backbone. High-level methods including solvent and potential effects have been employed to study electrochemical networks up to C_2 species, including the formation of acetaldehyde from CO_2.[16] However, the multiple conformations of C_4 molecules and the complexity of the reaction network with chemical (bulk solvent and interface) and electrochemical steps limits us to the use of DFT coupled to the Computational Hydrogen Electrode (CHE). The formation of the C_4 backbone comprises two steps: the C−C coupling between two acetaldehyde molecules to form 3-hydroxybutanal ($CH_3CH(OH)CH_2CHO$), and the subsequent dehydration of the latter to crotonaldehyde (Figure 2c). In solution, the aldol condensation starts with the stripping of an α-hydrogen from acetaldehyde by OH^- to form an ethenyloxy anion (CH_2CHO^-, Figure 2c(i)). The α-carbon of CH_2CHO^- then attacks the carbonyl group of a second acetaldehyde molecule, which is subsequently protonated to 3-hydroxybutanal (Figure 2c(ii)). Then, 3-hydroxybutanal loses an α-hydrogen as a proton, to generate a carbene species (Figure 2c(iii)), which forms crotonaldehyde by hydroxyl elimination. Con-

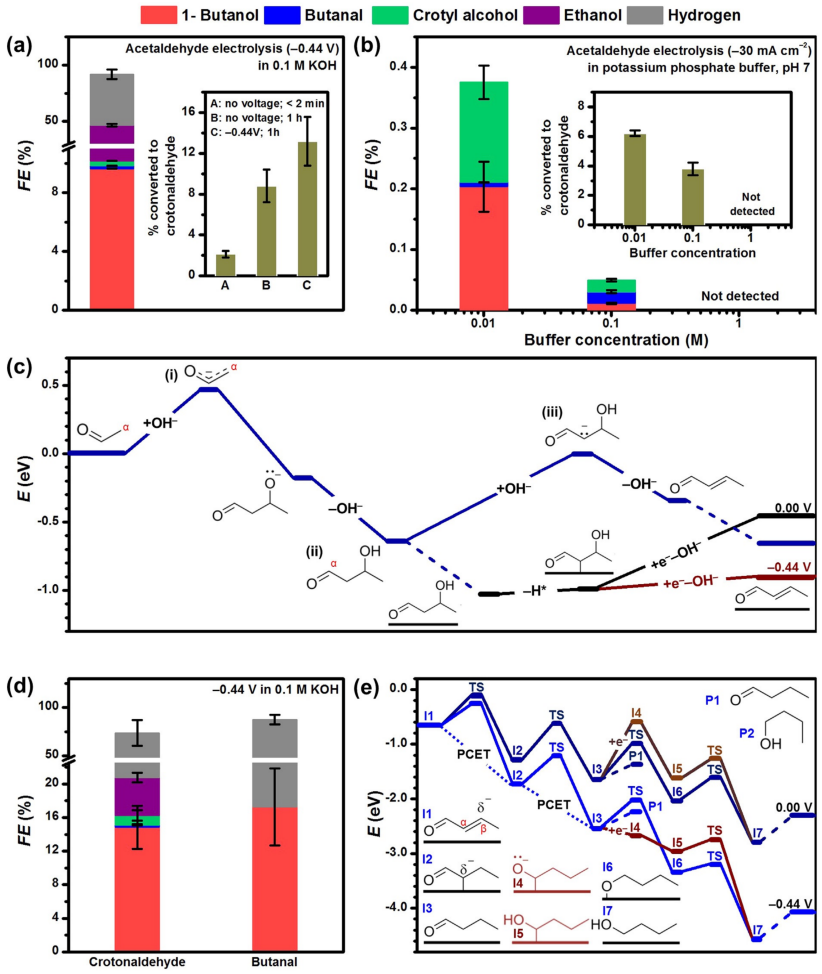


Figure 2. Faradaic efficiencies of products from 1 h electrolysis of 50 mM acetaldehyde on CuO-derived Cu (a) in alkaline and (b) neutral buffer (62 mol% K_2HPO_4 + 38 mol% KH_2PO_4) electrolyte. The insets of (a) and (b) show the percentage of acetaldehyde that was converted to crotonaldehyde in the electrolyte in each case. c) Potential energy diagram of acetaldehyde condensation to crotonaldehyde in solution (dark blue) and mediated by the surface (black). The final OH^- removal, which can be assisted by one electron donated from the surface, is promoted by reductive potentials (dark red). Water molecules were omitted for clarity. d) Faradaic efficiencies of products from alkaline electrolyses of 50 mM crotonaldehyde and 50 mM butanal on CuO-derived Cu. e) Potential energy diagram of crotonaldehyde reduction to 1-butanol via butanal (blue). Under negative potentials, butanal can be adsorbed via a one-electron transfer (dark red), which promotes its further reaction instead of desorption. Dashed lines represent adsorptions/desorptions. Dotted lines represent proton-coupled electron transfers (PCET). Additional (electro)chemical routes are shown in Sections S3 and S6.

sistent with findings from the literature,[19] the latter is the rate-determining step in solution. We further note that for the case of CO_2 reduction, adsorbed ethenyloxy species is formed as a precursor of acetaldehyde and ethanol,[16,20] and thus can also readily react with an acetaldehyde molecule in solution and be subsequently hydrogenated to form 3-hydroxybutanal (Section S3).

The amount of acetaldehyde converted to crotonaldehyde during electrolysis (13.2%) is larger than the 8.8% conversion (or 2.2 mM crotonaldehyde) when 50 mM acetaldehyde was aged in 0.1M KOH for 1 h without applied potential (inset in Figure 2a). This observation suggests that the Cu surface can promote the aldol condensation at cathodic potentials. DFT analysis reveals that this alternative pathway starts with 3-hydroxybutanal adsorbing exothermically on Cu and losing an α-hydrogen as an adsorbed H (dark red in Figure 2c). The hydroxyl group is then eliminated, in parallel with an electron transfer, to give crotonaldehyde. On the Cu surface, this step is promoted by negative applied potentials. Overall, the DFT investigation reveals that the alkaline electrolyte promotes the initial C–C coupling step between two acetaldehyde molecules, while the Cu surface promotes the subsequent dehydration step to crotonaldehyde.

To elucidate the fate of C_4 species after the aldol condensation step, we electrolyzed 50 mM crotonaldehyde on CuO-derived Cu in 0.1M KOH at −0.44 V (Figure 2d, Section S5). The major carbonaceous product was 1-butanol ($FE = 14.8\%$), while small amounts of butanal ($FE = 0.3\%$) and crotyl alcohol ($FE = 1.1\%$) were also detected. The Faradaic selectivity of 1-butanol ($FE_{1\text{-Butanol}}$ normalized by the FE of all the C_4 products) from crotonaldehyde electrolysis was 91.4%, which is similar to the case of acetaldehyde (94.7%, Table S10). This reinforces the role of crotonaldehyde as the main intermediate in the electrosynthesis of acetaldehyde to C_4 oxygenates. The presence of ethanol ($FE = 4.5\%$) was attributed to the reduction of acetaldehyde present due to the hydroxide-catalyzed retro-aldol reaction of crotonaldehyde, which is known to occur at room temperature.[21] This observation highlights the complexity of crotonaldehyde chemistry under aqueous alkaline conditions, and leads us to infer that the low total Faradaic efficiency of 72.2% is a consequence of undetected products from other side reactions of crotonaldehyde in the alkaline electrolyte (Figure S10).

Butanal and crotyl alcohol are known intermediates in the gas-phase hydrogenation of crotonaldehyde to 1-butanol,[22] and their presence during crotonaldehyde electrolysis suggests that they are potential electrochemically-active intermediates to 1-butanol. Therefore, we electrolyzed butanal and crotyl alcohol under the same conditions (Figure 2d, Table S9). 1-Butanol was the sole product from the electroreduction of butanal ($FE = 17.3\%$; the balance product is H_2). Only hydrogen was detected during the electrolysis of crotyl alcohol, which indicates that the latter is electrochemically inert, in good agreement with theoretical calculations in Figure S11.

Theoretical analysis of crotonaldehyde reduction reveals that butanal is formed by sequential hydrogenation of the β- and α-carbons of crotonaldehyde (Figure 2e). Once formed, butanal tends to desorb rather than further react. However, at potentials more reductive than −1.02 V vs. SHE (standard hydrogen electrode), butanal receives an electron from the cathode surface to form the $CH_3CH_2CH_2C^*HO^-$ anion (I4 in Figure 2e). As this adsorption does not involve proton transfers, it is independent of the electrolyte pH in the SHE scale. $CH_3CH_2CH_2C^*HO^-$ is subsequently protonated to yield 1-hydroxybutyl (I5 in Figure 2e), which is further hydrogenated in a chemical step ($E_a = 0.39$ eV) to produce 1-butanol. These theoretical findings are corroborated by our results from crotonaldehyde electrolysis performed at pH 7 and pH 13 (Table S9). At −1.20 V vs. SHE, 1-butanol was the most selective product in both electrolytes, consistent with the pH-independent adsorption of butanal. However, at −0.90 V vs. SHE, butanal was the most selective product, indicating that this potential was insufficient for its further reduction to 1-butanol. Alternative chemical routes from butanal to 1-butanol are shown in Figure S12.

Aldehydes can be hydrated to geminal diols in aqueous alkaline solution. Signals belonging to hydrated crotonaldehyde ($CH_3CH{=}CHCH(OH)_2$) were observed in nuclear magnetic resonance (NMR) spectroscopic analyses of 50 mM crotonaldehyde or acetaldehyde dissolved in 0.1M KOH (Figure 3a, Figure S13). We therefore considered the possibility of hydrated crotonaldehyde as an intermediate to 1-butanol. DFT suggests that hydrated crotonaldehyde cannot be further electrochemically reduced due to a larger activation barrier (+0.41 eV) compared to its desorption energy (+0.26 eV), as shown in Figure 3b. This result was corroborated by the $FE_{1\text{-Butanol}}$ of 46-47% observed from the crotonaldehyde and butanal electrolyses in 0.1M potassium phosphate buffer (pH 7) at −0.79 V (Figure 3c, Table S9), which was almost three times more than the values from electrolyses in 0.1M KOH. Figure 3a shows that in a neutral medium, only peaks corresponding to crotonaldehyde were observed. The neutral buffer electrolyte therefore likely suppressed the hydration process and increased the availability of unhydrated crotonaldehyde for reduction.

To find a better electrocatalyst to enhance the FE of 1-butanol from acetaldehyde and crotonaldehyde reduction, we proceeded to identify an activity descriptor. To this end, we electrolyzed acetaldehyde in 0.1M KOH and crotonaldehyde in 0.1M potassium phosphate buffer on different metal discs (Section S8). Since we have demonstrated that both the formation of C_4 oxygenates via the aldol condensation pathway, and the reactivity of crotonaldehyde are affected by the (local) pH, we employed a constant-current electrolysis at $-10\ mA\,cm^{-2}$ to identify the different reactivities of the metals. For acetaldehyde electrolysis, we found that the selectivity of Cu, Fe, Co, Ni, Ag, and Au towards 1-butanol (Figure 4a) and all C_4 products (Figure S14a) can be correlated to the cathode metal-oxygen bond strength, with Fe showing the highest selectivity towards 1-butanol ($FE = 4.0\%$). A similar trend was observed for the six above-mentioned metals and also Pd and Pt for crotonaldehyde reduction to 1-butanol (Figure 4b), with Fe also showing the highest FE of 26.3%. The linear-scaling relationships end sharply in a selectivity cliff.[23] The origin of the discontinuity is likely due to a phase transformation. According to their

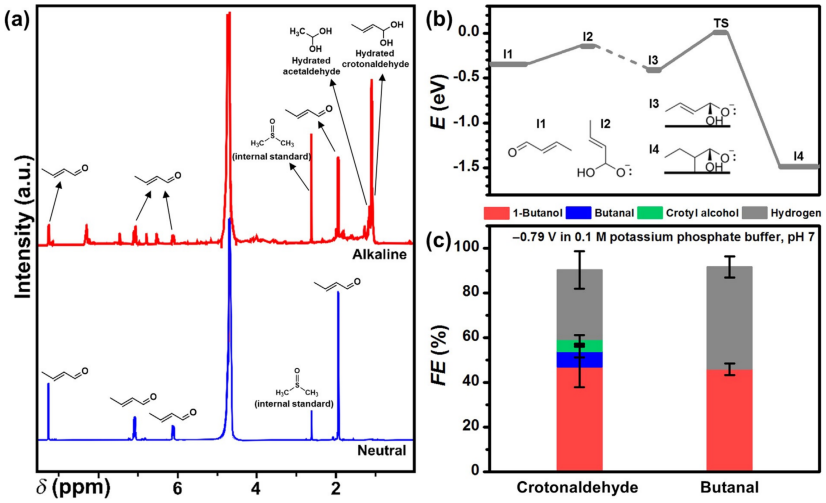


Figure 3. a) ^{1}H NMR spectra of 50 mM crotonaldehyde (δ refers to the chemical shift) in alkaline environment (0.1 M KOH, red line) and neutral (ultrapure deionized water, blue line). Phenol (7.2 ppm) and dimethylsulfoxide (2.6 ppm) were dissolved in D_2O (residual H_2O peak at 4.8 ppm) and added as internal standards. b) Potential energy diagram of crotonaldehyde hydration (I1 to I2) and subsequent hydrogenation. Dashed lines represent adsorption/desorption. c) Faradaic efficiencies of crotonaldehyde and butanal electrolysis products on CuO-derived Cu in neutral buffer electrolyte.

Pourbaix diagrams,[24] Zn, Ti, Cr, and Mo may have surface oxide layers under the working cathodic potentials and thus have poor yields towards 1-butanol and other C_4 products (Section S9). Incidentally, these metals were more selective for reducing acetaldehyde to crotyl alcohol (Table S11), and crotonaldehyde to butanal (Table S12), probably because the dominating linear-scaling relationships differ from those of the pure metals. Based on these results, we put forward that the C_4 product selectivity is influenced by the affinity of the catalyst to oxygen.

Collectively, our results gave the reasons for the low *FE* of 1-butanol during the electroreduction of CO_2 on oxide-derived copper. The first factor is the low activity for acetaldehyde production on copper materials ($FE_{CH_3CHO} \leq 0.1\,\%$, $j_{CH_3CHO} \leq -0.41\,mA\,cm^{-2}$ in Table S1, $FE_{CH_3CHO} < 2.1\,\%$, $j_{CH_3CHO} < 1\,mA\,cm^{-2}$ in the literature[4,25]). The second factor is that the conversion of acetaldehyde to ethanol on copper is kinetically facile and strongly competing ($FE_{ethanol} = 36.5\,\%$, while $FE_{1\text{-Butanol}} = 9.6\,\%$ at -0.44 V vs. RHE, Figure 2a).[17,26] The third factor is that the formation of the C_4 backbone and its subsequent reduction to 1-butanol are promoted by conflicting experimental conditions. While an alkaline environment facilitates aldol condensation between acetaldehyde molecules to crotonaldehyde, its hydration to an unreactive form is also promoted.

Conclusion

In summary, we detected for the first time C_4 oxygenates from CO_2 electroreduction on CuO-derived Cu, with 1-butanol being the most favored product amongst them ($FE = 0.056\,\%$, $j = -0.080\,mA\,cm^{-2}$ at -0.48 V). The quantification was made possible by a combination of the high-rate electrolysis, achieved by using a GDE in a flow cell, and sensitive analytical techniques. We ruled out the formation of 1-butanol from the C–C coupling of four individual C_1 adsorbates, such as *CO. Instead, the combination of experimental and theoretical studies established a rich reaction mechanism that combines chemical and electrochemical steps, where the electrocatalytically generated acetaldehyde plays a prominent role. Its base-catalyzed aldol condensation, promoted by high local pH and the catalyst surface, produces crotonaldehyde, which is subsequently electroreduced to 1-butanol.

This study further highlights the challenges associated with the one-pot approach to converting a low molecular weight feedstock like CO_2 into complex functionalized molecules. We discover that contrasting catalysts and conditions are required to maximize the yield of each step. In addition to operating under highly alkaline conditions, we also note that a single electrocatalytic surface can hardly optimize all the required steps. For example, among the metal discs tested, Fe was identified as the most selective catalyst for acetaldehyde reduction to 1-butanol, but it is not active per se

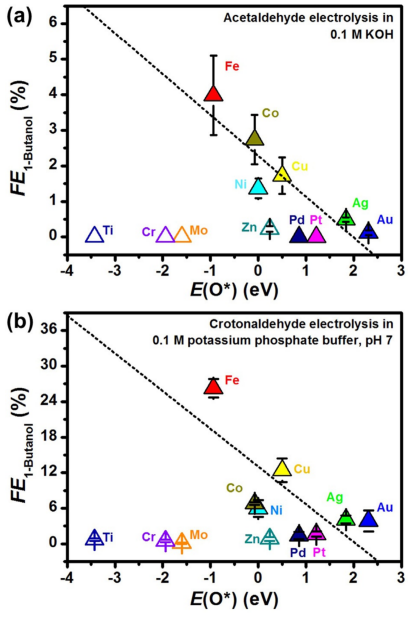


Figure 4. Faradaic efficiencies of 1-butanol from −10 mA cm^{-2} constant-current electrolysis of a) acetaldehyde and b) crotonaldehyde on selected metals as a function of the DFT-computed adsorbed oxygen stability on these metals with respect to water and hydrogen. Metals that are typically oxides at 0 V vs. RHE at the pH of the supporting electrolyte are shown as hollow symbols. 1-Butanol was not detected from acetaldehyde electrolysis on Ti, Cr, Mo, Pd, and Pt.

for CO_2 reduction. Therefore, designing a one-pot reactor for the electrosynthesis of large molecules would inevitably be associated with low performance. Instead, we propose that a more viable synthetic strategy would be to deconvolute the multi-step process into sequential operation units. Therein, chemical or electrochemical reactions with different process conditions could be independently optimized. The separate stepwise-optimized reactors could then be placed in tandem for the efficient conversion of each intermediate, leading to increased yield of the desired product.

Acknowledgements

We thank the National University of Singapore Flagship Green Energy Program (R143-000-A64-114, R143-000-A55-733 and R143-000-A55-646), Ministry of Education, Singapore (R143-000-683-112), Spanish Ministry of Science and Innovation RTI2018-101394-B-I00 project for financial support and the European Union through the A-LEAF project (732840-A-LEAF). The Barcelona Supercomputing Center—MareNostrum (BSC-RES) is acknowledged for providing generous computer resources. The authors are grateful to Dr. R. Hauert and S. Büchele for XPS measurements and Dr. R. Verel for NMR assistance.

Conflict of interest

The authors declare no conflict of interest.

Stichwörter: 1-butanol · carbon dioxide reduction · density functional theory · electrochemistry · reaction mechanisms

[1] O. S. Bushuyev, P. De Luna, C. T. Dinh, L. Tao, G. Saur, J. van de Lagemaat, S. O. Kelley, E. H. Sargent, *Joule* **2018**, *2*, 825 – 832.
[2] D. Ren, Y. Deng, A. D. Handoko, C. S. Chen, S. Malkhandi, B. S. Yeo, *ACS Catal.* **2015**, *5*, 2814 – 2821.
[3] C.-T. Dinh, T. Burdyny, M. G. Kibria, A. Seifitokaldani, C. M. Gabardo, F. P. G. de Arquer, A. Kiani, J. P. Edwards, P. De Luna, O. S. Bushuyev, *Science* **2018**, *360*, 783 – 787.
[4] K. P. Kuhl, E. R. Cave, D. N. Abram, T. F. Jaramillo, *Energy Environ. Sci.* **2012**, *5*, 7050 – 7059.
[5] M. Rahaman, A. Dutta, A. Zanetti, P. Broekmann, *ACS Catal.* **2017**, *7*, 7946 – 7956.
[6] D. Ren, N. T. Wong, A. D. Handoko, Y. Huang, B. S. Yeo, *J. Phys. Chem. Lett.* **2016**, *7*, 20 – 24.
[7] S. Lee, D. Kim, J. Lee, *Angew. Chem. Int. Ed.* **2015**, *54*, 14701 – 14705; *Angew. Chem.* **2015**, *127*, 14914 – 14918.
[8] W. Roberto da Silva Trindade, R. Gonçalves dos Santos, *Renewable Sustainable Energy Rev.* **2017**, *69*, 642 – 651.
[9] T. Haas, R. Krause, R. Weber, M. Demler, G. Schmid, *Nat. Catal.* **2018**, *1*, 32 – 39.
[10] X. Chang, A. Malkani, X. Yang, B. Xu, *J. Am. Chem. Soc.* **2020**, *142*, 2975 – 2983.
[11] D. Ren, J. Fong, B. S. Yeo, *Nat. Commun.* **2018**, *9*, 925.
[12] R. Kas, R. Kortlever, H. Yılmaz, M. T. M. Koper, G. Mul, *ChemElectroChem* **2015**, *2*, 354 – 358.
[13] R. Kortlever, I. Peters, C. Balemans, R. Kas, Y. Kwon, G. Mul, M. T. M. Koper, *Chem. Commun.* **2016**, *52*, 10229 – 10232.
[14] K. J. P. Schouten, Y. Kwon, C. J. M. van der Ham, Z. Qin, M. T. M. Koper, *Chem. Sci.* **2011**, *2*, 1902 – 1909.
[15] S. Nitopi, E. Bertheussen, S. B. Scott, X. Liu, A. K. Engstfeld, S. Horch, B. Seger, I. E. L. Stephens, K. Chan, C. Hahn, J. K. Nørskov, T. F. Jaramillo, I. Chorkendorff, *Chem. Rev.* **2019**, *119*, 7610 – 7672.
[16] A. J. Garza, A. T. Bell, M. Head-Gordon, *ACS Catal.* **2018**, *8*, 1490 – 1499.
[17] E. Bertheussen, A. Verdaguer-Casadevall, D. Ravasio, J. H. Montoya, D. B. Trimarco, C. Roy, S. Meier, J. Wendland, J. K. Nørskov, I. E. L. Stephens, I. Chorkendorff, *Angew. Chem. Int. Ed.* **2016**, *55*, 1450 – 1454; *Angew. Chem.* **2016**, *128*, 1472 – 1476.
[18] J. Li, F. Che, Y. Pang, C. Zou, J. Y. Howe, T. Burdyny, J. P. Edwards, Y. Wang, F. Li, Z. Wang, P. De Luna, C.-T. Dinh, T.-T. Zhuang, M. I. Saidaminov, S. Cheng, T. Wu, Y. Z. Finfrock, L. Ma, S.-H. Hsieh, Y.-S. Liu, G. A. Botton, W.-F. Pong, X. Du, J. Guo, T.-K. Sham, E. H. Sargent, D. Sinton, *Nat. Commun.* **2018**, *9*, 4614.
[19] J. P. Guthrie, *J. Am. Chem. Soc.* **1991**, *113*, 7249 – 7255.
[20] R. Kortlever, J. Shen, K. J. P. Schouten, F. Calle-Vallejo, M. T. M. Koper, *J. Phys. Chem. Lett.* **2015**, *6*, 4073 – 4082.
[21] J. P. Guthrie, *Can. J. Chem.* **1974**, *52*, 2037 – 2040.
[22] M. Englisch, A. Jentys, J. A. Lercher, *J. Catal.* **1997**, *166*, 25 – 35.
[23] J. Pérez-Ramírez, N. López, *Nat. Catal.* **2019**, *2*, 971 – 976.

[24] M. Pourbaix, *Atlas of Electrochemical Equilibria in Aqueous Solutions*, 2nd ed., National Association of Corrosion Engineers, Houston, **1974**.
[25] Y. Hori, I. Takahashi, O. Koga, N. Hoshi, *J. Mol. Catal. A* **2003**, *199*, 39 – 47.
[26] Y. Hori, R. Takahashi, Y. Yoshinami, A. Murata, *J. Phys. Chem. B* **1997**, *101*, 7075 – 7081.

Generalizing Performance Equations in Heterogeneous Catalysis from Hybrid Data and Statistical Learning

Sergio Pablo-García,[a,‡] Albert Sabadell-Rendón,[a,‡] Ali J. Saadun,[b] Santiago Morandi,[a] Javier Pérez-Ramírez,[b,*] Núria López[a,*]

[a] Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, Av. Països Catalans 16, 43007, Tarragona, Spain.

[b] Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.

*Corresponding authors. Email: jpr@chem.ethz.ch; nlopez@iciq.es

‡These authors have contributed equally to the present work

Abstract

Activity equations trying to mimic experimental catalytic performance derived from reaction profiles and microkinetic models have been the state-of-the-art in modeling in the last decades. This approach has been able to reproduce semi-quantitatively activity volcano plots leading to successful catalyst optimization through the use of descriptors. As systems become more complex (both catalysts and reactants) these methods face increasing limitations. Statistical Learning (SL) techniques can overcome these limitations and improve the search for descriptor-based performance equations. However, the black-box nature of SL techniques difficults the physical interpretation of the so-obtained models. To advance in the integration of these methodologies to real problems we have merged experimental activity and selectivity presented as a function of chemical descriptors from Density Functional Theory for a particularly complex family of hydrodehalogenation reactions CH_2X_2 (for X=Br, Cl) presenting three main products. The employed Bayesian procedure is able to identify robust equations for activity and selectivity as a function of only two descriptors. This work presents a starting point to solve complex reaction networks using a set of statistical learning tools and hybrid data.

Introduction

Equations describing catalytic performance are at the core of industrial heterogeneous catalysis.[1] Stock and Bodenstein found the first kinetic power law rate by investigating reaction rates at variable pressure.[1] This heuristic approach was soon complemented by the mechanistic insight developed by Langmuir, suggesting the existence of equivalent non-interacting active sites on the surface and which evolved in the so-called Langmuir-Hinshelwood-Hougen-Watson (LHHW) kinetics. In this approach, a reaction was described by a set of elementary steps, the rate of which was derived from a LHHW equation. The LHHW rate predictions were compared to the experimentally obtained kinetics data and, if the LHHW model was able to fit that data the mechanism was meaningful. As such, the equation was deemed useful to design a reactor for the application at the specified operating conditions, yet the model is too ideal. Alternatives considering anisotropy were introduced by Temkin,[2] with limited success.

Microkinetic modelling (MK) emerged in the second half of 20th century. [3,4] In the MK procedure, a mechanism was proposed, and each step had an associated kinetic coefficient *k,* defining a system of Ordinary Differential Equations (ODE) with initial conditions. The one-to-one mapping to experiments provided estimates for the kinetic coefficients and equilibrium constants (*K*) for the steps in the mechanism and the time evolution of intermediates either as concentration or coverage. Several alternative mechanisms could be tested following this procedure and the quality of the fitting and the reliability of the *K* and *k* parameters served to retain or discard a proposed mechanism. The emergence of massive atomistic simulations based on Density Functional Theory (DFT) allowed to investigate reaction profiles in model systems (particularly metals), and changed the traditional use of MK procedures. DFT finds the minimum energy configuration for intermediates on surfaces and the transition states linking the elementary steps. Thus the corresponding reaction energies, ΔE and activation barriers, E_a could be employed coupled to Statistical Thermodynamics to obtain estimates for the kinetic coefficients of each

elementary step[5] through the Arrhenius and Eyring equations and even for adsorption/desorption processes, where pressure (Hertz-Knudsen) is the relevant variable. Microkinetic models assumptions imply that the number of sites is preserved all through the reaction and the reactivity in all these centers is identical. However, as MK is a mean-field approach, it fails to fully describe highly anisotropic systems in which directionality is crucial. Consequently, even higher resolution methods, as kinetic Monte Carlo (kMC) are required instead. In the kMC approach, all possible steps are simulated on a surface (lattice) explicitly using a stochastic procedure, in which the probability to occur of each process is estimated according to the corresponding *k*.[6,7] Of course, the more elementary steps are introduced in the kMC simulations, the larger the number of explicit DFT evaluations is needed and thus, in non-obvious cases kMC becomes impractical although recent efforts in code parallelization could improve the predictions.[8] MK and kMC are used to predict macroscopic parameters such as Turn Over Frequency (TOF),[9–18] selectivity,[9,10,16–20] apparent activation energies,[18,19,21–26] and reaction orders.[10,12,19,23,26,27] Additionally, from the MK and kMC results analysis of the most relevant steps can be performed through the Degree of Rate Control (DRC),[14,19,28–31] and the Degree of Selectivity Control (DSC). From the DRC and DSR simplified model equations, known as surrogate models, can be inferred employing exponential dependences to key steps and then expanding the first term of the Taylor expansion of Arrhenius equation to make them linear.[32] In practice MK and kMC require intensive path investigation to identify the mechanism and some fine tuning of some of the DFT computed parameters in order to be directly mapped to experimental results for complex systems.[33–36] The consequence is that in reactor development power-law equations have been the functions of choice to represent experimental rates. Under this approach, similar functional forms suggest identical mechanisms,[37] whereas the reaction orders are often linked to the participation of a given species in the reaction network, particularly in before the rate-determining step. These equations have been employed in the chemical industry to find suitable operation conditions for over a century.

Catalyst development has been linked to the Sabatier principle[38,39] the "not-too-strong-not-too-weak" rule for the optimization within a family of catalytic materials. This rule was formulated deriving in the volcano-shaped functions of a single energy value acting as a catalytic activity descriptor that populate the research works in heterogeneous catalysis. Actually, this generalization of the catalytic activity within a family of catalysts or a family of reactants is rooted on the phenomenological observations by Hammett[40] and Hammond postulates,[41] and Brønsted-Evans-Polanyi (BEP)[42,43] equations. Again, atomistic simulations based on DFT unraveled the nature of these dependencies through the so-called Linear-Scaling Relationships (LSR), linking thermodynamics to thermodynamics and thermodynamics to kinetics.[23,36,44–51] In summary, on a metal surface (the most commonly investigated catalyst family) the binding energy of a simple intermediate can be traced back to another with the same heteroatom. The transition state for an elementary step can be linked to initial and final intermediates in a similar manner. Thus, the reaction energies and activation barriers can be estimated only by knowing the slopes for these linear dependencies and the energy parameter of the descriptor.[44,52,53] The dependences can be directly plugged as equations in the MK or kMC modeling, reducing the dimensionality of the catalytic performance problem and leading to computed activity volcanos. In summary, activity volcano plots are obtained from *ab initio* principles with Density Functional Theory (DFT)[37,52] data of a single catalysts, from which the full reaction network is computed. Typically, the energies of the intermediates and transition states are condensed through LSR and the transition state theory to provide the kinetic coefficients, after which they are embedded in the microkinetic models.

This approach has been the gold standard in heterogeneous catalysis[37,54] since the beginning of the 21st century. Catalytic phase changes, dynamics and deposits of poisons are limiting the progress in the field as they are not rightly accounted for in the microkinetic models. Thus, accumulated uncertainties make predictions less accurate when increasing in complexity,[55] which is mainly caused by: (*i*) coverage effects with concomitant adsorbate reorientation or adsorbate

energy changes, (*ii*) catalyst phase or surface transformations; (*iii*) large reaction networks, where elementary steps grow exponentially;[23] and (*iv*) highly dynamic materials that have various possible configurations.[56] Consequently, descriptors are chosen based on, to some extent, arbitrary heuristics,[57–59] and the computed MK(DFT) (or kMC(DFT)) rates are still far from experimental values and fail in describing key experimental observables such as selectivity.[59] Recent studies show that the energy profiles require corrections as large as 0.5 eV with respect to the computed DFT values for some intermediates (the asymmetry leads to further difficulties in assigning meanings to these corrections) to explain experimental observations, particularly in hydrodechlorination processes.[60,61]

In summary, the inference of robust and accurate mathematical expressions to predict activity and product selectivity and finding generally applicable rates within a given family of compounds, are far from obvious. Pioneering approaches employing hybrid data (i.e. from experiments and DFT) have been applied to Hydrogen Evolution Reaction (HER) on transition metal surfaces,[62] or the methanation reaction on alloys,[63] thus rather simple model catalysts. In these cases, LSRs were employed to correlate experimental activity to descriptors based on physical intuition rather than in statistics and are accordingly difficult to generalize.[64]As large experimental and computational datasets[65] are being made available through high throughput techniques, the systematically generated data is amenable for Statistical Learning (SL) treatments.[64,66–74] In general, the introduction of data approaches has been steered from the DFT community. Thus, approaches combining DFT and SL have been recently introduced in: (*i*) derivation of approaches to simplify the DFT computational burden of calculating the energies of many configurations;[75–78] (*ii*) applying these energies to traditionally (MK(DFT)) derived volcanoes providing candidates for electrochemical processes and (*iii*) searching for optimized descriptors employed in theoretical studies.[78,79] In parallel, attempts to link experimental catalytic performance can be found in recent literature. As for example the work of Foppa et al., in which around 40 tabulated experimental

observables were used to predict the annotated consistent conversion and selectivity of nine vanadium-based catalysts using a symbolic regression protocol (SISSO) to derive a non-trivial ensemble of models.[80]

Among the most relevant procedure for catalytic optimization descriptors identification, equivalent to dimensionality reduction in SL, has been the most investigated *via* t-distributed Stochastic Neighbor Embedding (t-SNE),[81] Least Absolute Shrinkage and Selection Operator (LASSO) [82] and Principal Component Analysis (PCA).[79,83] However, many of these methods are hard to interpret and the optimization parameters are difficult to map to macroscopic variables. Bayesian techniques hold the key to solving many of the interpretability and optimization issues of experimental variables and uncertainties quantification.[84–86] Attempts to generate modeling equations for spectroscopic, spectrometric and chromatographic,[87] using experimental-only datasets has been performed using Bayesian Statistics. In electrochemical modeling, Bayesian techniques and Gaussian regressor have been employed to predict best compositions for the Oxygen Reduction Reaction in binary and ternary alloys from hybrid data.[88] Functional exploration has been recently been proposed for some properties linking to material degradation.[72,89] In our case the search for symbolic equations to build parsimonious models we have employed the Bayesian Machine Scientist (BMS)[90] that allows functional exploration using Markov Chain Monte Carlo.

Product selectivity in homogeneous catalysis has been analyzed through various SL techniques, where in some studies computed parameters were taken as variables,[83,91–93] even with small set of experimental data.[94] This type of heuristics with hybrid data has not been attempted in heterogeneous catalysis, which have a “space of ligands”, defined as set of atoms or molecules able to be attached to the surrounding of the active metal centers, less continuous than homogenous systems. Because of this, the experimental data is sparser, observations are fewer, and the characterization needs to account for larger space and time scales. In our first attempt to

leverage hybrid data approaches to heterogeneous catalysis, we reproduced the conversion of CH_2Br_2 in metal-catalyzed (Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir and Pt) hydrodebromination using SL techniques, and demonstrated the stability of the employed algorithms[95]. The hydrodehalogenation reactions constitute a benchmark for SL techniques as the classical MK(DFT) modeling fails due to phase and surface changes (see below), as carbides form for Co (halides for Cu or Ag). In addition, the experimental observation of three different products, some of them ill-defined composition as coke. Therefore, setting a robust framework for the use of SL techniques on a technologically challenging reaction for a family of compounds can help us setting the key questions, like equation transferability, generalization to a family of reactants, and robustness of the functional forms.

In order to develop a robust framework for the deployment of SL with the aim to obtain reactivity equations of heterogeneous catalysts, we have used the technologically relevant CH_2X_2 (X=Br, Cl) hydrodehalogenation reaction and compared them to MK(DFT) models. This transformation is a key step in halogen-mediated methane upgrading processes and clearly displays the selectivity issues of the metal catalysts.[95–98] By combining the descriptor identification from PCA and BMS, we present a unique equation search for the reaction rate, CH_2X_2 conversion, and selectivity to different products, thus compiling the performance of the metal catalyst in a generalized form (procedure depicted in **Figure 1**). Our approach links experimental and theoretical data and sets a robust methodology that could be extrapolated to other heterogeneously catalyzed reactions.

Materials and Methods

Density Functional Theory and microkinetic model setup

Density Functional Theory (DFT) implemented in the Vienna Ab Initio Simulation Package (VASP 5.4.4)[99] was employed to describe the chemical systems involved in this work. The exchange-correlation energies were obtained using the Generalized Gradient Approximation with the Perdew-Burke-Ernzerhof (GGA PBE-D2) functional,[100,101] reparametrizing the C_6 values for the metals.[102] The inner electrons were represented using the Projected-augmented wave (PAW) and the valence electrons by plane waves with a cutoff energy of 450 eV.[103] The Monkhorst-Pack method was used to create the Γ-centered k-point mesh.[104] A four-layer p(3x3)-(111) face centered cubic slab (fcc) was used to model the metallic surfaces, except for Co and Ru, for which a p(3x3)-(0001) body centered cubic slab (hcp) was used. A box of 15×15×15 $Å^3$ was used to model the molecules in gas phase. The vacuum of the slab was implemented with a space in the cell of 15 Å in the z-direction, including the dipole correction due to the asymmetry of the system.[105] For the electronic and ionic relaxations, the threshold criteria were set as 10^{-5} eV and 0.03 eV $Å^{-1}$, respectively. During optimization, the two upper metal layers and the adsorbates were allowed to relax, while the rest of the atoms were fixed. The stability of the Cl, F and I halogenated intermediates was checked via a frequency analysis, while Br and non-halogenated structures and energies were retrieved from ioChem-BD (http://dx.doi.org/10.19061/iochem-bd-1-150 and http://dx.doi.org/10.19061/iochem-bd-1-152).[65,95] Further technical details on DFT simulations can be found in the **Supporting Information** (**SI**), **Note S1** and **Equations S1-2**.

In the microkinetic model (mechanism in **Figure S1** and **Table S1**), first we assessed the hermodynamic consistency (**Table S4**) and analyzed the reversibility for all metals (**Table S5**). The results of the thermodynamic consistency analysis are in a good agreement with those reported from the third millennium database.[106] The microkinetic runs for all metals were performed using a differential reactor model at the experimental reaction conditions: T=523 K, P=1 atm (70% inert gases), initial CH_2Br_2 : H_2 ratio = 1 : 4. Further details on the microkinetic setting can be found on **Note S2** on the **SI**, **Equations S3-7**, and **Figures S2-7**.

Catalyst Preparation

SiO_2-supported metal catalysts were synthesized following the protocol reported by Saadun *et al.*[95] Commercially available SiO_2 (Evonik, AEROPERL 300/30, S_{BET} = 257 m^2 g^{-1}, V_{pore} = 0.95 cm^3 g^{-1}, > 99.0%) was calcined at 973 K for 5 h in static air (heating rate 5 K min^{-1}) prior to its use as support in the catalyst preparation (see **Note S3** for further information). The resulting solids were dried at 373 K for 12 h and optionally calcined in static air at 623 K (heating rate 5 K min^{-1}) to obtain the SiO_2-supported metal oxides. Subsequently, all samples underwent a reductive treatment in 20 vol% H_2/He (PanGas, purity 5.0) flow for 3 h (heating rate of 10 K min^{-1}) at elevated temperatures (573-968 K) prior to their use in catalytic tests. The catalysts were referred to as M/SiO_2, where *M* denotes the metal (*i.e.* Co, Ni, Cu, Ru, Rh, Ag, Ir, or Pt).

Catalyst Testing

The hydrodehalogenation of the dihalomethanes (CH_2X_2, X=Cl, Br) was performed at ambient pressure in a home-made continuous-flow fixed-bed reactor set up. H_2 (PanGas, purity 5.0), He (Carrier gas, PanGas, purity 5.0), Ar (internal standard, PanGas, purity 5.0) were supplied by a set of digital mass flow controllers (Bronkhorst) and liquid CH_2Br_2 (Acros Organics, 99%) or CH_2Cl_2 (Sigma-Aldrich, >99.9%) was dosed by a syringe pump (Fusion 100, Chemyx) equipped with a water-cooled syringe to a vaporizer unit operated at 393 K. The quartz reactor (internal diameter, d_i = 12 mm) containing the reduced catalyst (catalyst weight, W_{cat} = 0.1-1 g, particle size, d_p = 0.4-0.6 mm) was heated to the desired temperature (T = 423-623 K) in an electrical oven under He flow. The catalyst bed was allowed to stabilize for at least 10 min at desired temperature before the reaction mixture was fed at a total volumetric flow (F_T) of 20-350 cm^3 STP min^{-1} and desired feed composition of CH_2X_2:H_2:Ar:He = 6:24:4.5:65.5 (vol%, X=Cl, Br). Downstream linings were heated at 393 K to prevent the condensation of unconverted reactants and/or products. The content of carbon containing compounds (CH_2X_2, CH_3X, CH_4,

C_2H_4, C_2H_6, C_3H_6, and C_3H_8) and of Ar in the reactor outlet gas stream was quantified online via a gas chromatograph equipped with a GS Carbon PLOT column coupled to a mass spectrometer (GC MS, Agilent GC 6890, Agilent MSD 5973N). After the GC MS analysis, the gas stream was passed through two impinging bottles in series containing an aqueous solution of NaOH (1 M) for neutralization prior to its release in the ventilation system.

The conversion of the dihalomethane, $X_{CH_2X_2}$, was calculated using **Equation (1)**,

$$X_{CH_2X_2} = \frac{n(CH_2X_2)_{in} - n(CH_2X_2)_{out}}{n(CH_2X_2)_{in}} \cdot 100, \% \quad \textbf{(1)}$$

where $n(CH_2X_2)_{in}$ and $n(CH_2X_2)_{out}$ are the molar flows of the reactant at the reactor inlet and outlet, respectively. The selectivity, S_i, to product *I* (*i*: CH_3X, CH_4, C_2H_4, C_2H_6, C_3H_6, and C_3H_8) was calculated according to **Equation (2)**,

$$S_i = \frac{\sigma \cdot n(J)_{out}}{n(CH_2X_2)_{in} - n(CH_2X_2)_{out}} \cdot 100\,, \% \quad \textbf{(2)}$$

where $n(j)_{out}$ is the molar flow of product *j* at the reactor outlet. To take the number of carbon atoms in the products into account, σ equals 1, 2, and 3, for C_1, C_2, and C_3 products, respectively. The error of the carbon balance, ε_C, used to specify the selectivity to coke, was determined using **Equation (3)**,

$$\varepsilon_c = \frac{n(CH_2X_2)_{in} - n(CH_2X_2)_{out} - \sigma \cdot n(J)_{out}}{n(CH_2X_2)_{in}} \cdot 100, \% \quad \textbf{(3)}$$

Evaluation of the dimensionless moduli based on the criteria of Carberry, Mears, and Weisz-Prater[107,108] indicated that the catalytic tests were performed in the absence of mass and heat transfer limitations.

Statistical Analysis details

The Random Forest Regressor (RF) and Gaussian Process Regressor (GR) were applied using the Scikit-Learn (0.23.1) package.[109] Principal Component Analysis (PCA) and Bayesian Machine Scientist (BMS) were implemented according to our previous work.[79,90,95] The RF was applied with one-hundred-and-fifty estimator trees with a maximum depth of 3 and optimized using the Mean

Squared Error and by bootstrapping the samples. For the GR, the Radial-Basis Function kernel was chosen with a noise parameter (α) of 10^{-5} (see **Note S4**, **Equations S8** to **S12**, **Figure S8**). The reasons to apply the RF are: (*i*) RF divides the variable space in regions according to the response magnitude (how large is the response according to the variables), and (*ii*) it presents reasonably accurate predictions on the explored zones. The choice of the GR responds to: (*i*) GR can separate the variable space in regions and the predictions are more accurate than those from the Random Forest, (*ii*) volcano plots have a similar shape as the Gaussian regressor output (a joint-Gaussian distribution), and (*iii*) the Gaussian regressor, as the BMS, is based on a Bayesian process.

For each run of the BMS for the 2-variables dataset, we choose a prior corresponding to two variables and two parameters (two fitting constants), and a maximum of ten-thousand steps (the first one-thousand steps were burnt), with a thinning of one-hundred was set. For each non-burnt step, the BMS exported: (*i*) the complexity of the current function, (*ii*) the Bayesian Information Criterion (BIC) of the current function, (*iii*) the Sum of Squared Errors (SSE) of the current function, (*iv*) the values of the fitting constants, and the function itself. Each run took from one to two days. Then, the simplest equations were refitted using the same fitting constants obtained by the BMS to check the output accuracy. Finally, the result was plotted and visually examined. The number of replicates needed for finding appropriate functional forms depends on the observable: for conversion and selectivity to CH_4 two replicates were enough, while for selectivity to coke, up to twenty runs were required. The accuracy of the BMS, RF and GR techniques depend on the observable, but as a general rule their SSE values follows the following trend: SSE(GR)<SSE(BMS)<SSE(RF). However, the SSE values for selectivity to coke and CH_3X of BMS (510 and 856 respectively) are much better than those obtained with GR (2925 and 3745 respectively).

Results and Discussion

The approach followed in the present work is illustrated in **Figure 1**, traditional methods (marked in gray) take advantage of the use of reaction profiles and microkinetic modeling to try to obtain rates and other parameters comparable to experiments. The alternative path (indicated by colors) shows the potential of Statistical Learning techniques containing hybrid data for the analysis of the performance descriptors of complex catalytic systems.

Catalytic gas phase hydrodehalogenations is a family of reactions in which halogen elimination from an organic compound is driven by hydrogen addition. The selective transformation of CH_2X_2 into CH_3X is of particular interest, since it is an important step in halogen-mediated natural gas upgrading processes,[95–98] though it presents selectivity issues. In the reaction mechanism, as shown in **Figure S1**, the target CH_3X is formed by the removal of a single halogen and the addition of an H-atom. Alternatively, both halides are eliminated and followed by the sequential addition of H- and halogen-atoms. Coke is formed as a side product from the CH_2 intermediate, whereas CH_4 is generated after hydrogenation of the CH_3 species.

Density Functional Theory results

The DFT (PBE, see **Methods**) dataset containing the geometries and adsorption energies of 74 intermediates derived from CH_2X_2 (for X=F, Cl, Br, and I) on 8 metallic surfaces (total 592) is compiled as a comma separated values (csv) file (see **Density Functional Theory** in **Methods** section, **Equations (S1-2)** in the **Note S1**, and **Tables S2-3** in the **Supporting Information** (**SI**). The geometries and raw adsorption energies are available in the ioChem-BD, in http://dx.doi.org/10.19061/iochem-bd-1-152 and https://iochem-bd.iciq.es/browse/review-collection/100/26403/180f4ec356725fa2cc79c21e).

Experimental Data

The next step, as shown in **Figure 1**, is the acquisition of experimental data. A full dataset implies the evaluation of reactivity while taking the entire CH_2X_2 family into account. However, from a practical point of view, dihalomethanes with X=F or I were not included due to handling issues: HF formation or the high boiling point of CH_2I_2. The experimental dataset contains the reactivity data of SiO_2-supported Fe-, Co-, Ni-, Cu-, Ru-, Rh-, Ag-, Ir-, and Pt-based catalysts (1.0 wt % metal basis) in the hydrodehalogenation of CH_2X_2 (X=Cl, Br), using identical reaction conditions (T = 523 K, P = 1 bar, and time on stream = 15 min) to generate a comprehensive and consistent dataset. The systematic evaluation ensures that: (*i*) only single metal component forms are considered; (*ii*) the nanoparticle size distribution is comparable as shown in the characterization of the fresh catalyst;[95] (*iii*) all generated products (CH_3X, coke, CH_4, and other gas phase side products) are measured with comparable accuracy. Briefly, the utilized experimental dataset consists of: (*i*) CH_2X_2 conversion ($X_{CH_2X_2}$, X=Cl, Br, **Figure 2 a)**), (*ii*) product selectivity (**Figure 2 b)**) and (*iii*) yield to the halomethane, coke, and methane (denoted as S_i and Y_i, respectively, where *i* = (CH_3X, coke, and CH_4)), and (*iv*) the reaction rate, $r_{CH_2X_2}$.

Microkinetic modeling

The state-of-the-art methodologies to determine rates and other parameters, take into account the information condensed in reaction energy profiles. In our case, the full DFT networks towards all the products were obtained only for the hydrodebromination, summarized in **Table S1**.[95] The paths to CH_3Br, CH_4 and coke were considered for all the metals in our study. The reversibility analysis shows that metals cluster in three different groups: Co and Ni, Ru and Rh, Ir, Pt, Cu and Ag (see **Note S2**) according to the most relevant elementary steps. Therefore, differently from traditional microkinetic models, where one metal system bears the information for the mechanistic preferences of the full metal dataset, we observe that at least three different mechanisms and

energy profiles are needed, leading to the selectivity maps that depend on the chosen cluster of metals (see **Figures S3-7**). This points towards the fact that the small errors in the LSR get amplified when trying to address the small energy variations involved in selectivity through the path selection towards different products.

Even if applying the three different microkinetic models, the selectivity patterns depart significantly from experimental observations, (**Figure S2**, panels **a)**, **b)** and **c)**) likely due to the high coverages of poisons as coke and Br. Coke would deposit on the catalyst surface, leading to a decrease over time of the total available active sites, thus limiting the predictability of MK models. Therefore, only systems lean on adsorbed carbons like Ru, Ir and Pt could be qualitatively represented, in line with previous observations.[61] If Br is the poison, strong lateral interactions due to the amount of charge dragged from the surface make adsorption energies highly dependent on the particular coverage, thus breaking the mean-field restriction of identical sites. The hydrodehalogenation system is thus a very good example to test Machine Learning approaches.

Principal Component Analysis

Principal Component Analysis (PCA) was applied to the DFT energy dataset, presented in **Figure 3**, to reduce the dimensionality by finding the main contributors to the adsorption of the intermediates. **Figure 3 a)**, shows the resulting dimensionless eigenvectors, ***w***, that store the information related to the intermediates (*loading* vectors). The metal characteristics ***t*** (*score* vectors) in **Figure 3 b)** represent the energy of the metals (in eV) and are obtained as the product between the centered input matrix and the loading vectors. The three significant principal components cover 89.9, 7.6, and 1.6% of the total variance, while weights were 92.8, 5.2, and 1.2 % (considering only the X=Br dataset).[95] Therefore, when expanding the size of the sampling space by four times, we observe that the contributions are robust but the relative weight between the first and the second component rebalance slightly (89.9 and 7.6 % with respect to 92.8 and

5.2 % for Br only and for all the halogens datasets, respectively). Moreover, the results are halogen independent, allowing generalization. The model can thus estimate the adsorption energy of any intermediate in the network as $E_{ads} = t_1w_1+t_2w_2$, accounting for 97.5% of the values. Selecting the minimum optimal set of intermediate energies that can represent the whole database implies that the selected intermediate should be described exclusively by one of the two principal components, and should provide a higher accuracy when predicting the energies.[79,110] CH_{hcp} and the single halogen atoms X_{hcp} were used as the intermediate descriptors of the first and second component, respectively. The first component is associated with the covalency and can be represented by the E_{ads} of CH, while the second component stands for the redox terms, more univocally described by the adsorption of the atomic halogen species.[79] Thus, the PCA term for metals are transferable, while the weights for the adsorbates, ***w***, follow the trends according to their position in the Periodic Table, see **Figure 3 a)**. The principal components condense all the information stored in the full DFT intermediate energy dataset and act as descriptors when finding catalyst performance equations for activity and product selectivity.

Bayesian Learning

We have applied the Bayesian Machine Scientist (BMS) to identify the functional forms for the experimental macroscopic observables as a function of the DFT-derived atomistic Principal Components. BMS was benchmarked against Random Forest and Gaussian Regressors (RF and GR, respectively). RF constitutes the state-of-the-art while GR follows Bayesian inference. Both reference methods provide reasonable accurate predictions within explored zones, separate space regions according to the response magnitude, and the GR presents functional shapes that are similar to volcano plots, though both techniques belong to the non-explainable class of SL. The experimental observables ($X_{CH_2X_2}$, S_i, Y_i, and $r_{CH_2X_2}$) are all fitted with BMS and the other

methods to the simplest possible meaningful equation taking the two principal components as only variables, as is illustrated in **Figure 1**.

The BMS methodology was first employed in the bromine reaction subset including the conversion of CH_2Br_2 ($X_{CH_2Br_2}$) and selectivity to CH_3Br (S_{CH_3Br}). While the functional forms of conversion and yield are very simple, showing a direct dependence on the Br adsorption energy and an inverse one with CH adsorption energy (see **Note S5, Equations (S13)** and **(S14)**, **Figure S9 a)** and **c)**), the equation for selectivity contained inverse and polynomial terms (up to degree 3) and four constants, which is an indication of potential overfitting (**Equation (S15)**, **Figure S9 b)**). Despite the complexity, the selectivity equation was able to split the {E_{ads}(Br), E_{ads}(CH)}-space in two areas, separating the catalysts depending on coke generation.

A natural extension of this approach is to assess whether the functional forms identified for bromine can be extended to the chlorine-containing reactant. Mathematically, this corresponds to the search of an isomorphism between the conversion, selectivity, and yield surfaces for the halides. In doing so, the metal descriptors are fixed t_i, while the halide and fragment contributions w_i are updated (particularly, C-only containing fragments are identical, as the interaction between C-only fragments and halogens is not considered now, see the **Table S2**). The predictions for the observables are denoted with a super index p (as for example: $X^{p}_{CH_2Cl_2}$). These follow the experimental points once the accompanying constants are refitted (see **Figure S9 d-i)**, **Table S6** and the **Supporting Note S5**). Thus, the conversion (and to some extent the yield) maintain the functional form, recalling the similarities in the volcano shapes found by traditional methods but with two descriptors, as found for the Deacon process,[54] and pointing towards the existence of a generalized equation that can express the performance set for the full halogen family. In turn, the analysis of selectivity, particularly to ${S_{CH_3Cl}}^{p}$, shows that although we do estimate these values correctly, the differences in the fitting constants (proportion of 1 to -20 for Br and Cl, respectively,

in c_2) changes the shape of the surface, visible when comparing **Figures S9 b)** and **e)**. In the latter, an artificial compression along the x-axis, corresponding to the CH adsorption, is shown.

Intrinsically, the different ***w*** values obtained for the different halogens are not enough for the Br-equations to be representative of Cl properties as halogen presence modifies CH adsorption. Adsorbed halogens drag density form the surface **Figure S10,** and modify the work functions (*WF*) of metals, which severely affects covalent contributions to the adsorption energy of coadsorbates, **Figure S11-12**.[111,112] The energies of the organic moieties would be affected most, from 0.02 to 0.12 eV per halogen atom. Therefore, rescaling the ***w*** parameters corresponding to the CH_n fragments is required. The difference between the metal-only *WF* and the electron affinity (E_{ea}) of the isolated halogen atom is taken as a proxy for the penalty of the CH_n fragment adsorption energy $\omega_1' = E_{\mathrm{ads}}(\mathrm{CH}_{hcp}) + WF - E_{\mathrm{ea}}$, this allows aligning the energies of different halogens (bromine and chlorine values are reported in **Table S7**).

Analysis of the Activity Equations

To extend the validation of the approach the BMS was further applied to the full experimental dataset containing $X_{\mathrm{CH_2X_2}}$, S_i, Y_i, and $r_{\mathrm{CH_2X_2}}$ (16 experiments and 4 sets of parameters (ω_1 and ω_2); halide adsorption and CH corrected energy adsorptions; see **Note S6** for further information on benchmark datasets). RF and GR were also performed for comparison (**Note S7** fand **Figures S13** to **S17**). Each of these searches for performance observable equations was run independently and analyzed separately. The criteria for the best function are: (*i*) simplicity, (*ii*) interpretability from a physical point of view, and (*iii*) performance in terms of the lowest sum of square errors, SSE, the acceptance range of which depends on the magnitude of the observables: for conversion and the selectivity, the SSE range is between 100 and 1000 (observables between 0 and 100 percent of product), while for the reaction rate it is between 100 and 2000 (observables

between 0 and 220 s^{-1}). The BMS equations and their associated SSE values are listed in **Table 1**, and their associated fitting constant are reported in **Table S8**. The selected ensemble of equations, together with the corresponding SSE values and fitting constants for each observable, can be found in the **Supporting Notes S6** and **S8**, **Tables S9** to **S24** and **Equations (S16)** to **(S70)**. The validation of the BMS predictions was performed via a Leave-One-Out test (for the results and brief discussion, see **Figures S18** and **S19**, and **Supporting Note S9**).

The BMS result for conversion is expressed as **Equation (4)** in **Table 1,** $X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$, showing a simple function (SSE = 260) with a marked volcano character formed by a direct dependence with the adsorption energy of the halogen and an inverse dependence with carbon fragments ω_1. It is worth to remark that BMS predictions domain is enclosed in the near range of the input variables (as any other SL technique), thus the risk of any discontinuity by means of the inverse dependence with ω_1 is prevented. The $X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$ dependencies point to a possible poisoning of the metal surfaces with the most exothermic halogen-metal bonds, while conversion improves at lower CH adsorption values, as illustrated in **Figure 4 a)**.[113] According to this interpretation, $X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$ is higher for Cl than for Br due to the higher electronegativity. **Equation (4)** can be interpreted as the surrogate models presented by Campbell et al.[32] To test the hypothesis, we have formulated an exponential model (**Equation (10)**) for conversion ($X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$).

$$X^{p}_{CH_2X_2} = c_1 + c_2 e^{-\omega_2} + c_3 e^{-\omega_1} \qquad \textbf{(10)}$$

The exponential expression assumes that conversion is related to the desorption rates of poisoning species, that is, CH and Br, and are in agreement with the BMS model, as illustrates **Figure 4 b)**. Indeed, the inverse dependence of conversion with ω_1 found for BMS can be thought in terms of $e^{-x}=(e^{x})^{-1}$ thus mapping the equations under the same variable space span. Finally, the SSE values of both models are very similar (SSE(BMS)=276, SSE(Exponential)=331), see **Table S9**). These results support the interpretability of the BMS function obtained for conversion

(**Equation (4)**) as the surrogate model of the reaction rates taking into account that the BMS derived expressions are only valid within the descriptors span and cannot be employed for extrapolations. On the other hand, selectivity patterns are too complex to apply to the exponential model satisfactorily.

To test the performance of the BMS algorithm with alternative datasets of the DFT-generated reaction profile (see **Note S6**), we have used two different DFT-variables and corresponding experiments: (*i*) a set of nine adsorption energies (X, H, and the C fragments: CX, CH_nX, and CH_n, where X=Cl, Br, and *n*={1,2,3} on the hcp sites), and (*ii*) the non-zero barriers of dehydrobromination to fit $X_{CH_2Br_2}{}^{p}$ and $S_{CH_3Br}{}^{p}$, corresponding to reactions R3 to R5, R7 and R8 in **Table S1**. In all the cases we have used the same number of variables and fitting constants. For the 9-variables case, the BMS algorithm successfully discards four variables (and parameters) for the simplest models. However, the resulting 5-variable function (**Equation (11)**) is less compact and elegant than the two variables equation, has a slightly better predictive power (SSE of 223 *vs*. 260), requires 2-to-3 times higher computational cost (one-two *vs*. two-three days), and more importantly, it is difficult to interpret from a physical point of view, even if it has a negative exponential term depending on hydrogen adsorption energy, which partially resembles to the Arrhenius equation (indeed, **Equation (S33)** on **Table S16** from the 2-variables $X^{\mathrm{p}}_{CH_2X_2}$ ensemble also presents exponential dependences with ω_1 and ω_2). Particularly, the rest of the ensemble of the 9-variables functions seem to entangle the different variables. The 9-variables functions ensemble are reported in **Table S10**, **Eqs. (S16)** to **(S20)**, their fitting constants and SSE values are reported in **Table S11**.

For the $X^{\mathrm{p}}_{CH_2Br_2}$ when the input space are the reaction barriers of Br-only dataset (only those that are non-zero), after some iterations, the BMS algorithm converges towards an ensemble of functions (**Table S12**, **Equations (S21)** to **(S25)**) depending on all five variables, thus, overfitting.

Even worse is the performance for $S^{\mathrm{p}}_{\mathrm{CH_3Br}}$ as the algorithm diverges (**Table S13**, **Equations (S26)** to **(S30)**). Therefore, the reduction of the dimensionality of DFT-dataset is thus crucial to achieve meaningful and consistent functional forms, thus reinforcing the role of descriptors in heterogeneous catalysis.

$$X^{\mathrm{p}}_{\mathrm{CH_2X_2}} = \frac{E_{\mathrm{ads}}(\mathrm{CH_{2hcp}}) + {c_1}^2}{E_{\mathrm{ads}}(\mathrm{CH_{hcp}})c_2 \frac{E_{\mathrm{ads}}(\mathrm{CHX_{hcp}})}{c_3} e E_{\mathrm{ads}}(\mathrm{H_{hcp}})} \qquad \textbf{(11)}$$

In summary, employing a full energy dataset (including transition states), generates unnecessary issues as the algorithm could generate long sets of coupled descriptors, which increases the optimization time, and is prone to overfitting. In addition, by using PCA, the rationalization and interpretation of the leading terms (descriptors) becomes more transparent.

Analysis of the Selectivity Equations

Next, we addressed selectivity patterns. The selectivity is governed by competing steps in very small adsorption energy ranges at product distribution switches (the so-called cliffs).[91] Three main potential products appear in the hydrodehalogenation process: when hydrogenation dominates, the major product is CH_4 (with a small minority of volatile C2 coupling products), if C-C bond formation prevails, the catalyst cokes. However, the desired product is the semi-hydrogenated CH_3X that mechanistically shares intermediates with both hydrogenation and coking routes. Therefore, the selective regime to CH_3X corresponds to a narrow area and its selectivity equation is the steepest, hence most difficult to fit. As a rule, and for any complex reaction with a different product distribution, a semi-reaction path would be the most difficult to approach mathematically. Therefore, the extremes of the catalytic products (coke and methane) are taken as more robust functions than the CH_3X equation. Consequently, the expressions for the selectivity to coke and methane are analyzed first, and the halomethane selectivity is obtained by subtraction from the

total. Notice the differences in the predicted selectivities from BMS and the different microkinetic models only for the Br system, **Figure S13**.

The prediction for selectivity to CH_4, $S^{\mathrm{p}}_{\mathrm{CH_4}}$ in **Equation (5)** in **Table 1**, presents a quadratic polynomial dependence with ω_1 and ω_2. The interpretation of these results points to a trade-off between both variables. Relatively strong metal-CH bonds promote CH_4, while a large halogen adsorption energy inhibits CH_4 production. Only Pt, Ir and Rh (for Cl) present the right combination between relatively low exothermic halogen adsorption and relatively strong CH bonding (see **Figure 4 c)**). This model, as in the conversion case, is simple and the SSE for **Equation (5)** in **Table 1** is 284. It is significant that both selectivity terms are polynomials of second degree, this closely follows the traditional modeling of atomistic potential energy surfaces as combinations of parabolas even in Marcus developments.[114] Remarkably, CH_4 selectivity arises from a balance between both main descriptors.

The $S^{\mathrm{p}}_{\mathrm{coke}}$ expression (**Equation (6)** in **Table 1**), shows a direct dependence with the exothermicity of ω_2 and an inverse dependence with the stability of ω_1. These results points to the following: a strong halogen-metal bond results in the complete dehalogenation of CH_2X_2, whereas a strong CH-metal bond leads to lower probabilities to generate coke by C-C coupling. In addition, the $S^{\mathrm{p}}_{\mathrm{coke}}$ expression clearly divides the $\{\omega_1, \omega_2\}$-space in four different clusters based on the principal product obtained, recover the main CH_3Br result reported in our previous work,[95] and extends it to CH_3Cl, as illustrated in **Figure 4 d)**. Inside the regions, it is possible to see a volcano shape in the ω_2 direction, which limits the maximal coke production sub-regions (maximal ω_2 and minimal ω_1 modulus). The cosine term modulates the frontier between several Br and Cl points without altering the general dependencies.[115] This is explained by the fact that the computed $\omega_{1,2}$ variables for Br-Ru and Cl-Br are very similar, but the selectivity to coke differ by one order of

magnitude (79 and 7% for Cl and Br, respectively). Due to the complexity of predicting the S_{coke} behavior, the SSE value is 510.

A direct attempt to predict selectivity to CH_3X, $S^{\mathrm{p}}_{\mathrm{CH_3X}}$, leads to complex and difficult to interpret functional forms (see **Table S19**). However, it can be obtained from the carbon-balance, including the use of selectivity of the other two main compounds from $S^{\mathrm{p}}_{\mathrm{CH_4}}$ and $S^{\mathrm{p}}_{\mathrm{coke}}$ (**Equation (7)** in **Table 1**). The main contribution to the SSE value (856) comes from the propagation of the ${S_{\mathrm{coke}}}^{p}$ errors. The ${S_{\mathrm{CH_3X}}}^{p}$ obtained in this way is illustrated in **Figure 4 e)**. From the found dependencies, the more exothermic the halogen adsorption is, the narrower the selective range or productive sub-regions in the $\{\omega_1, \omega_2\}$-space for CH_3Br or CH_3Cl is. The area of those productive sub-regions depends on ω_2, as shown in **Figure 4 e)**. Thus, larger productive sub-regions areas imply higher probabilities of finding points with higher associated selectivity values. Physically, this is interpreted as follows: the stronger the halogen is adsorbed to the surface, the lower is the selectivity to CH_3X. However, this trend does not apply to CH_3Br selectivity over Ru and Ni, which explains why the BMS divides the different regions using maximum selectivity discontinuities (yellow stripes in **Figure 4 e)**).

Analysis of the Yield and Rate Equations

The yields to each of the three main products have been estimated using **Equations (8)** from **Equation (4)** and their respective selectivity expression (**Equations (5)** to **(7)**), all listed in **Table 1**. If we examine the ensembles of yields (**Tables S20** to **S22**, **Equations (S51)** to **(S65)**), the SSE obtained with **Equation (8)** is equal or even lower compared with the other candidate equations (95,193, and 119 for CH_3X, coke, and CH_4, respectively). In general, the separation between regions as shown for the selectivity surfaces are smoothed for yields. The surface for $Y^{\mathrm{p}}_{\mathrm{CH_3X}}$ shows a mixed dependence with ω_2 and ω_1 between $S^{\mathrm{p}}_{\mathrm{CH_4}}$ and $X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$, as illustrated in **Figure**

S13 m). $Y^{\mathrm{p}}_{\mathrm{coke}}$ presents a remarkable similarity to the surface obtained for $Y_{\mathrm{CH_3Br}}{}^{p}$ (**Figure S13 p)**), which points to a direct squared dependence with ω_2 and an inverse squared dependence with ω_1. The $Y^{\mathrm{p}}_{\mathrm{CH_4}}$ surface allows a more direct interpretation, similar to $X^{\mathrm{p}}_{\mathrm{CH_2X_2}}$, as reported in **Figure S13 s)**.

Finally, we have extrapolated our methodology and results to predict the reaction rate of CH_2X_2, $r^{\mathrm{p}}_{\mathrm{CH_2X_2}}$. The resulting function is reported in **Table 1**, **Equation (9)**, which is a second-degree polynomial with two terms: *(i)* the squared difference between ω_1 and ω_2, which is always positive and *(ii)* ω_1, which is negative. Thus, favored CH adsorption and unfavored X adsorption or vice versa (*i.e.*, the difference between ω_1 and ω_2 is high) leads to high rates. The obtained shape is a cliff (**Figure 4 f)**). The Cl-Ru point is the main responsible of the error (60 % of the 1687 error), similarly to the $S^{\mathrm{p}}_{\mathrm{coke}}$ case.

Benchmark of BMS versus common Statistical Learning techniques

The Random Forest and Gaussian Regressors divide the CH_2X_2 conversion, product selectivity and yield, and rate spaces qualitatively in the same manner as BMS (**Figures S13** and **S14** for $r^{\mathrm{p}}_{\mathrm{CH_2X_2}}$, and **Figures S15** to **S17**), particularly for conversion (**Figures S13 a)**, **b)** and **c)**). The robustness of conversion (**Equation (4)**) further reinforces the heuristic knowledge of the strength of volcano plots in catalysis. For product selectivity and yield, and rate, the RF was able to place the frontiers of the regions on the $\{\omega_1, \omega_2\}$-space with similar ω_2 values. The RF is difficult to interpret as no equation is derived (within the given space); GR is even less interpretable, as discerning selectivity with Gaussian functions is difficult due to the sharp nature of selectivity cliffs. Furthermore, the accuracy of BMS prediction is much better than Random Forest or Gaussian regressors for the selectivity to CH_3X and coke.

On the comparison of BMS to classical methodologies

In this final section we would like to discuss the advantages, limitations and challenges for future hybrid computational models to represent experimental observables, summarized in **Figure 5**. Microkinetic models have found extreme success in addressing the reactivity of metals for simple reaction networks leading to one product. Hybrid data in this case, further reinforces the soundness of these models and thus, a complete understanding of the reactivity can be rooted in first principles data.

As on the interpretability of the models, the equations derived from BMS are only qualititavely interpretable, given indications of the most demanding process or the most likely poisons. For instance, we can extrapolate some qualitative insight from the analysis of the conversion surface. The regions in which we have a maximum poisoning (from CH or the halogen), the most abundant species at the steady state would be carbon fragments or halogens. Under these conditions, we can assume that the rate determining step is the adsorption of the reactants.[37] In this regard, the insight provided is of less quality that the microkinetic model, in part because the complex DFT analysis profile is summarized into only the key steps. The BMS equations are only useful in the ranges expanded by the variables and thus extrapolations can lead to non-physically meaningful results. However, there are many aspects where SL methodologies provide less insightful predictions at the atomistic scale. Nevertheless, SL predictions are more robust and useful than standard MK results. Particularly as reactions become more complex and reactivity involves more elementary steps the relevance of MK is more limited.

MK models are no longer able to handle the issues associated to different phase (in some cases even mixed ones), dynamic rearrangements, site blocking due to poisoning and in general related to the material gap problem. Although we have presented our results for metal-only systems, the extension to structural differences should be the next step in generalization. Another limitation of MK regards thegeneralization to a family of reactants, as shown here. This is possible by a

rescaling in the variable for BMS but it was less evident for MK models, even if evidence within a reactant family are as old as Brønsted studies.

Regarding the type of data employed by both methods, hybrid sets are very valuable as they constitute the true benchmark. From the MK standpoint, a single rate or yield and the full reaction profile for one single metal system has been the state of the art. ML techniques can better benefit from the extended systematization of the experimental data, as automatization provides accessibility to larger amounts of identically generated kinetic data, while diminish the burden of the DFT part as they only depend on the adsorption energies. Thus, automatically generated data are less affected by the intrinsic DFT errors. Besides, algorithms to identify transition states allow the unbiased identification of the descriptors as they can be rooted in robust SL techniques and not in heuristic priors.

When considering accuracy, MK(DFT) models are excellent in presenting qualitatively the activity, but have severe issues in reproducing selectivity patterns. This is due to the fact that the Linear-Scaling Relationships employed to simplify MK equations typically present small errors that do not provide enough accuracy to address the small energy differences involved in the prediction of selectivity. A clustering technique applied to our PCA descriptors for instance would tend to group elements right where the cliff changing selectivities appears (Ni and Co, for instance). Thus, even if very accurate DFT values could be obtained, it is unlikely that sharp selectivity patterns for multiproduct reactions can be obtained using MK models. In contrast, BMS clearly identifies the patterns when built on robust variables as those derived by PCA. Finally, BMS derived functions could be used as surrogate equations in multiscale approaches and even as a systematic unbiased way to approach standard microkinetic models to experimental data.

Conclusions

Modeling in catalysis has its early origin on phenomenological observations, giving empirical equations that were simple, such as the power laws. With the introduction of microkinetics, such models became more complex and could only be solved numerically. The gold standard in reaction modeling over the last decades has been coupling Density Functional Theory reaction profiles and mechanisms to microkinetic modeling if needed, the energy profiles are massaged to account for errors due to set up or intrinsic to DFT. However, going further from metal catalytic systems as the material complexity, dynamicity and the number of elementary steps in the reactions increase in an exponential way and thus these systems might be not fully addressable by traditional DFT based methods making classical microkinetic methods fail in their predictions. The availability of consistent, systematic, annotated experimental and computational datasets provides the needed clean raw data to apply SL to apprehend this complexity. The present work shows the possibility to go beyond traditional schemes based on DFT energy profiles for a reaction for which the standard microkinetic modeling fails. To this end, we have qualitatively addressed rates and other experimental parameters using multiscale approaches, in a case that imply a different phase for some types of metals (carbides for Co or bromides for Ag and Cu): hydrodehalogenation (X=Cl, Br). The methodology presented is robust enough to generalize the equations to an entire family of reactants.

We have taken a hybrid data set composed of the activity, selectivity and rate of hydrodehalogenation reactions on identically prepared metal nanoparticles with similar sizes. The energies of the reaction intermediates were evaluated by DFT and the overall hybrid dataset was employed to feed the SL models (Bayesian Machine Scientist) and derive performance equations that can be physically interpreted. A first step requires identifying the descriptors through the dimensionality reduction of the problem, which is mandatory to derive robust functions. The traditionally identified volcanos are retrieved with data-driven methodologies

indicating the robustness of this functional shape to describe heterogeneous catalyst problems. Generalized performance date can be found for a family of transformations, in the present case Cl or Br-hydrodehalogenations. The procedure can be expanded to other sets of catalysts and can be particularly successful when investigating families of similar reactants or larger number of atoms where standard DFT approaches fail, such as (de)polymerizations or biomass conversion. Selectivity involves a more elusive set of parameters and while the extremes of the phase reaction space are easier to describe (semi-reactions are more ill-defined due to the intricacy of the reaction networks). This problem should be assessed in detail by future studies as SL techniques have been focused on clustering, whereas selectivity is about defining differences in very narrow energy spans. The equations derived from BMS can be used in coarse-grain models and reactor design to avoid the instabilities and error propagation observed when employing microkinetic modeling on the DFT results. Overall we have shown how the results from BMS can be mapped to previous surrogate models in the literature and outperforms MK(DFT) thus having the potential to fill the gap where microkinetic modeling is not possible, due to phase transitions, exceedingly complex reaction networks, multiple products and selectivity issues. Finally, this hybrid data approach can be used to identify and explain the descriptors in future catalyst design.

Associated content

Supporting Information, which contains all the details on the computational methods and complementary results reported on the **Supporting Figures S1** to **S19**, **Supporting Tables S1** to **S24**, **Supporting Equations (S1)** to **(S70)**, the **Supporting Notes S1** to **S9.** DFT structures and adsorption energies can be found on the ioChem-BD platform in the following links: (*i*) the complete hydrodehalogenation set in http://dx.doi.org/10.19061/iochem-bd-1-152 and https://iochem-bd.iciq.es/browse/review-collection/100/26403/180f4ec356725fa2cc79c21e, and (*ii*) the hydrodebromination set in http://dx.doi.org/10.19061/iochem-bd-1-150 and http://dx.doi.org/10.19061/iochem-bd-1-152. All other relevant source data are available from the corresponding author upon reasonable request.

Authors Information:

Corresponding authors:

Javier Pérez-Ramírez - *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.* Email: jpr@chem.ethz.ch

Núria López - *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, Av. Països Catalans 16, 43007, Tarragona, Spain.* Email: nlopez@iciq.es

Authors:

Sergio Pablo-García – *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, Av. Països Catalans 16, 43007, Tarragona, Spain.*

Albert Sabadell-Rendón – *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, Av. Països Catalans 16, 43007, Tarragona, Spain.*

Ali J. Saadun - *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.*

Santiago Morandi - *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, Av. Països Catalans 16, 43007, Tarragona, Spain.*

Acknowledgements:

This work was supported by the ETH Research Grant ETH-43181 and Ministerio de Ciencia e Innovación (Ref. RTI2018-101394-BI00). This publication was in part created in NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation. The authors thank BSC-RES for generously providing computational resources. The authors also would like to explicitly thank Reviewer 3 for their valuable suggestions.

Notes:

The authors declare no competing interests.

Figures and Tables

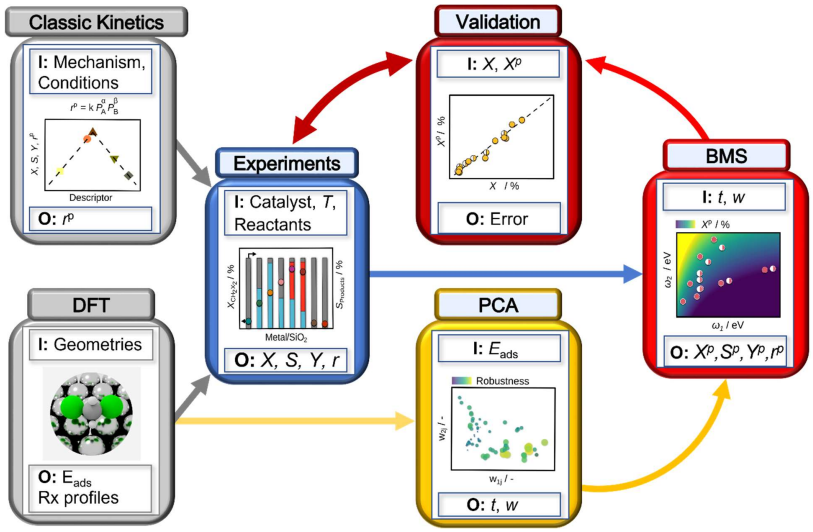


Figure 1: Workflow of the present study, where grey arrows represent classical approaches to catalysis modeling. Colored pathways indicate the route and methods applied for analyzing conversion (X), selectivity (S), yield (Y) and rate (r) of CH_2X_2 hydrodehalogenation. Predicted values of the experimental observables are denoted with the super-index p (as example: observable X, prediction X^p).

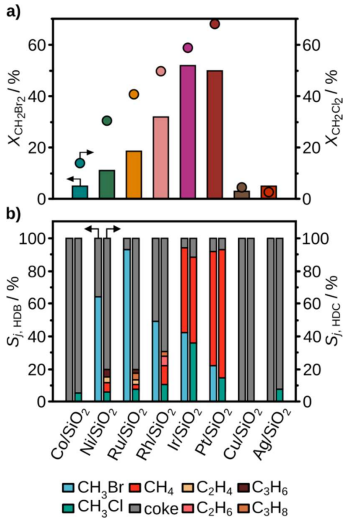


Figure 2: **a)** Conversion of CH_2Br_2 and CH_2Cl_2, and **b)** product selectivity of the catalysts in CH_2X_2 hydrodehalogenation. In panel **a)**, the conversion was assessed at a constant space velocity of F_T/W_{cat} = 40 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in hydrodebromination (HDB) and F_T/W_{cat} = 100 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in hydrodechlorination (HDC), while product selectivities in **b)** were determined at *ca.* 20% CH_2X_2 conversion achieved by adjusting the space velocity in the range of F_T/W_{cat} = 20-150 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in HDB and 70-350 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in HDC. Other reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:4.5:65.5 (X=Br, Cl), T = 523 K, P = 1 bar, and *tos* = 15 min.

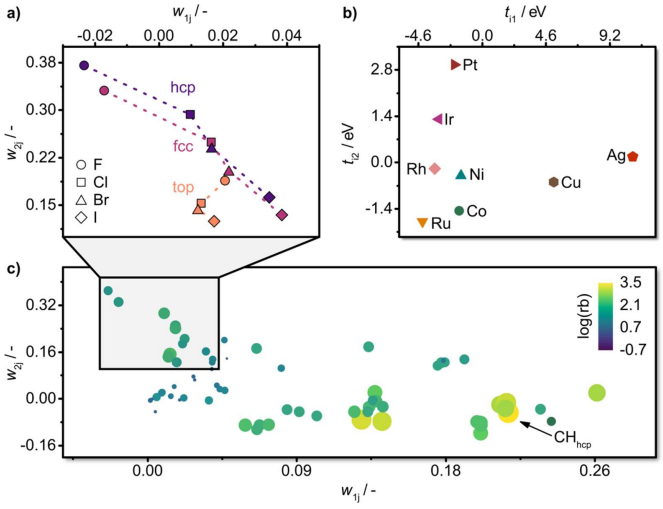


Figure 3: Principal component analysis contributions of **a)** the halogens to the two principal components depending on the adsorption position, **b)** the metallic surfaces to the principal components, and **c)** the absorbed intermediates to the two principal components. The color code and the size represent the robustness term associated with the accuracy of the intermediate as descriptor.

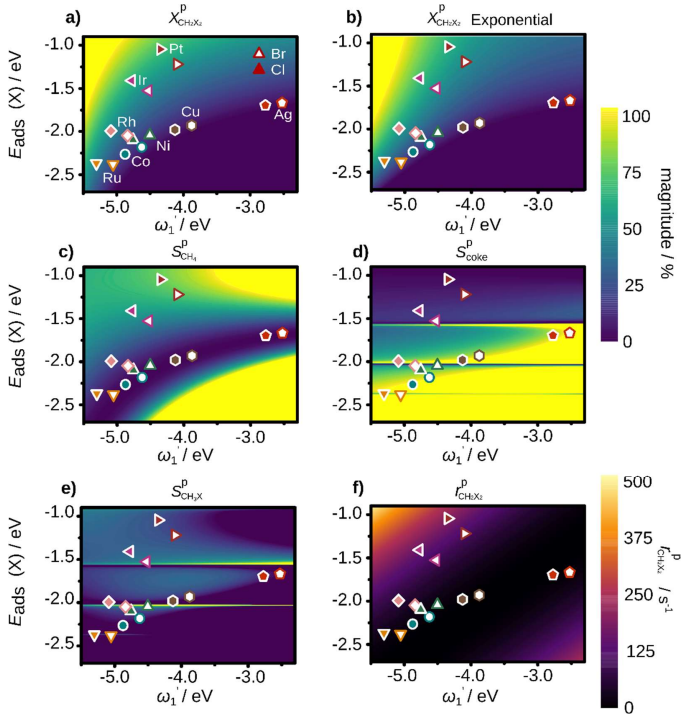


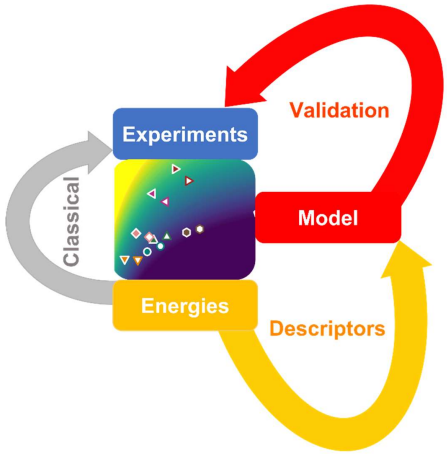
Figure 4: Predicted surfaces using the Bayesian Machine Scientist (Bayesian, **Equations (1)** to **(6)** and **(11)**), of: CH_2X_2 conversion, ($X^{p}_{CH_2X_2}$, **a)**), CH_2X_2 conversion using the exponential model, ($X_{CH_2X_2}{}^{p}$ Exponential, **b)**) selectivity to CH_3X, coke, and CH_4 ($S^{p}_{CH_4}$, **c)**; S^{p}_{coke}, **d)**; $S^{p}_{CH_3X}$, **e)**;), and CH_2X_2 rate ($r^{p}_{CH_2X_2}$, **f)**). All the predictions are presented as a function of the adsorption energy of Br, Cl (denoted as X), and CH. E_{ads}(CH) are corrected with the corresponding metal Work Function (*WF*) and halogen Electron Affinity (E_{ea}) (ω_1') on the hcp sites ($E_{ads}(X_{hcp})$ and $E_{ads}(CH_{hcp})$, respectively) of Co, Ni, Ru, Rh, Ir, Pt, Cu, and Ag.

Model feature	Microkinetic Modeling	Bayesian Machine Scientist
Reactivity complexity	Best in single product reactions	Multiproduct reaction network
Interpretability	Physically interpretable Direct mechanistic insights	Physically interpretable Not direct mechanistic insights
Dataset content and size	Full reaction profile One experimental measure	Only key adsorption energies All experimental data available
Generalizability	Only for simple systems No issues with variables span	For an entire family of reactants Values near to initial variables
Accuracy	Reasonable for activity Poor for selectivity Might need DFT-profiles fitting	Excellent for activity Excellent for selectivity Might need energy rescaling

Figure 5: Comparison between the key features of microkinetic models and the BMS approach. (In green if the feature is better or equal for compared to the alternative model, in red if the feature is worst).

Table 1: BMS equations for the conversion of CH_2X_2 ($X_{CH_2X_2}{}^p$, X=Cl, Br), selectivity to CH_3X, coke, and CH_4 ($S_i{}^p$, *i* = CH_3X, coke, CH_4), corresponding yields ($Y_i{}^p$), and reaction rate, $r_{CH_2X_2}{}^p$. $\omega_2{}' = E_{\mathrm{ads}}(\mathrm{X_{hcp}}) + \mathrm{c_3}$, $\omega_1 = \omega_1' + \mathrm{c_4}$, and $\omega_1{}' = E_{ads}(\mathrm{CH}_{hcp}) + WF - E_{\mathrm{ea}}$, and statistical error (SSE). The c_i coefficients and individual SE are presented in **Tables S4** and **S5.** The comparison plots between experimental and predicted data are in **Figure S9**.

Observable	Equation	SSE	**Eq.**
$X_{CH_2X_2}$	$X^{\mathrm{p}}_{CH_2X_2} = -\omega_2 \mathrm{c}_1 + \frac{\mathrm{c}_2{}^2}{\omega_1}$	260	**(4)**
S_{CH_4}	$S^{\mathrm{p}}_{CH_4} = \left(\omega_2 + \omega_1 + \omega_2\left(\omega_2 - \frac{1}{\mathrm{c}_1}\right)(\omega_1 + \mathrm{c}_2)^2\right)^2$	284	**(5)**
S_{coke}	$S^{\mathrm{p}}_{\mathrm{coke}} = \frac{\omega_2 \mathrm{c}_1}{\omega_2 \mathrm{c}_2 - \omega_1}(\omega_2 + \mathrm{c}_2)\left(\omega_2 + \mathrm{c}_2 + \frac{\mathrm{c}_2}{\omega_1 \cos\left(\frac{\omega_2{}^2}{\mathrm{c}_2}\right)}\right)$	510	**(6)**
S_{CH_3X}	$S^{\mathrm{p}}_{CH_3X} = 100 - S_{\mathrm{coke}}{}^p - S_{CH_4}{}^p$	856	**(7)**
$Y_{i\in\left\{\substack{CH_3X,\\ \mathrm{coke},CH_4}\right\}}$	$Y^{\mathrm{p}}_{i\in\{CH_3X,\,\mathrm{coke},CH_4\}} = \frac{X_{CH_2X_2}{}^p S_i{}^p}{100}$	{95,193,119}	**(8)**
$r_{CH_2X_2}$	$r^{\mathrm{p}}_{CH_2X_2} = \left(\frac{\omega_1 + \mathrm{c}_2\omega_2}{\mathrm{c}_1}\right)^2 + \omega_1$	1687	**(9)**


Validation
Experiments
Classical
Model
Energies
Descriptors

1
2
3 **TOC**

36

References

(1) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*; Wiley, 2003.

(2) Temkin, S. I.; Yakobson, B. I. Diffusion-Controlled Reactions of Chemically Anisotropic Molecules. *J. Phys. Chem.* **1984**, *88* (13), 2679–2682.

(3) Boudart, M. Kinetics on Ideal and Real Surfaces. *AIChE J.* **1956**, *2* (1), 62–64.

(4) Aris, R.; Mah, R. H. S. Independence of Chemical Reactions. *Ind. Eng. Chem. Fundam.* **1963**, *2* (2), 90–94.

(5) Dumesic, J. A.; Aparicio, L. M.; Rekoske, J. E.; Treviño, A. A.; Rudd, D. F. *The Microkinetics of Heterogeneous Catalysis*; American Chemical Society, 1993.

(6) Chatterjee, A.; Vlachos, D. G. An Overview of Spatial Microscopic and Accelerated Kinetic Monte Carlo Methods. *J. Comput. Mater. Des. 2007 142* **2007**, *14* (2), 253–308.

(7) Andersen, M.; Panosetti, C.; Reuter, K. A Practical Guide to Surface Kinetic Monte Carlo Simulations. *Front. Chem.* **2019**, *0* (APR), 202.

(8) Ravipati, S.; Savva, G. D.; Christidi, I.-A.; Guichard, R.; Nielsen, J.; Réocreux, R.; Stamatakis, M. Coupling the Time-Warp Algorithm with the Graph-Theoretical Kinetic Monte Carlo Framework for Distributed Simulations of Heterogeneous Catalysts. *Comput. Phys. Commun.* **2021**, *270*, 108148.

(9) Hansen, M. H.; Nørskov, J. K.; Bligaard, T. First Principles Micro-Kinetic Model of Catalytic Non-Oxidative Dehydrogenation of Ethane over Close-Packed Metallic Facets. *J. Catal.* **2019**, *374*, 161–170.

(10) Stegelmann, C.; Schiødt, N. C.; Campbell, C. T.; Stoltze, P. Microkinetic Modeling of Ethylene Oxidation over Silver. *J. Catal.* **2004**, *221* (2), 630–649.

(11) Salciccioli, M.; Chen, Y.; Vlachos, D. G. Microkinetic Modeling and Reduced Rate Expressions of Ethylene Hydrogenation and Ethane Hydrogenolysis on Platinum. *Ind. Eng. Chem. Res.* **2010**, *50* (1), 28–40.

(12) Konstantinos, A.; G. Vlachos, D. Surface Chemistry Dictates Stability and Oxidation State of Supported Single Metal Catalyst Atoms. *Chem. Sci.* **2020**, *11* (6), 1469–1477.

(13) Chutia, A.; Thetford, A.; Stamatakis, M.; Catlow, C. R. A. A DFT and KMC Based Study on the Mechanism of the Water Gas Shift Reaction on the Pd(100) Surface. *Phys. Chem. Chem. Phys.* **2020**, *22* (6), 3620–3632.

(14) Andersen, M.; Plaisance, C. P.; Reuter, K. Assessment of Mean-Field Microkinetic Models for CO Methanation on Stepped Metal Surfaces Using Accelerated Kinetic Monte Carlo. *J. Chem. Phys.* **2017**, *147* (15), 152705.

(15) Pineda, M.; Stamatakis, M. Beyond Mean-Field Approximations for Accurate and Computationally Efficient Models of on-Lattice Chemical Kinetics. *J. Chem. Phys.* **2017**, *147* (2), 024105.

(16) Lian, Z.; Ali, S.; Liu, T.; Si, C.; Li, B.; Su, D. S. Revealing the Janus Character of the Coke Precursor in the Propane Direct Dehydrogenation on Pt Catalysts from a KMC Simulation. *ACS Catal.* **2018**, *8* (5), 4694–4704.

(17) Jørgensen, M.; Grönbeck, H. Selective Acetylene Hydrogenation over Single-Atom Alloy Nanoparticles by Kinetic Monte Carlo. *J. Am. Chem. Soc.* **2019**, *141* (21), 8541–8549.

(18) Huš, M.; Grilc, M.; Pavlišič, A.; Likozar, B.; Hellman, A. Multiscale Modelling from Quantum Level to Reactor Scale: An Example of Ethylene Epoxidation on Silver Catalysts. *Catal. Today* **2019**, *338*, 128–140.

(19) Singh, S.; Li, S.; Carrasquillo-Flores, R.; Alba-Rubio, A. C.; Dumesic, J. A.; Mavrikakis, M. Formic Acid Decomposition on Au Catalysts: DFT, Microkinetic Modeling, and Reaction Kinetics Experiments. *AIChE J.* **2014**, *60* (4), 1303–1319.

(20) Kopač, D.; Huš, M.; Ogrizek, M.; Likozar, B. Kinetic Monte Carlo Simulations of Methanol Synthesis from Carbon Dioxide and Hydrogen on Cu(111) Catalysts: Statistical Uncertainty Study. *J. Phys. Chem. C* **2017**, *121* (33), 17941–17949.

(21) Teschner, D.; Novell-Leruth, G.; Farra, R.; Knop-Gericke, A.; Schlögl, R.; Szentmiklósi, L.; Hevia, M. G.; Soerijanto, H.; Schomäcker, R.; Pérez-Ramírez, J.; López, N. In Situ Surface Coverage Analysis of RuO2-Catalysed HCl Oxidation Reveals the Entropic Origin of Compensation in Heterogeneous Catalysis. *Nat. Chem. 2012 49* **2012**, *4* (9), 739–745.

(22) Nikbin, N.; Caratzoulas, S.; Vlachos, D. G. A First Principles-Based Microkinetic Model for the Conversion of Fructose to 5-Hydroxymethylfurfural. *ChemCatChem* **2012**, *4* (4), 504–511.

(23) Li, Q.; García-Muelas, R.; López, N. Microkinetics of Alcohol Reforming for H\-(2) Production from a FAIR Density Functional Theory Database. *Nat. Commun.* **2018**, *9* (1), 1–8.

(24) Frei, M. S.; Mondelli, C.; García-Muelas, R.; Kley, K. S.; Puértolas, B.; López, N.; Safonova, O. V.; Stewart, J. A.; Curulla Ferré, D.; Pérez-Ramírez, J. Atomic-Scale Engineering of Indium Oxide Promotion by Palladium for Methanol Production via CO2 Hydrogenation. *Nat. Commun. 2019 101* **2019**, *10* (1), 1–11.

(25) Piccinin, S.; Stamatakis, M. Steady-State CO Oxidation on Pd(111): First-Principles Kinetic Monte Carlo Simulations and Microkinetic Analysis. *Top. Catal. 2016 601* **2016**,

60 (1), 141–151.

(26) Huš, M.; Hellman, A. Ethylene Epoxidation on Ag(100), Ag(110), and Ag(111): A Joint Ab Initio and Kinetic Monte Carlo Study and Comparison with Experiments. *ACS Catal.* **2018**, *9* (2), 1183–1196.

(27) Ovesen, C. V.; Clausen, B. S.; Hammershøi, B. S.; Steffensen, G.; Askgaard, T.; Chorkendorff, I.; Nørskov, J. K.; Rasmussen, P. B.; Stoltze, P.; Taylor, P. A Microkinetic Analysis of the Water–Gas Shift Reaction under Industrial Conditions. *J. Catal.* **1996**, *158* (1), 170–180.

(28) Campbell, C. T. Future Directions and Industrial Perspectives Micro- and Macro-Kinetics: Their Relationship in Heterogeneous Catalysis. *Top. Catal. 1994 13* **1994**, *1* (3), 353–366.

(29) Campbell, C. T. The Degree of Rate Control: A Powerful Tool for Catalysis Research. *ACS Catal.* **2017**, *7* (4), 2770–2779.

(30) Cortright, R. D.; Dumesic, J. A. Kinetics of Heterogeneous Catalytic Reactions: Analysis of Reaction Schemes. *Adv. Catal.* **2001**, *46*, 161–264.

(31) Kozuch, S.; Shaik, S. A Combined Kinetic−Quantum Mechanical Model for Assessment of Catalytic Cycles: Application to Cross-Coupling and Heck Reactions. *J. Am. Chem. Soc.* **2006**, *128* (10), 3355–3365.

(32) Wolcott, C. A.; Medford, A. J.; Studt, F.; Campbell, C. T. Degree of Rate Control Approach to Computational Catalyst Screening. *J. Catal.* **2015**, *330*, 197–207.

(33) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J.*

Phys. Chem. C **2017**, *121* (46), 25847–25863.

(34) Brezny, A. C.; Landis, C. R. Development of a Comprehensive Microkinetic Model for Rh(Bis(Diazaphospholane))-Catalyzed Hydroformylation. *ACS Catal.* **2019**, *9* (3), 2501–2513.

(35) Jaraíz, M.; Rubio, J. E.; Enríquez, L.; Pinacho, R.; López-Pérez, J. L.; Lesarri, A. An Efficient Microkinetic Modeling Protocol: Start with Only the Dominant Mechanisms, Adjust All Parameters, and Build the Complete Model Incrementally. *ACS Catal.* **2019**, *9* (6), 4804–4809.

(36) Sutton, J. E.; Vlachos, D. G. Building Large Microkinetic Models with First-Principles′ Accuracy at Reduced Computational Cost. *Chem. Eng. Sci.* **2015**, *121*, 190–199.

(37) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1* (1), 37–46.

(38) Sabatier, P. The Method of Direct Hydrogenation by Catalysis. *Nobel Lect.* **1912**.

(39) Balandin, A. A. Modern State of the Multiplet Theor of Heterogeneous Catalysis1. *Adv. Catal.* **1969**, *19* (C), 1–210.

(40) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96–103.

(41) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77* (2), 334–338.

(42) Brönsted, J. N.; Pedersen, K. Die Katalytische Zersetzung Des Nitramids Und Ihre Physikalisch-Chemische Bedeutung. *Z Phys. Chem.* **1924**, *108* (1), 185–235.

(43) Evans, M. G.; Polanyi, M. Further Considerations on the Thermodynamics of Chemical

Equilibria and Reaction Rates. *Trans. Faraday Soc.* **1936**, *32* (0), 1333–1360.

(44) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. J. H. Universality in Heterogeneous Catalysis. *J. Catal.* **2002**, *209* (2), 275–278.

(45) Mazeau, E. J.; Satpute, P.; Blöndal, K.; Goldsmith, C. F.; West, R. H. Automated Mechanism Generation Using Linear Scaling Relationships and Sensitivity Analyses Applied to Catalytic Partial Oxidation of Methane. *ACS Catal.* **2021**, *11* (12), 7114–7125.

(46) Majumdar, P.; Greeley, J. Generalized Scaling Relationships on Transition Metals: Influence of Adsorbate-Coadsorbate Interactions. *Phys. Rev. Mater.* **2018**, *2* (4), 045801.

(47) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier Principle to a Predictive Theory of Transition-Metal Heterogeneous Catalysis. *J. Catal.* **2015**, *328*, 36–42.

(48) Valter, M.; Santos, E. C. dos; Pettersson, L. G. M.; Hellman, A. Selectivity of the First Two Glycerol Dehydrogenation Steps Determined Using Scaling Relationships. *ACS Catal.* **2021**, *11* (6), 3487–3497.

(49) Jørgensen, M.; Grönbeck, H. Scaling Relations and Kinetic Monte Carlo Simulations To Bridge the Materials Gap in Heterogeneous Catalysis. *ACS Catal.* **2017**, *7* (8), 5054–5061.

(50) Rankin, R. B.; Greeley, J. Trends in Selective Hydrogen Peroxide Production on Transition Metal Surfaces from First Principles. *ACS Catal.* **2012**, *2* (12), 2664–2672.

(51) Wu, H.; Sutton, J. E.; Guo, W.; Vlachos, D. G. Volcano Curves for in Silico Prediction of Mono- and Bifunctional Catalysts: Application to Ammonia Decomposition. *J. Phys.*

Chem. C **2019**, *123* (44), 27097–27104.

(52) Pérez-Ramírez, J.; López, N. Strategies to Break Linear Scaling Relationships. *Nat. Catal.* **2019**, *2* (11), 971–976.

(53) Gu, G. H.; Mullen, C. A.; Boateng, A. A.; Vlachos, D. G. Mechanism of Dehydration of Phenols on Noble Metals via First-Principles Microkinetic Modeling. *ACS Catal.* **2016**, *6* (5), 3047–3055.

(54) Toftelund, A.; Man, I. C.; Hansen, H. A.; Abild-Pedersen, F.; Bligaard, T.; Rossmeisl, J.; Studt, F. Volcano Relations for Oxidation of Hydrogen Halides over Rutile Oxide Surfaces. *ChemCatChem* **2012**, *4* (11), 1856–1861.

(55) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-Principles-Based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2* (2), 659–670.

(56) Zhang, Z.; Zandkarimi, B.; Alexandrova, A. N. Ensembles of Metastable States Govern Heterogeneous Catalysis on Dynamic Interfaces. *Acc. Chem. Res.* **2020**, *53* (2), 447–458.

(57) Falsig, H.; Hvolbæk, B.; Kristensen, I. S.; Jiang, T.; Bligaard, T.; Christensen, C. H.; Nørskov, J. K. Trends in the Catalytic CO Oxidation Activity of Nanoparticles. *Angew. Chemie - Int. Ed.* **2008**, *47* (26), 4835–4839.

(58) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9* (4), 2752–2759.

(59) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catal.* **2019**, *9* (8), 6624–6647.

(60) Bhandari, S.; Rangarajan, S.; Mavrikakis, M. Combining Computational Modeling with Reaction Kinetics Experiments for Elucidating the in Situ Nature of the Active Site in Catalysis. *Acc. Chem. Res.* **2020**, *53* (9), 1893–1904.

(61) Xu, L.; Stangland, E. E.; Dumesic, J. A.; Mavrikakis, M. Hydrodechlorination of 1,2-Dichloroethane on Platinum Catalysts: Insights from Reaction Kinetics Experiments, Density Functional Theory, and Microkinetic Modeling. *ACS Catal.* **2021**, *11*, 7890–7905.

(62) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Kitchin, J. R.; Chen, J. G.; Pandelov, S.; Stimming, U. Trends in the Exchange Current for Hydrogen Evolution. *J. Electrochem. Soc.* **2005**, *152* (3), J23.

(63) Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Nørskov, J. K. Toward Computational Screening in Heterogeneous Catalysis: Pareto-Optimal Methanation Catalysts. *J. Catal.* **2006**, *239* (2), 501–506.

(64) Artrith, N. Learning What Makes Catalysts Good. *Matter* **2020**, *3* (4), 985–986.

(65) Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The IoChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55* (1), 95–103.

(66) Bo, C.; Maseras, F.; López, N. The Role of Computational Results Databases in Accelerating the Discovery of Catalysts. *Nat. Catal.* **2018**, *1* (11), 809–810.

(67) Computation and Machine Learning for Chemistry https://www.nature.com/collections/gcijejjahe (accessed 2021 -06 -22).

(68) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**,

10 (3), 2260–2297.

(69) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning:Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365.

(70) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555.

(71) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3* (12), e1701816.

(72) Naik, R. R.; Tiihonen, A.; Thapa, J.; Batali, C.; Liu, Z.; Sun, S.; Buonassisi, T. Discovering Equations That Govern Experimental Materials Stability under Environmental Stress Using Scientific Machine Learning. *arXiv* **2021**, https://arxiv.org/abs/2106.10951v1.

(73) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater. 2017 31* **2017**, *3* (1), 1–13.

(74) Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater. 2019 51* **2019**, *5* (1), 1–6.

(75) Liu, X.; Xiao, J.; Peng, H.; Hong, X.; Chan, K.; Nørskov, J. K. Understanding Trends in Electrochemical Carbon Dioxide Reduction Rates. *Nat. Commun.* **2017**, *8* (1), 1–7.

(76) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *J. Phys. Chem. Lett.* **2020**, *11* (9), 3185–3191.

(77) Tran, K.; Neiswanger, W.; Broderick, K.; Xing, E.; Schneider, J.; Ulissi, Z. W. Computational Catalyst Discovery: Active Classification through Myopic Multiscale Sampling. *J. Chem. Phys.* **2021**, *154* (12), 124118.

(78) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem* **2020**, *6* (11), 3100–3117.

(79) García-Muelas, R.; López, N. Statistical Learning Goes beyond the D-Band Model Providing the Thermochemistry of Adsorbates on Transition Metals. *Nat. Commun.* **2019**, *10* (1), 1–7.

(80) Foppa, L.; Ghiringhelli, L. M.; Girgsdies, F.; Hashagen, M.; Kube, P.; Hävecker, M.; Carey, S. J.; Tarasov, A.; Kraus, P.; Rosowski, F.; Schlögl, R.; Trunschke, A.; Scheffler, M. Materials Genes of Heterogeneous Catalysis from Clean Experiments and Artificial Intelligence. *arXiv* **2021**, https://arxiv.org/abs/2102.08269v1.

(81) Meyer, B.; Sawatlon, B.; Heinen, S.; Von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069–7077.

(82) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction Trends between Single Metal Atoms and Oxide Supports Identified with Density Functional Theory and Statistical Learning. *Nat. Catal.* **2018**, *1* (7), 531–539.

(83) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424).

(84) Felton, K. C.; Rittig, J. G.; Lapkin, A. A. Summit: Benchmarking Machine Learning

Methods for Reaction Optimisation. *Chemistry–Methods* **2021**, *1* (2), 116–122.

(85) Xiong, J.; Shi, S. Q.; Zhang, T. Y. A Machine-Learning Approach to Predicting and Understanding the Properties of Amorphous Metallic Alloys. *Mater. Des.* **2020**, *187*, 108378.

(86) Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem. 2016 84* **2016**, *8* (4), 331–337.

(87) Hibbert, D. B.; Armstrong, N. An Introduction to Bayesian Methods for Analyzing Chemistry Data: Part II: A Review of Applications of Bayesian Methods in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *97* (2), 211–220.

(88) Pedersen, J. K.; Clausen, C. M.; Krysiak, O. A.; Xiao, B.; Batchelor, T. A. A.; Löffler, T.; Mints, V. A.; Banko, L.; Arenz, M.; Savan, A.; Schuhmann, W.; Ludwig, A.; Rossmeisl, J. Bayesian Optimization of High-Entropy Alloy Compositions for Electrocatalytic Oxygen Reduction**. *Angew. Chemie Int. Ed.* **2021**, *60* (45), 24144–24152.

(89) Rudy, S. H.; Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Data-Driven Discovery of Partial Differential Equations. *Sci. Adv.* **2017**, *3* (4).

(90) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. A Bayesian Machine Scientist to Aid in the Solution of Challenging Scientific Problems. *Sci. Adv.* **2020**, *6* (5), eaav6971.

(91) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879–6889.

(92) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377.

(93) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), 015016.

(94) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134–1140.

(95) Saadun, A. J.; Pablo-García, S.; Paunović, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J. Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* **2020**, *10* (11), 6129–6143.

(96) Saadun, A. J.; Zichittella, G.; Paunović, V.; Markaide-Aiastui, B. A.; Mitchell, S.; Pérez-Ramírez, J. Epitaxially Directed Iridium Nanostructures on Titanium Dioxide for the Selective Hydrodechlorination of Dichloromethane. *ACS Catal.* **2020**, *10* (1), 528–542.

(97) Saadun, A. J.; Kaiser, S. K.; Ruiz-Ferrando, A.; Pablo-García, S.; Büchele, S.; Fako, E.; López, N.; Pérez-Ramírez, J. Nuclearity and Host Effects of Carbon-Supported Platinum Catalysts for Dibromomethane Hydrodebromination. *Small* **2021**, *17* (16), 2005234.

(98) Zichittella, G.; Pérez-Ramírez, J. Status and Prospects of the Decentralised Valorisation of Natural Gas into Energy and Energy Carriers. *Chem. Soc. Rev.* **2021**, *50* (5), 2984–3012.

(99) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6* (1), 15–50.

(100) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(101) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(102) Almora-Barrios, N.; Carchini, G.; Błoński, P.; López, N. Costless Derivation of Dispersion Coefficients for Metal Surfaces. *J. Chem. Theory Comput.* **2014**, *10* (11), 5002–5009.

(103) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953–17979.

(104) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **1976**, *13* (12), 5188–5192.

(105) Neugebauer, J.; Scheffler, M. Adsorbate-Substrate and Adsorbate-Adsorbate Interactions of Na and K Adlayers on Al(111). *Phys. Rev. B* **1992**, *46* (24), 16067.

(106) Burcat, A.; Ruscic, B.; Chemistry. Third Millenium Ideal Gas and Condensed Phase Thermochemical Database for Combustion (with Update from Active Thermochemical Tables). **2005**.

(107) Anderson, J. R.; Boudart, M. *Catalysis*; Anderson, J. R., Boudart, M., Eds.; CATALYSIS—Science and Technology; Springer Berlin Heidelberg: Berlin, Heidelberg, 1996; Vol. 10.

(108) Mears, D. E. Diagnostic Criteria for Heat Transport Limitations in Fixed Bed Reactors. *J. Catal.* **1971**, *20* (2), 127–131.

(109) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.;

Dubourg, V.; Passos, A.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(110) The robustnes measures the sensitivity and the precision of the descriptor. The robustness parameter is calculated the standard error of the descriptor prediction, and gives information on its variability.

(111) Roman, T.; Groß, A. Periodic Density-Functional Calculations on Work-Function Change Induced by Adsorption of Halogens on Cu(111). *Phys. Rev. Lett.* **2013**, *110* (15), 156804.

(112) Roman, T.; Gossenberger, F.; Forster-Tonigold, K.; Groß, A. Halide Adsorption on Close-Packed Metal Electrodes. *Phys. Chem. Chem. Phys.* **2014**, *16* (27), 13630–13634.

(113) Regions over 100% in **Figure 4** are unexplored zones far from the input points. Furthermore, there is no halogen-metal in our dataset (including F and I) contained in the over 100% domain.

(114) Guthrie, J. P. Multidimensional Marcus Theory: An Analysis of Concerted Reactions. *J. Am. Chem. Soc.* **1996**, *118* (51), 12878–12885.

(115) The cosine term indicates that our predictions for S_{coke} is overfitted due to the difference between Ru-Br and Ru-Cl values, as the RF and GR predictions. The only way to avoid overfitting is increasing the sample size, which is not possible due to the limited points in the Periodic Table.

ARTICLE

Mechanistic Routes Toward C_3 Products in Copper-Catalysed CO_2 Electroreduction

Sergio Pablo-García[a,†], Florentine L. P. Veenstra[b,†], Louisa Rui Lin Ting[c,d,†], Rodrigo García-Muelas[a], Antonio J. Martín[b], Federico Dattila[a], Boon Siang Yeo[c,d,*], Javier Pérez-Ramírez[b,*], Núria López[a,*]

The electrocatalytic CO_2 reduction (eCO_2R) reaction powered by renewable electricity holds promise for the sustainable production of multi-carbon chemicals and fuels. On Cu-based catalysts, ethylene and ethanol (C_2) have been produced in appreciable amounts. However, C_3 products (mostly terminal oxygenates) have limited yields, whereas propylene, is puzzlingly absent. Herein, we devise a *divide-and-conquer* strategy to explain the formation of the C_3-backbone and elucidate the mechanisms responsible for the observed selectivity by combining network graphs, Density Functional Theory, and experiments to prune the network and benchmark the identified path. Our approach concludes that the most frequently reported products, propionaldehyde and 1-propanol, originate from the coupling of CH_2CH with C(H)O. While propylene and 1-propanol share common intermediates, the former is barely produced due to the unfavourable formation of allyl alkoxy (CH_2CHCH_2O), whose precursor nature was confirmed experimentally. This work paves the way for tailoring selective routes towards C_3 products *via* eCO_2R.

Introduction

Developing functional catalysts for the electrochemical CO_2 reduction (eCO_2R) to complex products lies at the core of new efforts to develop sustainable technologies[1]. Among available materials, copper-based electrocatalysts occupy a pivotal role due to their ability to form the C_{2+} backbone for high-value fuels and commodity chemicals[1,2]. The type and amount of products formed are sensitive to the applied potential, electrolyte, and the preparation protocol of Cu[3–6]. The established mechanism to the C_2 fraction advocates that CO_2 first reduces to CO, which dimerises to $OCCO^-$ and subsequently reduces to hydrocarbons and alcohols. Typically, the main C_2 product is ethylene (up to *ca*. 74% Faradaic efficiency *FE*),[7] although an exceptional ethanol selectivity (*ca*. 91% *FE*)[8] has been reported on Cu clusters. For C_4 products, the aldol condensation of acetaldehyde (C_2) gives crotonaldehyde, which reduces to 1-butanol, albeit with low yields[9]. Among C_3 compounds, 1-propanol can be produced with appreciable yields (~23% *FE*),[10,11] whereas propylene (the corresponding C_3 olefin; 0.36 eV less stable than 1-propanol, **Table S1**) has only been detected as a trace product (<0.1% *FE*)[9]. This puzzling outcome contrasts with the vast formation of ethylene, which is less stable than ethanol by 0.47 eV, **Table S1**. Furthermore, 2-propanol, which is the most thermodynamically stable C_3 alcohol (0.17 eV lower than 1-propanol, **Table S1**), has never been observed in eCO_2R.[12] Reports have indicated that the formation of the C_3 backbone at high CO concentrations and relatively mild applied potentials (–0.36 to –0.56 V *vs.* RHE)[11,13] requires asymmetric sites on the oxide-derived Cu (OD-Cu) catalyst. Nonetheless, the mechanistic understanding of the reasons behind the low formation of C_3 products in eCO_2R at a molecular level is very limited, due to the large number of elementary steps (>10^3) that prevent the use of standard reaction sampling tools based on explicit Density Functional Theory (DFT) and reaction profile analysis.

Herein, we analyse electrocatalytic routes towards C_3 products through a *divide-and-conquer* strategy based on the generation of the network graph, computational reaction profiles combined with key electrochemical experiments involving C_1, C_2, and C_3 reagents. This new methodological approach enables us to (i) identify the most likely C_1–C_2 coupling steps towards C_3 intermediates; (ii) elucidate the bifurcation points to different C_3 products; and (iii) pinpoint kinetic bottlenecks, by propylene.

Experimental and Computational Methods

The electrocatalytic reactor used for our experiments is a gas-tight cell consisting of two compartments separated by a Nafion 211 membrane with gas-flow inlet and outlet ports. The cell has an OD-Cu working electrode, a gas diffusion layer (GDL) carbon paper counter electrode, and a leak-free Ag/AgCl (3 M KCl) reference electrode. Triplicate measurements were done, with

[a] *Institute of Chemical Research of Catalonia, ICIQ, The Barcelona Institute of Science and Technology. Av. Països Catalans 16, 43007 Tarragona, Spain.*
[b] *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zürich, Vladimir-Prelog-Weg 1, 8093 Zürich, Switzerland.*
[c] *Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543.*
[d] *Solar Energy Research Institute of Singapore, National University of Singapore, 7 Engineering Drive 1, Singapore 117574.*
† *These authors contributed equally to this work.*
**Corresponding authors: chmyeos@nus.edu.sg; jpr@chem.ethz.ch; nlopez@iciq.es*
Electronic Supplementary Information (ESI) available: Experimental and computational details, supplementary notes, figures, and tables. See

ARTICLE

Table 1. Summary of C_3 products and their corresponding formation rates (in parenthesis) observed experimentally from the electrolysis of C_1 or a mixture of C_1 and C_2 compounds on OD-Cu. The C_1 and C_2 compounds used are listed in the topmost row and leftmost column respectively, while the experimental conditions are indicated in the footnotes. A colour code has been added to highlight the more favourable experiments for C_3 formation. For reference, the formation rates of C_3 products from eCO_2R are given in parenthesis, coloured scale is indicated in the lower bar. The full set of experiments and product distributions are shown in **Tables S2-S14**.

C_2 \ C_1	Carbon monoxide, CO	Formaldehyde, CH_2O	Methanol, CH_3OH	Carbon dioxide, CO_2
No C_2	1-Propanol (0.02)[A] 1-Propanol (2.2)[B] Allyl alcohol (2.1)[B] Propionaldehyde (0.4)[B]	No C_3[A,B]	No C_3[A]	1-Propanol (75.8)[C] Allyl alcohol (25.2)[C] Propionaldehyde (20.3)[C] Acetone (3.2)[C]
Oxalate, C_2O_4	1-Propanol (0.4)[A]	No C_3[A]	1-Propanol (0.1)[A]	X
Glyoxal, CHOCHO	1-Propanol (0.3)[A]	No C_3[A]	1-Propanol (0.1)[A]	X
Acetate, CH_3COO^-	1-Propanol (0.2)[A]	No C_3[A]	No C_3[A]	X
Ethylene glycol, CH_2OHCH_2OH	1-Propanol (0.5)[A]	No C_3[A]	No C_3[A]	X
Ethanol, CH_3CH_2OH	1-Propanol (0.2)[A]	No C_3[A]	No C_3[A]	X
Acetaldehyde, CH_3CHO	1-Propanol (12.1)[B] Propylene (trace)[B]	Propylene (1.4)[B] Allyl alcohol (2.0)[B]	X	X

[A] 0.1 M KOH at –0.40 V *vs.* RHE
[B] 0.1 M PPB at –1.00 V *vs.* RHE
[C] 0.1 M $KHCO_3$ at –0.95 V *vs.* RHE
[X] Not performed

lower — rate / μmol cm^{-2} h^{-1} — higher

the average values and standard deviations presented in the ESI. Extended details about reagents, catalysts preparation, electrochemical measurements, detection limits, and product analyses are shown in **Experimental and Computational Details** in the ESI. The OD-Cu catalyst was obtained from CuO (see **Figure S1** for XRD analysis). The physicochemical and catalytic properties of this material have been discussed elsewhere[9,14].
We initially reduced CO_2 in 0.1 M $KHCO_3$ at –0.95 V *vs.* RHE to maximise the production of multi-carbon products[14], as shown in **Figure 1** (see **Figure S2** for polarisation curves).

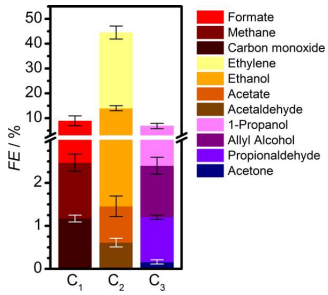


Figure 1. C_1-C_3 products formed from the electrocatalytic CO_2 reduction on oxide-derived copper in 0.1 M $KHCO_3$ at –0.95 V *vs.* RHE. Detailed results are presented in **Table S2**. Hydrogen *FE* is 37.4%.

C_1 and C_2 products account for 53% of the *FE*, whereas 7% corresponds to C_3 products and the balance is H_2. Our observed product distribution agrees with selectivity trends presented in the literature, which are summarised in **Figure S3**, and the paths to C_2 products are shown in **Figures S4–S5**. To unravel the selectivity patterns observed from both the literature and our experiments, our workflow entails: (i) building the reaction network by encoding the corresponding structural graphs; (ii) sampling the intermediates by DFT; (iii) computing all C_1–C_2 backbone couplings by DFT; (iv) pruning the network of non-viable backbone formation routes by probing the products from CO, formaldehyde, and methanol co-reduction with C_2 reactants (**Table 1**), with particular attention to missing products; (v) computing all routes from the C_3 backbone to the final products using DFT and Linear-Scaling Relationships (**Table S15**) to identify the best routes towards propanol and propylene; (vi) experimental benchmarking of the main predicted routes via electrocatalytic tests with key intermediates.
While the routes to C_1 and C_2 products can be probed manually, as shown in the literature[15–20], the analysis of routes to C_3 products demands automation. The full network containing all C_1, C_2, and C_3 intermediates contains 463 elements, represented here as nodes in a graph (see "Graph representation of the reaction network" in **Experimental and Computational Details**). The energies of intermediates (referenced to CO_2, H_2O, and H_2) were evaluated with the computational hydrogen electrode (CHE)[21,22] containing the DFT energy obtained with a PBE-D2 formulation[23–25] (corrected for metal overbinding), our in-house developed implicit solvation model[26,27], and the polarisation term[9,28] (see **Eq. S1-S4** in the **Experimental and Computational Details**). The contribution of the D2 of H* and CO* adsorption is small: 0.04 and 0.14 eV, respectively. Intermediates are linked by 2266 steps (edges linking the nodes in the graph): 55 C_1–C_1 and 636 C_1–C_2 couplings, 683 C–H and 305 O–H hydrogenations,

and 587 C–O(H) cleavages. To ensure the desired accuracy, 586 out of all C_1–C_2 couplings (**Tables S16-S19**, 92% of total), 10 C–O(H) breakings (**Table S20**), and 8 hydrogenations (**Table S21**) were explicitly obtained via Nudged Elastic Band (DFT-NEB)[29] and confirmed by vibrational analysis. Initial guesses for NEB were generated automatically (**Experimental and Computational Details in Electronic Supplementary Information** and **Notes S1-S4**). Linear Scaling Relationships (LSR) were employed in the initial fast-sampling of C–H and O–H hydrogenations as they are reliable for these cases[20,30,31] (**Table S15**). The transition states for key hydrogenation steps in the main path were further refined with Nudged Elastic Band (DFT-NEB). Heyrovsky-type reactions for C–OH breakings and C–H formations were considered (**Note S5**). Tests on density functionals, LSR, and charge analysis are described in **Note S6** and **Figures S6-S9,** and demonstrate that our strategy provides an excellent cost-efficiency balance.

Benchmark electrochemical experiments (**Table 1**) involving the reduction of selected C_1 and C_2 compounds and their mixtures were conducted at mild overpotentials (–0.4 V *vs*. RHE) in alkaline pH, where the production of multi-carbon products is expected to be boosted.[32,33] In the case of aldehydes, which undergo side reactions in alkaline media[9], electrolysis was performed in neutral potassium phosphate buffer (PPB) at −1.0 V *vs.* RHE as the optimum condition for the production of propylene (**Table S3**). To avoid interference of the parasitic hydrogen evolution reaction (HER) in assessing reactivity, we compared production rates, instead of Faradaic efficiencies, of the carbonaceous products formed under different conditions. Additional information on the experimental conditions can be found in **Note S7.**

Results and Discussion

Routes to main C_1 and C_2 products and precursors

The main outcome of previous computational studies on the manually-analysed network for eCO_2R to C_1 and C_2 products over Cu(111) or Cu(100) can be summarised as follows:[15–20] CO_2 adsorbs on the metal and, through two proton-coupled electron transfers, reduces to CO, a precursor for methanol and methane (C_1). Alternatively, CO dimerises to $OCCO^-$, which has been claimed to be the rate-determining step in the formation of the C_2 backbone[15–19]. On Cu(100), this process is endergonic by 0.84 eV at –0.4 V *vs*. RHE, (with an activation barrier of 1.42 eV), but can be promoted on defective surfaces[34]. Additionally, $OCCO^-$ can be stabilised by electrolyte cations[35,36] and the electric field at the cathode[28,36,37]. $OCCO^-$ is further protonated and then dehydrated to CCO upon a proton-coupled electron transfer step. The latter intermediate is subsequently hydrogenated to CH_2CHO[1,38]. If hydrogenolysis occurs on the remaining O atom, CH_2CH is formed, leading to ethylene. If hydrogenation occurs on the α-C ($C_\alpha H_2CHO$), ethanol is formed (**Figures S4-S5**).

Exploring C_1–C_2 coupling reactions

To unravel the most likely reaction routes towards C_3 backbone formation, a sequential analysis combining available reports, experiments and theory was applied. The possible C_1–C_2 couplings shown in **Figure 2a** and **Tables S16-S18** emerge from the combination of 10 C_1 (–CH_xO* or –CH_x*) and 70 C_2 precursors that can be hydrocarbons (CH_yCH_z*–) or oxygenates (CH_yOCH_zO*–, CH_yOCH_z*–, CH_yCH_2O*–). This gives rise to oxygenate-to-oxygenate, hydrocarbon-to-oxygenate, and hydrocarbon-to-hydrocarbon C–C coupling reactions. Thus, six families of intermediates, namely C_3H_x, 1-C_3OH_x, 1,3-$C_3O_2H_x$, 2-C_3OH_x, 1,2-$C_3O_2H_x$ and 1,2,3-$C_3O_3H_x$, x = 0-8 (families **1**-**6** in **Figure 2a**) are formed, where the prefix numbers show the position of the oxygenated functional group, irrespective of it

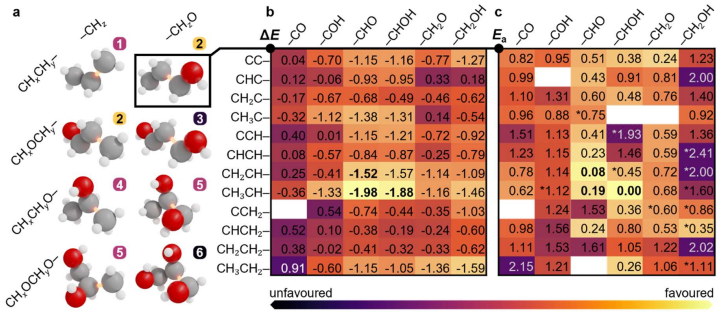


Figure 2. Screening process to narrow down the C_1–C_2 coupling steps. **a** $C_xH_yO_z$ backbones that can be obtained through C_1–C_2 couplings. Only the fully hydrogenated product is shown with the bond formed marked in orange. The numbers label six families of molecules and the colour stands for their abundance found in CO_2 reduction experiments, found either from in literature or this work. High, low, or zero relative abundance is shown in orange, purple, and black, respectively. Backbones 3 and 5 can be formed from two combinations of C_1 and C_2 intermediates. **b–c** Reaction and activation energies for CH_xCH_y–CH_zO couplings between a C_1 oxygenate and a C_2 hydrocarbon, ΔE and E_a in **Eq. S3S5-S4S6**. Most likely steps in bold. The full set of 10 C_1 × 70 C_2 coupling reactions is shown in **Tables S16-S18**. Further details can be found in **Notes S1-S2**, **S8-S9**.

having alcohol, alkoxy, aldehyde, or ketone character. Molecular fragments with carboxylate, carboxylic acid, ethers, or cyclic backbones were not considered, as these functionalities have not been found experimentally in the pool of C_3 products.

We then compared the C_3 products formed from the electrolysis of mixtures of CO with different C_2 molecules (glyoxal, ethylene glycol, oxalate, acetate, ethanol) at open-circuit potential and –0.4 and –1.0 V *vs.* RHE (**Tables 1**, **S6, S8-S9**). At –0.4 V *vs.* RHE, all these mixtures generate 1-propanol at rates much larger than the reduction of CO alone (**Table S4**). Mixtures with ethylene glycol gave the highest yield, while those with ethanol and acetate gave the lowest. Propylene was not detected (detection limits of gaseous products are equivalent to 0.5 μmol cm^{-2} h^{-1}, see **ESI Experimental and Computational Details**). Overall, if a set of products with a given C_xO_y backbone is not observed experimentally, then such couplings can be considered unlikely, and the routes pruned from the network.

To further verify the nature of the active C_1 fragment leading to 1-propanol, electrocatalytic tests of the C_2 compounds with either methanol or formaldehyde (**Error! Reference source not found.**) were conducted. 1-Propanol was not detected in some of these experiments, though we observed allyl alcohol and propylene from the electrolysis of a formaldehyde and acetaldehyde mixture. Moreover, the reduction of CH_2O itself produced only CH_3OH and CH_4 (**Tables 1**, **S4**), but it does not produce C_2 and C_3 compoundsas it is hardly broken into the more reactive CH2 and CHO species (**Table** , **S22**). CH_3OH, on the other hand, was electrochemically inert (**Tables S12–S13**). The unique predominance of 1–propanol in experiments using CO indicates that *CO (or a derivative like *CHO) is instrumental in promoting 1-propanol formation. This is further confirmed by the absence of 1-propanol in experiments starting with CH_2O or CH_3OH (**Tables 1**).

After considering the experimental input, we then switched to theory to explore the C_1–C_2 coupling reactions based on the reaction energies (ΔE, **Table S16**), activation energies obtained by DFT-NEB (E_a, **Table S17**), and complemented by the electrochemical driving force computed as the polarisation variation upon reaction ($\Delta\Delta G_B$ in **Eq. S6**, **Table S18**). More favourable values are shown in brighter colours in **Figure 2b-c**. The most likely candidates were then selected among all couplings, which reduced the set to CH_2CH–CHO and CH_3CH–CHO. In the following paragraphs, we describe how the different coupling families are retained or discarded during the analysis of the network based on abovementioned literature, experimental, and theoretical analyses :

1,2,3-$C_3O_3H_x$ backbone Early computational studies proposed that the C_3 backbone was formed via trimerisation of *CO[39,40] (family **6** in **Figure 2a**, 1,2,3-$C_3O_3H_x$). We computed this potential reaction and found a relatively high activation barrier of E_a = 0.96 eV. Furthermore, the CO-trimer reverts 0.14 e^- to the surface, so the net reaction is therefore expected to be hindered under reductive potentials (**Table S19**). Alternatively, a sequential process can be envisaged where CO dimerises to $OCCO^-$, which further reacts with CHO (E_a = 0.73 eV, $\Delta\Delta G_B$ = –0.81 e^-), or COCHO with CO (E_a = 0.78 eV, $\Delta\Delta G_B$ = –0.39 e^-) to form COCOCHO as the base of the 1,2,3-$C_3O_3H_x$ backbone. Only the latter reaction would be promoted at more reducing potentials. However, should this reaction occur, glycerol would likely appear as a product of CO_2 reduction, but this has not been reported in the literature. Thus, the absence of glycerol as a product combined with the medium to high computed barriers and electrochemical penalties suggest that 1,2,3-$C_3O_3H_x$ (**6** in **Figure 2a**) intermediates are unlikely to participate in the main mechanistic route.

1,2-$C_3O_2H_x$ and 1,3-$C_3O_2H_x$ backbones There are two families of $C_3O_2H_x$ intermediates, with O atoms in different positions: 1,2-$C_3O_2H_x$ (**5**) and 1,3-$C_3O_2H_x$ (**3**, **Figure 2a**). Among the products derived from 1,2-$C_3O_2H_x$, only 1-hydroxyacetone has been reported in the literature, albeit in trace quantities.[2] Indeed, some CH_xCH_yO*–CH_2O* pairs have low coupling barriers, such as the coupling of CH_2CO, CH_3CO, and CH_3COH with CHO (up to 0.32 eV, **Table S17**). Remarkably, the HO coupling is expected to be strongly promoted under reductive potentials ($\Delta\Delta G_B$ = –0.52 |e^-|, **Table S18**). However, these C_2 intermediates have higher potential energies than other structural isomers (**Figure S5**). As such, their concentration is expected to be too low at –0.4 V *vs.* RHE to form any significant amount of 1,2-$C_3O_2H_x$ products, as confirmed by the range of products observed in our experiments (**Tables S8-S13**). The formation of 1,3-$C_3O_2H_x$ products is expected to proceed from CH_2OCH–CHO(H) ($E_a \leq$ 0.34 eV) or CH_2OHCH–CHO (E_a < 0.16 eV). These reactions would occur as chemical steps and are not favoured under eCO_2R conditions (according to computational charge considerations, **Table S18**). Our simulation results may also explain why 1,3-$C_3O_2H_x$ products (*viz.* 1,3-propanediol and 3-hydroxypropanal) have not been experimentally observed for Cu-based catalysts.

2-C_3OH_x backbone via H_xC_1-C_2H_yO coupling We are now left with the paths that generate mono-oxygenated C_3 intermediates. 2-C_3OH_x products can be produced from the coupling of a CH_xCH_yO*– fragment, such as CH_3CHO*– and CH_2CHO*–, with a C_1 hydrocarbon, –CH_x* (family **4** in **Figure 2a**, **Tables S16-S18**). This pathway may expectedly yield 2-propanol, which is 0.17 eV more stable than 1-propanol (**Table S1**), whereas intermediates leading to these two species show similar stabilities (**Figure S10**). However, 2-propanol was not experimentally detected (**Error! Reference source not found.**). Previous experiments on Cu-based catalysts have only detected small amounts of acetone[2], in line with our present results (0.2% *FE*, **Figure 1**, **Table S2**). Acetone is likely produced by coupling CH_3CO with CH_2, (ΔE = –1.57 eV; E_a = 0.28 eV) and the further hydrogenation of the unsaturated aliphatic carbon atom.

1-C_3OH_x backbone Most of the C_3 products detected in our experiments belong to the 1-C_3OH_x family (**2** in **Figure 2a**), namely 1-propanol ($CH_3CH_2CH_2OH$, 1-C_3OH_8), with propionaldehyde and allyl alcohol (CH_3CH_2CHO and CH_2CHCH_2OH, 1-C_3OH_6) produced at smaller rates (**Table 1**).[2,41] Considering the experimentally observed scarcity of $C_3O_2H_x$ and $C_3O_3H_x$ products, we infer that there is only one oxygen atom present during the coupling, either on the C_1 or the C_2 moiety. Reported experiments[42–45] show that during eCO_2R, the

maximum production of 1-propanol occurs when high amounts of CO and C_2H_4 are formed simultaneously. Indeed, the lowest activation barriers are found for the highly exothermic $CH_2CH–CHO$ and $CH_3CH–CHO(H)$ couplings ($\Delta E < -1.50$ eV, $E_a \leq 0.19$ eV, **Figure 2 b-c**). As CH_2CH is a precursor of C_2H_4, and CHO(H) is directly formed from CO, we conclude that all such paths are highly likely. Couplings involving C_2 moieties *less hydrogenated* than CH_2CH or CH_2CHO (another C_2H_4 precursor) are therefore less likely. In the remaining region, most $C_1–C_2$ couplings are highly activated (**Figure 2c**). Thus, CH_2CHCHO^* and CH_3CHCHO^* intermediates are common precursors for C_3 products. To a lesser degree, CH_2CHCO^* can also be formed if the coupling starts with –CO instead of –CHO, **Figure 3**. Finally, reactions involving C_2 oxygenated precursors (**Tables S16-S18**) have higher barriers, such as the $CH_2–CCH_2O$ coupling (while the process is exothermic, $\Delta E = -1.61$ eV, it exhibits a non negligible activation barrier, $E_a = 0.39$ eV). Upon reaction, part of the electronic density of CH_2CHCHO^* is returned to the surface (–0.30 e^-, **Table S18**). The reaction is therefore unfavoured at strongly reductive potentials, which explains the decrease in 1-propanol production as the potential becomes more negative[1,2].

The activation barriers of the transition states associated with

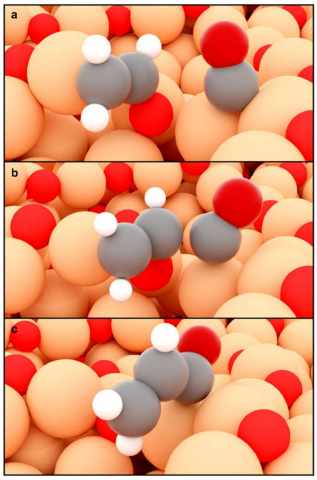


Figure 3. CH_2CHO^*CO coupling on OD-Cu models[34]. **a-c** Initial, transition, and final states respectively. A surface cavity with high oxygen affinity assists the C–O bond breaking of the CH_2CHO^* precursor. A neighbouring polarised site, $Cu^{\delta+}$, weakly adsorbs CO.

the formation of key C_3 intermediates are sensitive to surface geometry and ensembles. Since defective Cu surfaces have been reported selective to propanol formation[39,44], we assessed the role of defects on OD-Cu models[34] for the concerted coupling of $CH_2CH(O)^*–CO^*$ to yield the simplest C_3 precursor, CH_2CHCO, **Figure 3**. The CuO_x structures were optimised for 10 ps through *ab initio* molecular dynamics and recurrent morphological motifs occured upon surface reconstruction[34]. Out of twelve surface motifs assessed (**Figure S11**)[34,46], an active site composing of a surface cavity (**Figure 3a**, centre) paired with a neighbouring $Cu^{\delta+}$ site (**Figure 3a**, right) is the most suitable for promoting the coupling. While CH_2CHO^* is trapped at the surface cavity, the high oxygen affinity of this site leads to the breaking of its C–O bond to give CH_2CH^* (**Figure 3b**). On the other hand, CO adsorption is almost thermoneutral on the polarised copper site. Thus, the CH_2CH^* fragment can easily couple to the weakly bound CO* to form the C_3 backbone (exergonic by 0.13 eV; **Figure 3c**). In absence of polarized Cu sites (**Figure S11**), this step is endergonic by at least 0.6 eV, thus confirming the instrumental role of surface polarization. This coupling mechanism may explain the high selectivity toward 1-propanol (FE = 23% at –0.44 V vs. RHE) achieved on highly defective Cu surfaces containing a large number of surface cavities[11,39,44]. This concept can also be extended to other key $C_1–C_2$ coupling reactions from moieties directly derived from CH_2CHO and CO, such as $CH_2CH–CHO$ and $CH_3CH–CHO$. Going beyond pure copper catalysts, we propose that intermetallic alloys containing high oxygen affinity elements coupled with weak CO binding sites could be highly selective to C_3 as well. For instance, Cu-Ag alloys exhibited enhanced propanol selectivity depending on the silver atomic ratio[13,47], suggesting a CO spillover mechanism[48] from Ag domains to facilitate the formation of $CH_2CH–CO$.

Routes to C_3 products.

Routes to 1-propanol Once the C_3 backbone is formed, the C_3 subnetwork (**Figure 4**) starting from CH_2CHCHO^* (orange) and CH_2CHCO^* (black) can be employed to analyse selectivity trends. The colour code of the boxes in **Figure 4** represents the computed relative stability of the intermediates (thermodynamics), while the thickness of the lines linking the intermediates account for the barriers (thicker lines stand for faster steps). The hydrogenation of CH_2CHCO^* gives CH_2CHCHO^*, which then evolves via $CH_3CHCHO^* \rightarrow$ propionaldehyde ($CH_3CH_2CHO^*$) $\rightarrow$ propanoxy ($CH_3CH_2CH_2O^*$) $\rightarrow$ 1-propanol. The existence of this path is confirmed experimentally, since the electrochemical reduction of propionaldehyde on OD-Cu yielded predominantly 1-propanol (**Figure 5c, Table S23**). Alternatively, the 1-propanol formation can proceed through $CH_2CH_2CO^* \rightarrow CH_2CH_2CO^*H \rightarrow CH_3CH_2COH^* \rightarrow CH_3CH_2CHOH^* \rightarrow$ 1-propanol (bottom path in **Figure 4**, and **Figure 5a**).

Routes to propylene Mono-oxygenates can be converted to propylene *via* dehydration reactions starting from $CH_2CHCO(H)^*$, $CH_2CHCHO(H)^*$, $CH_3CHCHO(H)^*$, $CH_2CHCH_2O(H)^*$, and $CH_3CHCH_2O(H)^*$, where (H) represents an optional hydrogen. The corresponding barriers of these ten reactions were computed (**Table S20**). Most C–O(H) bonds are relatively difficult to activate ($E_a > 1.0$ eV), thus, we depict the ones showing relatively lower barriers ($CH_2CHCH_2–OH^*$ and $CH_3CHCH_2–OH^*$, $E_a = 0.17$ and 0.94 eV, respectively) in

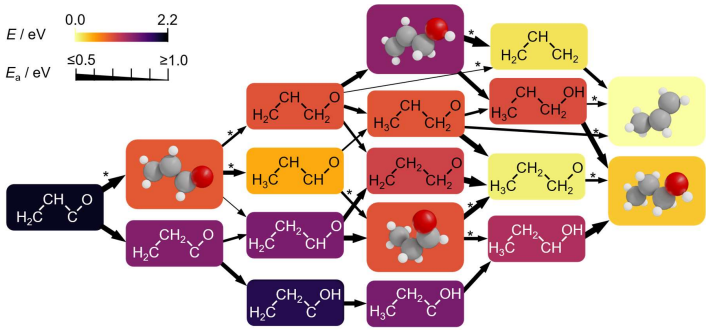


Figure 4. Computed subnetwork for CH_2CHCO and CH_2CHCHO conversion to propylene (C_3H_6) and 1-propanol (C_3H_7OH) at –0.4 V *vs.* RHE (full network in **Figure S12-S14**). The colors of the boxes scale with the relative DFT energy of their intermediates (**Eq. S3**). Relevant intermediates that can desorb and be used as probe molecules are drawn in 3*D*. We used allyl alcohol (C_3H_5OH) and propionaldehyde (C_3H_6O) as reactants in our experiments to confirm the pathways predicted by the network (**Tables S23–S24**). The thickness of the arrows connecting the intermediates account for the activation energies (E_a, obtained by LSR. Those obtained explicitly by DFT are denoted by *).

Figure 5a-b. A selectivity switch to propylene occurs when the aldehyde carbon on CH_2CHCHO* is hydrogenated to form CH_2CHCH_2O*, which in turn produces allyl alcohol (CH_2CHCH_2OH). To generate propylene, OH is eliminated from the allyl alcohol intermediate, which is then hydrogenated (**Figure 5b**). However, this path is not fully selective, as allyl alcohol can also undergo hydrogenation to CH_3CHCH_2OH* to form 1-propanol. The C–OH bond breaking in allyl alcohol (CH_2CHCH_2OH) has a low barrier of 0.17 eV and it is strongly promoted by reducing potentials, with a net charge gain of 0.87 e^- (**Table S20** and **Figure 5b**).

Therefore, the production of propylene could be traced to the allyl alkoxy (CH_2CHCH_2O) intermediate, which is also a direct precursor of allyl alcohol. This proposition was verified experimentally by reducing allyl alcohol on OD-Cu (**Figure 5d**), which gave noticeable amounts of propylene as theoretically predicted. Moreover, allyl alcohol (1.97 μmol cm^{-2} h^{-1}) was detected alongside propylene (1.44 μmol cm^{-2} h^{-1}) from the reduction of a mixture of acetaldehyde and formaldehyde (**Table 1**). In this combination, the most likely path occurs when acetaldehyde loses an acidic α-hydrogen (H_α–CH_2CHO)[9] to form CH_2CHO, which dehydrates to form CH_2CH. The latter compound reacts with CH_2O to form CH_2CHCH_2O* (**Figure 4**, **S12-S14**) which is mainly selective towards allyl alcohol and propylene, but not 1-propanol (**Table 1**). Interestingly, we note that CO_2 reduction produced 1-propanol (*FE* = 4.4%) and allyl alcohol (*FE* = 1.2%) (**Figure 1**), while propylene was absent. This can be rationalised by a mild eCO_2R interface alkalinisation, which occurs under reaction conditions,[45] favouring the desorption of allyl alkoxy (protonated in solution into allyl alcohol) and thus preventing propylene synthesis. Overall, these observations strongly suggest the key role of allyl alcohol in the route to propylene.

From a broader perspective, the low activity of Cu catalysts for eCO_2R to C_3 compounds, particularly propylene, could be improved through engineering at different scales. Currently, the most explored approaches to promote multi-carbon products include engineering catalyst surfaces with a high density of defects to improve activity, and optimizing the electrolyte and reactor conditions to alleviate mass transport limitations and tuning the environment at the electrode-electrolyte interface. Modifications at the process level could benefit from three different approaches: (i) one single reactor recycling C_2 (or C_1) eCO_2R products to ensure a high concentration of active intermediates; (ii) independently optimised catalysts and reactors to produce C_1 (Cu or Ag-based catalysts) and C_2 (on an oxide-derived Cu catalyst) intermediates, which mix in a third reactor dedicated to the coupling to form the C_3 backbone or alloys containing close domains of both; (iii) a process able to produce the relevant intermediate allyl alcohol (for which an effective catalyst is not yet known) that is then converted to propylene in a second unit. Although there are some experimental indications in the literature of the potential of strategies (i) and (ii)[13], the detailed understanding of the reaction network and the elucidation of key intermediates presented in this work ultimately serve to direct future works toward realizing these solutions.

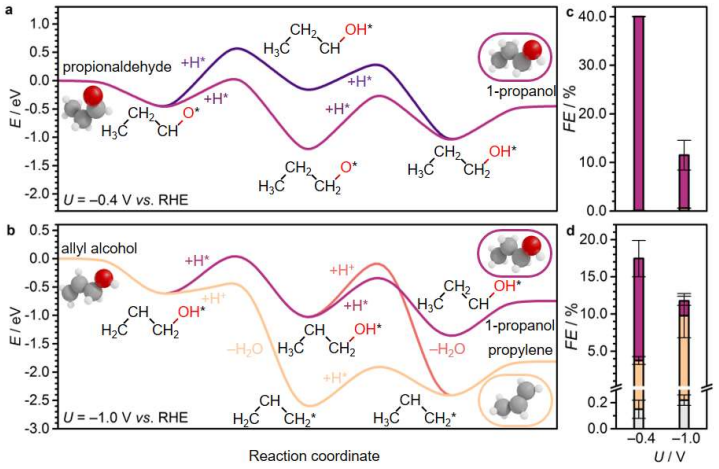


Figure 5. **a-b** Energy profiles for electrocatalytic reduction of key C_3 compounds on Cu(100), using H_2, CO_2, and H_2O as thermodynamic sinks, and shifting the energy reference to make **a** propionaldehyde and **b** allyl alcohol the zero. Corresponding products for experimental electroreduction of **c** propionaldehyde and **d** allyl alcohol at –0.4 V and –1.0 V *vs.* RHE on OD-Cu: propane (grey), propylene (orange), and 1-propanol (purple). Full product distributions are shown in **Tables S23-S24**. Other energy profiles at 0.0 V, –0.4 V, and –1.0 V *vs.* RHE are shown in **Figures S15-S17**. The (*) symbol refers to species adsorbed on the surface. Detailed DFT values can be found in **Tables S20–S21**.

Conclusions

In conclusion, we have performed an integrated mechanistic analysis of the eCO_2R to C_3 products. Methodological implementations including structural graph network generation, fast energy screenings, and network pruning of irrelevant paths through experimental input allow the effective sampling of the complex C_3 network. C_2 and C_3 products were found to share a common precursor, CH_2CHO*. Our findings rationalise the generally observed low selectivity of eCO_2R toward C_3 products, as well as their enhancement on nanostructured Cu catalysts: (i) C_3 backbones are formed *via* the sluggish coupling of CO or CHO with CH_2CH*, preferentially at defects and (ii) all C_3 precursors end up containing at least one O atom, *i.e.*, CH_x+C_2H_y couplings are highly unlikely. The most stable mono-oxygenated intermediate CH_2CHCHO* gives access to propylene, propionaldehyde, and 1-propanol. The inaccessible allyl alkoxy intermediate is identified as the most likely kinetic trap preventing propylene production as indicated by simulations and further reinforced with the electrolysis of allyl alcohol leading to propylene. Our mechanistic understanding paves the way towards the development of advanced electrocatalysts that promote C_3 products, particularly alkenes.

Author Contributions

S.P.-G., Software, Methodology, Data Curation, Investigation, Visualisation, Writing – original draft. F.L.P.V., L.R.L.T., Methodology, Data Curation, Investigation, Visualisation, Writing – original draft. R.G.-M., A.J.M., Methodology, Data curation, Investigation, Visualisation, Validation, Supervision, Writing – original draft. F. D., Data Curation, Investigation, Writing – review & editing. B.S.Y., J.P.-R., N.L., Conceptualisation, Funding acquisition, Supervision, Project administration, Writing – review & editing.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

This work was financed by the Spanish Ministry of Science and Innovation (RTI2018-101394-B-I00, Severo Ochoa CEX2019-000925-S 10.13039/501100011033), ETH Research Grant (ETH-47 19-1), the National University of Singapore Flagship Green Energy Program (R143-000-A64-114, R143-000-A55-733 and R143-000-A55-646), and Ministry of Education of Singapore (R143-000-B52-114). This publication was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation. The

Barcelona Supercomputing Centre – MareNostrum (BSC-RES) is acknowledged for providing generous computational resources. The authors thank Mavis Pei Lin Kang from NUS for assisting with XRD analyses.

Notes and references

1 S. Nitopi, E. Bertheussen, S. B. Scott, X. Liu, A. K. Engstfeld, S. Horch, B. Seger, I. E. L. Stephens, K. Chan, C. Hahn, J. K. Nørskov, T. F. Jaramillo and I. Chorkendorff, *Chem. Rev.*, 2019, **119**, 7610–7672.

2 K. P. Kuhl, E. R. Cave, D. N. Abram and T. F. Jaramillo, *Energy Environ. Sci.*, 2012, **5**, 7050.

3 B. Schmid, C. Reller, S. Neubauer, M. Fleischer, R. Dorta and G. Schmid, *Catalysts*, 2017, **7**, 161.

4 Y. Kwon, Y. Lum, E. L. Clark, J. W. Ager and A. T. Bell, *ChemElectroChem*, 2016, **3**, 1012–1019.

5 X. Wang, A. Xu, F. Li, S. F. Hung, D. H. Nam, C. M. Gabardo, Z. Wang, Y. Xu, A. Ozden, A. S. Rasouli, A. H. Ip, D. Sinton and E. H. Sargent, *J. Am. Chem. Soc.*, 2020, **142**, 3525–3531.

6 Y. Y. Birdja, E. Pérez-Gallent, M. C. Figueiredo, A. J. Göttle, F. Calle-Vallejo and M. T. M. Koper, *Nat. Energy*, 2019, **4**, 732–745.

7 J. Li, Z. Wang, C. McCallum, Y. Xu, F. Li, Y. Wang, C. M. Gabardo, C. T. Dinh, T. T. Zhuang, L. Wang, J. Y. Howe, Y. Ren, E. H. Sargent and D. Sinton, *Nat. Catal.*, 2019, **2**, 1124–1131.

8 H. Xu, D. Rebollar, H. He, L. Chong, Y. Liu, C. Liu, C. J. Sun, T. Li, J. V Muntean, R. E. Winans, D. J. Liu and T. Xu, *Nat. Energy*, 2020, **5**, 623–632.

9 L. R. L. Ting, R. García-Muelas, A. J. Martín, F. L. P. Veenstra, S. T.-J. Chen, Y. Peng, E. Y. X. Per, S. Pablo-García, N. López, J. Pérez-Ramírez and B. S. Yeo, *Angew. Chem. Int. Ed.*, 2020, **59**, 21072–21079.

10 D. Gao, I. Sinev, F. Scholten, R. M. Arán-Ais, N. J. Divins, K. Kvashnina, J. Timoshenko and B. Roldan Cuenya, *Angew. Chem. Int. Ed.*, 2019, **58**, 17047–17053.

11 J. Li, F. Che, Y. Pang, C. Zou, J. Y. Howe, T. Burdyny, J. P. Edwards, Y. Wang, F. Li, Z. Wang, P. De Luna, C. T. Dinh, T. T. Zhuang, M. I. Saidaminov, S. Cheng, T. Wu, Y. Z. Finfrock, L. Ma, S. H. Hsieh, Y. S. Liu, G. A. Botton, W. F. Pong, X. Du, J. Guo, T. K. Sham, E. H. Sargent and D. Sinton, *Nat. Commun.*, 2018, **9**, 4614.

12 L. Mandal, K. R. Yang, M. R. Motapothula, D. Ren, P. Lobaccaro, A. Patra, M. Sherburne, V. S. Batista, B. S. Yeo, J. W. Ager, J. Martin and T. Venkatesan, *ACS Appl. Mater. Interfaces*, 2018, **10**, 8574–8584.

13 X. Wang, Z. Wang, T. T. Zhuang, C. T. Dinh, J. Li, D. H. Nam, F. Li, C. W. Huang, C. S. Tan, Z. Chen, M. Chi, C. M. Gabardo, A. Seifitokaldani, P. Todorović, A. Proppe, Y. Pang, A. R. Kirmani, Y. Wang, A. H. Ip, L. J. Richter, B. Scheffel, A. Xu, S. C. Lo, S. O. Kelley, D. Sinton and E. H. Sargent, *Nat. Commun.*, 2019, **10**, 1–7.

14 D. Ren, J. Fong and B. S. Yeo, *Nat. Commun. 2018 91*, 2018, **9**, 925.

15 A. J. Garza, A. T. Bell and M. Head-Gordon, *ACS Catal.*, 2018, **8**, 1490–1499.

16 J. D. Goodpaster, A. T. Bell and M. Head-Gordon, *J. Phys. Chem. Lett.*, 2016, **7**, 1471–1477.

17 Y. Huang, Y. Chen, T. Cheng, L.-W. Wang and W. A. Goddard, *ACS Energy Lett.*, 2018, **3**, 2983–2988.

18 T. Cheng, H. Xiao and W. A. Goddard, *Proc. Natl. Acad. Sci. U.S.A*, 2017, **114**, 1795–1800.

19 Y. Zheng, A. Vasileff, X. Zhou, Y. Jiao, M. Jaroniec and S.-Z. Qiao, *J. Am. Chem. Soc.*, 2019, **141**, 7646–7659.

20 Q. Li, R. García-Muelas and N. López, *Nat. Commun.*, 2018, **9**, 526.

21 J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard and H. Jónsson, *J. Phys. Chem. B*, 2004, **108**, 17886–17892.

22 A. A. Peterson, F. Abild-Pedersen, F. Studt, J. Rossmeisl and J. K. Nørskov, *Energy Environ. Sci.*, 2010, **3**, 1311–1315.

23 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

24 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.

25 N. Almora-Barrios, G. Carchini, P. Błoński and N. López, *J. Chem. Theory Comput.*, 2014, **10**, 5002–5009.

26 M. Garcia-Ratés and N. López, *J. Chem. Theory Comput.*, 2016, **12**, 1331–1341.

27 M. Garcia-Ratés, R. García-Muelas and N. López, *J. Phys. Chem. C*, 2017, **121**, 13803–13809.

28 R. B. Sandberg, J. H. Montoya, K. Chan and J. K. Nørskov, *Surf. Sci.*, 2016, **654**, 56–62.

29 G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.

30 M. García-Mota, B. Bridier, J. Pérez-Ramírez and N. López, *J. Catal.*, 2010, **273**, 92–102.

31 S. Pablo-García, M. Álvarez-Moreno and N. López, *Int. J. Quantum Chem.*, 2021, **121**, e26382.

32 C. T. Dinh, T. Burdyny, G. Kibria, A. Seifitokaldani, C. M. Gabardo, F. Pelayo García De Arquer, A. Kiani, J. P. Edwards, P. De Luna, O. S. Bushuyev, C. Zou, R. Quintero-Bermudez, Y. Pang, D. Sinton and E. H. Sargent, *Science*, 2018, **360**, 783–787.

33 J.-J. Lv, M. Jouny, W. Luc, W. Zhu, J.-J. Zhu and F. Jiao, *Adv. Mater.*, 2018, **30**, 1803111.

34 F. Dattila, R. García-Muelas and N. López, *ACS Energy Lett.*, 2020, **5**, 3176–3184.

35 E. Pérez-Gallent, G. Marcandalli, M. C. Figueiredo, F. Calle-Vallejo and M. T. M. Koper, *J. Am. Chem. Soc.*, 2017, **139**, 16412–16419.

36 J. Resasco, L. D. Chen, E. Clark, C. Tsai, C. Hahn, T. F. Jaramillo, K. Chan and A. T. Bell, *J. Am. Chem. Soc.*, 2017, **139**, 11277–11287.

37 K. Jiang, R. B. Sandberg, A. J. Akey, X. Liu, D. C. Bell, J. K. Nørskov, K. Chan and H. Wang, *Nat. Catal.*, 2018, **1**, 111–119.

38 R. Kortlever, J. Shen, K. J. P. Schouten, F. Calle-Vallejo and M. T. M. Koper, *J. Phys. Chem. Lett.*, 2015, **6**, 4073–4082.

39 Y. Pang, J. Li, Z. Wang, C. S. Tan, P. L. Hsieh, T. T. Zhuang, Z. Q. Liang, C. Zou, X. Wang, P. De Luna, J. P. Edwards, Y. Xu, F. Li, C. T. Dinh, M. Zhong, Y. Lou, D. Wu, L. J. Chen, E. H. Sargent and D. Sinton, *Nat. Catal.*, 2019, **2**, 251–258.

40 H. Xiao, T. Cheng and W. A. Goddard, *J. Am. Chem. Soc.*, 2017, **139**, 130–136.
41 Y. Hori, A. Murata and R. Takahashi, *J. Chem. Soc. Faraday Trans.*, 1989, **85**, 2309–2326.
42 Y. Hori, I. Takahashi, O. Koga and N. Hoshi, *J. Phys. Chem. B*, 2002, **106**, 15–17.
43 D. Ren, N. T. Wong, A. D. Handoko, Y. Huang and B. S. Yeo, *J. Phys. Chem. Lett.*, 2016, **7**, 20–24.
44 T. Zhuang, Y. Pang, Z. Liang, Z. Wang, Y. Li, C. Tan, J. Li, C. T. Dinh, P. De Luna, P. L. Hsieh, T. Burdyny, H. H. Li, M. Liu, Y. Wang, F. Li, A. Proppe, A. Johnston, D. H. Nam, Z. Y. Wu, Y. R. Zheng, A. H. Ip, H. Tan, L. J. Chen, S. H. Yu, S. O. Kelley, D. Sinton and E. H. Sargent, *Nat. Catal.*, 2018, **1**, 946–951.
45 F. L. P. Veenstra, N. Ackerl, A. J. Martín and J. Pérez-Ramírez, *Chem*, 2020, 1–16.
46 F. Dattila, R. García-Muelas and N. López, *Dataset associated to: Active and Selective Ensembles in Oxide-Derived Copper Catalysts for CO_2 Reduction; DOI 10.19061/iochem-bd-1-165*, 2020.
47 A. Herzog, A. Bergmann, H. S. Jeon, J. Timoshenko, S. Kühl, C. Rettenmaier, M. Luna Lopez, F. T. Haase and B. Roldan Cuenya, *Angew. Chem. Int. Ed.*, 2021, **60**, 7426–7435.
48 P. Iyengar, M. J. Kolb, J. R. Pankhurst, F. Calle-Vallejo and R. Buonsanti, *ACS Catal.*, 2021, **11**, 4456–4463.

UNIVERSITAT
ROVIRA i VIRGILI