Final version published as Malek, A. , Eslamibidgoli, M. J. , Mokhtari, M. , Wang, Q. , Eikerling, M. H. and Malek, K. (2019), Virtual Materials Intelligence for Design and Discovery of Advanced Electrocatalysts. ChemPhysChem. DOI:10.1002/cphc.201900570

Virtual Materials Intelligence for Design and Discovery of Advanced Electrocatalysts

Ali Malek [a], [b], Mohammad Javad Eslamibidgoli [b], Mehrdad Mokhtari [b], Qianpu Wang [a] , Michael H. Eikerling [b], [c], Kourosh Malek *[a], [b]

Abstract

Similar to the advancements gained from big data in genomics, security, internet of things, and e-commerce, the materials workflow could be made more efficient and prolific through advances in streamlining data sources, autonomous materials synthesis, rapid characterization, big data analytics, and self-learning algorithms. In electrochemical materials science, data sets are large, unstructured/heterogeneous, and difficult to process and analyze from a single data channel or platform. Computer-aided materials design together with advances in data mining, machine learning, and predictive analytics are touted to provide inexpensive and accelerated pathways towards tailor-made functionally optimized energy materials. Fundamental research in the field of electrochemical energy materials focuses primarily on complex interfacial phenomena and kinetic electrocatalytic processes. This perspective article critically assesses AI-driven modeling and computational approaches that are currently applied to those objects. An application-driven materials intelligence platform is introduced, and its functionalities are scrutinized considering the development of electrocatalyst materials for CO_2 conversion as a use case.

[a] Dr. A. Malek, Dr. Q. Wang, Dr. K. Malek
NRC-EME, 4250 Wesbrook Mall,
Vancouver, BC,
V6T 1W5, Canada

[b] Dr. A. Malek, Dr. M. J. Eslamibidgoli, M. Mokhtari, Prof. Dr. M. H. Eikerling, Dr. K. Malek
Department of Chemistry, Simon Fraser University,
8888 University Drive,
Burnaby, BC
V5A 1S6, Canada

[c] Prof. Dr. M. H. Eikerling
Institute of Energy and Climate Research,
Forschungszentrum Jülich,
52425 Jülich,
Germany

Supporting information for this article is given via a link at the end of the document

1. Introduction

Discovery and design of tailor-made materials are crucial drivers of the transition towards a de-fossilized, highly efficient and environmentally benign global energy economy. However, especially in the realm of electrochemical energy technologies, the materials design space is high-dimensional in terms of intrinsic materials properties, structural parameters, dynamic reaction conditions, and the complex interplay of transport and reaction phenomena that must be considered [1-3]. Correspondingly, systematic forays in design, manufacturing, and testing of electrochemical materials demand an unrealistic amount of resources. Therefore, practicable efforts pursued to optimize performance, component integrity, and stability of electrochemical technologies are severely restricted in terms of materials classes and modification strategies, fabrication approaches and conditions, and the range of applications and operating conditions evaluated.

Particularly in the field of electrochemical energy technology, the complex nature of reaction pathways and interdependencies of structure, property, activity, selectivity, and stability of materials calls upon high-throughput albeit efficient semi-automated materials selection and design approaches, as recently explored by various research groups [4-6]. Recent years have witnessed an ever-escalating growth in the production of scientific material data, not only in terms of large data volumes, but also in terms of heterogeneity and format, as well as the computational and experimental methods and tools used for their analyses [7, 8]. This huge amount of material databases with collectively over 0.5 Trillion data points generated annually [7, 8], has caused an inherent complexity among generated data that necessitates rapid storing and probing of structure-process-property relationships using data-mining techniques and machine learning algorithms [8, 9].

One of the main challenges in materials informatics is the creation of material databases that meet requirements for advancing material discovery. Several materials databases are currently available in different database categories. For instance, the Citrination [10] and MatWeb [11] have a computational and experimental database for different types of material properties, e.g., mechanical, optical, electrical, thermal, electrochemical, and structural. The Crystal Structural Database (ICSD) contains crystallographic data [12] and the Cambridge Crystallography Data Centre [13] provides databases of small organic and metal-organic crystal structures. The Materials Project [14], automatic-flow for material discovery (AFLOWLIB) [15], the novel materials discovery (NOMAD) laboratory [16] and open quantum material database (OQMD) [17] have access to many computed structural, physical and chemical properties of materials through computational and various types of electronic structure packages. “Quantum-Machine” offers a specialized database for organic molecules of one billion components [18].

The databases mentioned above, however, insufficiently represent specialized application-based properties such as stability, selectivity, and catalytic activity of a specific electrochemical system for an envisaged application. For instance, electrochemical processes such as CO_2 conversion and H_2 production via electrolysis demand extensive theoretical and computational modeling to understand and predict activities, selectivity, and process performance of electrocatalysts under device-specific operating conditions and for a targeted reaction. In recent years, NRELMatDB (infrastructure of the High-

Performance Computing Center at the National Renewable Energy Laboratory) [19, 20] has been offering a computational materials database with a specific focus on materials for renewable energy applications, including photovoltaic materials, materials for photo-electrochemical water splitting, and thermoelectric materials. Another open repository for chemical reactions on the catalytic surface, became recently accessible at https://www.catalysis-hub.org [21].

2. Motivation for an application-driven platform

Material science is a diverse and interdisciplinary field, in which the advances are accomplished through complex and mutual interactions between data producers, i.e., researchers at universities, government and industry laboratories; funding agencies; lab equipment manufacturers; distributors of research results such as scientific publishers, and manufacturers of materials components or electrochemical devices. All enablers along the value chain play a significant role in the field of material commercialization, in proceeding from inception and conceptualization to lab testing to fabrication scale-up, manufacturing, and up to device integration. These enablers, however, often are not amenable to open data sharing. On the other hand, organizing the material database into standardized, machine-readable forms, such as 'comma-separated values' (CSV) or excel spreadsheet remains as one of the grand challenges of current material informatics. As the main substrate for material discovery and optimal usage of computational data, it is vital to identify meaningful and standardized patterns in massive, interlinked material datasets.

The main motivation for the application-driven approach presented in this perspective article is to offer a simple bridging among crucial gaps in the field of material informatics: i) **the gap in data provenance and management**, which is due to the existence of hugely diverse and uncorrelated data sources; this gap should be overcome by offering a standardized data description and linking data sources with analytics tools to import them with a standard format; ii) **the gap in semantic and materials linguistics,** which is due to multiple vocabularies and different ontologies in the field of materials sciences; to overcome this gap, users can attach existing ontologies and tools for the semantic annotations to data and tools; then the algorithm can provide statistics on which ontologies are most used and which are obsolete according to users' comments; iii) ***the gap in application-specific knowledge***, which is due to heterogeneous and fragmented knowledge sources that collectively support materials properties for a specific application field; to remedy this, the materials platform should link data, tools and workflows to published papers and online documentation; it should also allow rich text description and keywords to be attached to every object in the platform; relevant information should be made visible and searchable; and simple recommendations for data/analysis/workflows integration should be offered; iv) ***the gap in sustainable collaboration***, which is due to differences in user interests, field of expertise, research objectives and methodologies; to address this, users should be required to either provide their profile or import it from a collaborative medium such as Research Gate (www.researchgate.net), in which they fill in their expertise, publications, and tooling preferences.

Either at component (e.g., electrodes, membranes, or membrane electrode assembly), cell, or device levels, the design of electrocatalyst materials is ultimately driven by effectiveness and high performance of electrochemical reactions and related

physico-chemical processes. The kinetics of these reactions depends on the type of catalyst (e.g., metals, metal oxides, metal organic frameworks, or organometallic compounds), chemical structure and composition of the electrolyte, and operating conditions such as pH and applied electrode potential. The complexity of the chemical processes calls upon concerted approaches in data analytics and predictive algorithms, derived from micro-kinetics, component and device level modeling [22, 23] to predict electrocatalyst activity and cell performance under relevant operating conditions.

The rest of this perspective article is organized as follows. In the next section, we critically revisit existing Artificial Intelligence (AI) driven materials design and discovery, details of their start-of-the-art, and the critical gaps that need to be addressed by the materials research community. Thereafter, we introduce and discuss an application-driven Virtual Materials Intelligence (VIMI), emphasizing the importance of data intelligence and analytics. The practicality of the materials data intelligence is elaborated next in the context of new electrocatalyst materials for CO_2 conversion. Finally, a few perspective notes on current AI-driven modeling and computational approaches are provided.

3. Perspective on materials design and discovery

Design and implementation of efficient and cost-effective electrochemical materials is a complex challenge. It hinges on big-data driven knowledge at the frontiers of materials and surface science (electronic materials, catalysts, and polymers), materials synthesis, physical electrochemistry, computational physics and chemistry, and electrochemical engineering.

In recent years, emerging concepts driven from optimization theory and statistics such as predictive and prescriptive analytics have become an integral component of materials science and engineering. The main problem with “Material Informatics” today is the lack of an integrated platform that enables rapid predictions based on past data rather than by direct experimentation or by further computations and simulations, i.e., creating larger datasets. Many data-driven strategies that attempt to address the problem posed above are composed of two distinct steps: (i) numerically represent and reduce the various input material properties to a string of numbers (or “fingerprints”); and (ii) establish a mapping between the fingerprinted input and the target property. These massive data–driven processes, however, require intensive cognitive and thus expensive systems, including humans, to determine the best design decisions. A novel approach towards cognitive analytics, Artificial Intelligence, and Machine Learning algorithms can overcome the latter.

Figure 1 illustrates a generic ecosystem, comprising various data sources and physico-chemical processes which are used in materials discovery. The main distinction is between autonomous and de-centralized approaches. For the autonomous approach, the entire processes of precursor preparation, mixing, testing, and characterization are performed by an automated robotic equipment. In contrast, the de-centralized approach utilizes existing legacy equipment by employing advances in AI and the Internet of Things (IoT) connectivity. This enforces communication among different processes and equipment which can take place seamlessly via cloud computing. A cognitive process with accurate and distinct correlation functions between

structure, functional properties, and performance can enhance the de-centralized approach to materials discovery. The de-centralized approach can bring about a robust and rapid implementation in a more cost-effective fashion than that under an autonomous process.

Key pillars of materials discovery are profound structure-function relationship and relevant correlation functions that can accurately predict materials properties from physico-chemical databases and structural descriptors. Off-the-shelf machine learning (ML) algorithms can apply to big-databases from in-house and literature sources to predict and design advanced energy materials [24-26]. Certain properties can be optimized in view of performance, durability and longevity for energy applications by building AI-driven cognitive algorithms and utilizing widely recognized materials databases, such as Materials Project (MP) [12] , Materials Data Facility (MDF) [27] , and Materials Platform for Data Science (MPDS) [28] . Regardless of the discipline and the nature of materials databases, the goal of AI-driven design should be to support predictive capabilities in materials design, geared towards specific technology challenges and application targets. AI-driven design strategies are anticipated to accelerate and increase efficiencies of materials innovation by reducing the need for intensive human decisions. The first step in any ML process is the access to historical data. Thus, an effort is necessary to mine data from existing literature, create “low-cost” new datasets (e.g. using in-house computational materials techniques) or employ an Application Program Interface (API) to connect to open materials databases. The data storage, data analysis and advanced analysis algorithm therein need to enable efficient and secure dataflow between several different simulation and characterization tasks. We have carefully accounted for the above recipes and implemented design requirements to create and deploy a new materials discovery platform, called “Virtual Materials Intelligence” lab. An ideal materials discovery platform comprises several key processes and functionalities. First, one needs to identify Use-Case materials for a pre-determined application and implement a database of relevant properties, performance, and characterization data that are generally specific to that application. Common materials databases to be employed for manual or automated data extraction are, among others MP, MDF, and MPDS [28]. The next logical step is to create a library of materials properties that are identified in the previous step. In parallel, one needs to implement the API and retrieve data from external or third-party materials databases. This step is primary focused on data mining from existing literature. Testing, validation, cost analysis, and implementation of ML-Algorithm is performed at the next stage. Several algorithms and libraries, ranging from elementary (e.g., linear regression) to highly sophisticated [kernel ridge regression (KRR), decision trees (DT), deep neural networks (DNN), Keras] [29] are freely available from Python developer communities such as PyData, scikit-learn and Scipy stacks [30, 31]. Once the data size increases, there will be a need for managing the data. This last step includes data migration, data curation and warehousing activities, as well as providing services and tools to enable data governance. As described in the next section, VIMI enables seamless data discovery and analyses by leveraging the latest technology in data management, analytics, and services.

4. Virtual Materials Intelligence platform

This section introduces the VIrtual Material Intelligence (VIMI) platform that offers open access to experimental and computational databanks, materials intelligence descriptors, and machine learning algorithms, as well as big data and predictive analytics tools to accelerate the discovery of advanced energy materials.

In line with the requirements depicted in the previous section, one needs a sustainable platform that integrates and manages heterogeneous data and knowledge sources. Moreover, respective processes, methodologies, visualization, and embedded AI/ML tools, all need to be connected through a cognitive workflow editor and execution environment that augments the human-based decision-making process. VIMI is a flexible and scalable data management platform that provides storage, access, processing and visualization of a diverse set of data types and data sources. The main feature differentiating VIMI among similar materials data management platforms is in its application-driven workflow, where key target attributes such as performance, physico-chemical properties, cost and matching to targeted technical properties are defined by default at the outset. Target technical properties must be reviewed and altered as per user demand. The latter is the key governing boundary condition and requirement to guide ML algorithms and property optimization for materials discovery.

VIMI currently encompasses data assets for various applications within clean energy technologies such as 1) electrocatalysts for CO_2 conversion to fuels; 2) membranes and electrocatalysts for polymer electrolyte fuel cell technology and 3) materials for electrochemical energy storage technology [32-34].

Given its broad technology scope, VIMI has potential to create innovative networks and a multi-disciplinary community among researchers and materials developers. As an enterprise solution, it can also facilitate knowledge sharing, and collaboration, and expedites gaining insight for a specific clean energy technology within a single firm or on broader user-bases involving industry and academia.

The practical goal of VIMI is to support the materials industry by providing rapid data-access, cognitive insights, and capabilities in material modeling, materials data analytics, and AI-driven predictive tools, geared towards specific technology challenges and development targets. The intention is to become an industry standard for transferring and accelerating commercialization of new advanced materials.

4.1 VIMI description

Figure 2 illustrates the workflow of the VIMI platform. In the data collection layer, materials databases are formed and collected from a wide range of sources and user-types, namely 1) academic researchers, 2) figures, tables and text in published articles, and 3) industry collaborators.

In the second layer, all datasets are integrated into a master databank with predefined header and standard format that includes data processing and cleaning. The resulting data is stored in a relational database, e.g., MySQL [23] that allows tables to be joined together. In the third layer, the VIMI platform implements the supervised and unsupervised data science tools and techniques such as clustering, classification, statistical modeling, ML algorithms, regression, and big data visualization for

analysis and prediction of material properties. It also contains cognitive criteria for deviation from target indicators for a specific application, defined by default or by the user. Finally, the last layer provides the outcome of the analysis and prediction in the form of visualizations or recommendation to the user.

VIMI provides users not only with a vehicle to upload data, share, and explore datasets but also with an interactive data visualization environment that focuses on multi-dataset exploration. The focus of VIMI is on material datasets for use by the clean energy sector, with the aim of facilitating open science, data sharing and reuse, and analytics of data for industrial applications.

The currently accepted data formats in VIMI are in the form of CSV, Excel and tabular txt files, including binary files. The reason for this restriction is the diversity of information formats from different packages, databases, and researchers. It is necessary to organize a standardized format of data to make quicker analyses and avoid more filtering and data shaving for outlier's data and errors. Moreover, it is vital to enable data import for different scientific and industrial domains. VIMI can support the numerical data, and categorical data for visualization and analytics. It allows a user to create multiple charts to analyze material properties from different perspectives.

While the material research community already has access to a variety of material data resources [10, 12, 15], such as citrination.com, VIMI generates and classifies interactive databases for specific applications like fuel cells, batteries, and electrocatalysts for CO_2 conversion to fuel. VIMI can inherently mine data from the literature in order to better understand materials for thermoelectric [35, 36], Li-ion batteries [37], catalysis, kinetics, and more.

The process for creating an interactive database involves gathering appropriate publications, identifying key data in publications, and extracting figures and tables through semi-automated techniques [35]. The process involves entering numbers and features into text files and digitizing plots in the publications using freeware like *WebPlotDigitizer* (https://automeris.io/WebPlotDigitizer/) and extracting tables using Tabula. Finally, all text files and tables are read by python scripts to transfer to a master database. One of the main advantages of this approach is that a massive amount of information can be stored in tabular form using csv or excel spreadsheet format. This database can be visualized and interpreted using different plots. There is a great advantage of interactive data rather than static tables of data. In fact, by selecting parameters arbitrarily, users will explore previously unexpected, and unobserved correlations in material properties. This can be useful for model development and selection in applying machine learning algorithms. Other features of VIMI are explained in the APPENDIX.

4.2 Use Case: CO_2 Reduction Reaction (CO2RR) Electrocatalysts

This section demonstrates VIMI for the use case of electrocatalysts for CO_2 conversion to fuels. We leveraged the internally and externally enriched databases for hydrogen production as a baseline to train, improve, and optimize AI-driven learning

algorithms for discovery of electrocatalysts [38, 39]. Intercorrelation models already developed for water splitting electrocatalysts are extensively utilized as well [40, 41].

The platform establishes a seamless top-to-bottom data workflow enabled by an AI driven framework that harnesses (i) extensive synthesis and characterization data as well as (ii) predictive analytics based on physical-mathematical modeling of relevant processes and materials. Based on theoretical estimates, one can set out to demonstrate that utilizing such a framework may reduce discovery time and production cost of new materials for CO_2 conversion by factors of 2 and 5, respectively, while markedly improving performance in form of faradaic efficiency and product selectivity [42]. Because such a framework can be adapted for materials discovery in many other applications (by being application-agnostic), the uniqueness of such platform leads to opportunities for wider impacts. These key features make VIMI complementary to, rather than competing or duplicate with, the existing modules of *Material Acceleration Platforms* (MAP) and *Materials Project* currently being developed globally [38, 43, 44]

The first visualization scheme, shown in Figure 3, is a 4-dimensional chart (abscissa, ordinate, color, and marker) for experimental datasets of CO_2 conversion. Electrocatalytic CO_2 conversion is regarded as a prospective pathway for the recycling of carbon resource and the generation of sustainable fuels. The electrocatalyst plays an important role to the product selectivity of CO_2 RR. Different electrocatalysts generate specific carbonaceous compounds, such as HCOOH/$HCOO^-$, CO, formaldehyde (HCHO), hydrocarbons, and alcohols, as well as H_2 as a side product. However, the product selectivity of CO_2 reduction is mechanistically complex and sensitive to the reaction conditions. The sequence of reaction steps depends on several variables and parameters, including electrode potential, electrolyte composition, pH value, temperature, and pressure. We have collected data for a selection of 513 electroctalysts, obtained for a range of reaction conditions, and paired this information with measured product selectivity and faradaic efficiency from the literature [1].

A user can select different descriptive metadata as x, y, color and size within the drop-down list. The faradaic efficiency can be extracted for different products of CO_2 RR while values of current density and applied potential can be visualized as color and size of the circular disks, respectively

Hovering over a data point in Fig. 3 brings up a tooltip with information on the materials specifications of the electrocatalyst, current density, electrode potential, as well as the data source. This type of plot gives the user a rapid screening instrument to understand and compare vital performance trends and correlations for relevant classes of materials.

In Fig. 4 when a user selects a dataset for CO_2 conversion, a list of numerical and categorical descriptive metadata (features) is listed in the x-axis and y-axis sections belonging to the CO_2 experimental dataset. The user can then extract the corresponding values or indexes of measurements from the dataset for visualization and change the range of values for numerical features like faradaic efficiency or electrode potential. This feature helps users to select the best material with their most benign operating conditions. In addition, the user can select and sort the type of products (or any other categorical element) from the right panel of the 2-D plot application. This kind of plot provides a direct comparison between a family of

parameters for different major products and different categories of electrocatalysts for CO_2 RR. Like 4-D plots, a user can query more pertinent information, such as the type and composition of an electrocatalyst, product, and current density as well as the literature sources by hovering over a data point.

Electroatalysts with smaller over-potential, higher activity, and greater selectivity need to be developed to make electrochemical reduction of CO_2 commercially viable. Theoretical calculations could help identify the reaction mechanism and predict which materials are likely to be better electrocatalyst for CO_2 RR. DFT calculations have become a powerful tool for studying catalytic processes and have helped find improved catalyst materials [45, 46].

Within the embedded databases of electronic structure calculations, the VIMI platform currently stores more than 835 of intermediate reaction energies for different CO_2RR electrocatalysts. The catalytic materials of interest for this application includes transition metals, alloys, and metal-oxides. All these mechanisms are based on DFT calculations which were extracted from more than 100 published articles [1]. As a subset of these data, Fig. 5 displays the free energy diagrams for production of HCOOH, CO, CH_4, CHOH, C_2H_4, C_2H_5OH and HCCOH from CO_2 on 135 catalysts. The user can define (e.g. metal, metal oxide, organometallic) or select a specific category of catalyst material, e.g., Ag, Au, Cu, Fe, Ni, Pt and Rh, from the right panel and a product from the left panel. Understanding the mechanisms by which a given product is formed can provide crucial guidance for effective catalyst design strategies.

From an interactive free energy diagram like that shown in Fig. 5, the user can scrutinize the pathways for CO_2 RR on different electrocatalysts and identify the lowest free energy pathways to form a specific product among the selected group of materials. The user will have access to other information such as the adsorption energies of intermediate species in electron volt (eV), catalyst structure in view of crystallographic properties and surface configuration in view of lattice structure, as well as relevant references.

5 Path ahead: VIMI for ML-based catalyst design

Machine learning has proven to be promising both in terms of predicting the potential energy surface of chemical structures, as well as fast discovery and design of target structures. Extensive research is shifting towards employing machine learning in theoretical chemistry for developing inter-atomic potentials [47] or performing ab initio molecular dynamics (AIMD) simulations [48]. ML is also employed in protein design [49], molecular design [50, 51], drug design [52] or materials discovery [53-55].

Through several ground-breaking works by Norskov and his coworkers [56-60] on developing the d-band model and the computational hydrogen electrode (CHE) scheme, quantum mechanical calculations based on DFT have become the quintessential methodology for understanding reaction mechanisms and predicting activity and selectivity of catalyst materials for electrochemical energy storage and conversion (see review paper [22] and references therein). However, system size and complexity of the electrode-electrolyte interface and the immense parameter space it entails limit such methods in terms of consistency and accuracy, and their applicability in materials screening [61-64]. These parameters include nanoparticle size and

shape, composition of catalyst, layer thickness in thin film structures, solvent type and composition, ion concentration and pH, abundance of adsorbed species.

The computational cost of quantum mechanical simulations scales exponentially with the number of atoms considered [65, 66], which limits the size and complexity of systems to which DFT methods can be applied directly up to just a hundred of atoms. These methods are thus inept for electrochemical interfaces which consist of an electrode region, an electrolyte region and a specific interfacial region separating them. The number of degrees of freedom, variables, and parameters involved in determining physical properties, structural dynamics and electrochemical kinetics of this system exceeds the capabilities of quantum mechanical (QM) DFT approaches. Therefore, hierarchical approaches must be employed that have QM-DFT at its core.

As a proper use case for advanced data management platforms such as VIMI, a DFT-based machine learning approach requires a highly qualified dataset for training the model; in machine learning, data is expressed in the form of vectors associated to a high-dimensional feature space. The common task of a machine learning algorithm (whether it is a generative or a discriminative model [67, 68]) is to "learn" from data and make predictions on unseen data. The learning process typically involves an optimization routine, such as stochastic gradient descent [69], to find a set of hidden parameters (like weights associated to features) that minimize an objective function (e.g., minimizing the negative of log-likelihood function). Given the parameters, the predictive model can then be applied to unseen data.

In order to exploit a given set of raw data for a specific system, feature engineering is needed to be used by the research scientist who has the domain knowledge. Therefore, as illustrated in Figure 6, the first step is to generate a combinatorial dataset using DFT, AIMD, or DFT-based Monte-Carlo simulations in order to compute mixtures of structures and identify various surface or bulk properties of nanoparticles or slabs, with or without water layers. An order of tens of thousands of feature vectors are needed to capture the statistics of the system. Data should account for variables for the electronic structure of the solid electrode, solvent properties and ion distributions in the electrolyte as well as specific properties of a boundary region in-between. Such informative features should be provided in standardized form (mean as 0, standard deviation as 1). Appropriate encoding may be required to transform categorical features to numerical ones. To generate this large dataset, an efficient data harvesting method is needed along with human and computational resources.

The second step is to select the machine learning model and to train it for searching, classifying, or clustering the chemical space in terms of structure-property relationship. The latter is an efficient screening/filtering of structure-property relationships to prioritize the materials of interest out of hundreds of thousands good choices [70]. Selection of appropriate statistical analysis techniques such as classification, regression, clustering or dimensionality reduction methods, primarily depends upon size of the data set, (non)linearity of the data, data indexing, memory, scalability and time efficiency of the training algorithms, and accuracy of the predictive data models. Appropriate techniques like random search or cross validation should be employed on training and validation sets to tune the parameters and hyper parameters of the model. Likewise, feature engineering is

required to avoid issues such as highly bias and regularization or dimensionality reduction techniques may also be used to avoid high variance in the model that may result in overfitting problem.

The last step is to use DFT again to calculate the properties of the predicted structures and evaluate the performance an effectiveness of the machine learning algorithms using confusion matrix, area under curve, F1 score, mean absolute error, or mean squared error[71, 72]. Finally, the proposed materials may be experimentally synthesized and tested for further valuation and integration at the component and device levels.

6 Closing remarks

This article presented an application-specific VIMI platform for storing heterogeneous experimental, computational and industrial datasets for clean energy applications. VIMI is designed with a long-term vision to seamlessly integrate data sources through a combination of interfaces, like IoT, user interfaces and API, to build "standardized energy materials data" in real-time with advanced filtering, machine learning and data analytics functionalities. The ultimate goal is to accelerate output data for real-world applications that can greatly accelerate discovery and shorten the time to commercialization. We expect that the VIMI platform will advance the field of clean energy materials. As with any large-scale platform, the development should be a continued effort. It needs steady improvements of the tools employed in data mining/AI/ML. In these processes, we are looking forward to receiving feedback from the material community to advance this platform. In addition, we encourage users, experimentalist, theorists and industrial users, to collaborate and share their data.

Acknowledgments

This work was supported by German-NRC collaboration project. KM and QW would like to thank NRC international office for their financial support. The authors also greatly acknowledge Dr. Olivier Guillon and his team at the Forschungszentrum Jülich for valuable insights and contributions to this project. Contributions from Shervan Gheidi and Gagandeep Singh Bajwa at NRC-EME for the development VIMI are greatly acknowledged. MJE, MM, and MHE would like to acknowledge support of this work by the *Engineered Nickel Catalysts for Electrochemical Clean Energy* project administered from Queen's University and supported by Grant No. RGPNM 477963-2015 under the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Frontiers Program."

Keywords: Electrocatalyst, material discovery, AI-driven, machine learning, Internet of Things

References:

[1] M. Eikerling, A. Kulikovsky, Polymer electrolyte fuel cells: physical principles of materials and operation, CRC Press, **2014**.
[2] M. K. Debe, *Nature* **2012**, *486*, 43.
[3] M. A. Hannan, M. H. Lipu, A. Hussain, A. Mohamed, *Renewable Sustainable Energy Rev.* **2017**, *78*, 834-854.
[4] J. D. Benck, B. A. Pinaud, Y. Gorlin, T. F. Jaramillo, *PLoS One* **2014**, *9*, e107942.
[5] V. Stamenkovic, B. S. Mun, K. J. Mayrhofer, P. N. Ross, N. M. Markovic, J. Rossmeisl, J. Greeley, J. K. Nørskov, *Angew. Chem., Int. Ed.* **2006**, *45*, 2897-2901.
[6] N. Marković, P. Ross Jr, *Surf. Sci. Rep.* **2002**, *45*, 117-229.

[7] P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, E. Sargent, *Nature* **2017**, *552*, 23-25.
[8] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, *Mrs Bulletin* **2016**, *41*, 399-409.
[9] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, *Comput. Mater. Sci.* **2018**, *152*, 60-69.
[10] C. Citrine Informatics , available at https://citrination.com
[11] L. M. MatWeb, Material property data.[cited 2012 6th May]." Data base of materials data sheets.
[12] A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Cryst. B* **2002**, *58*, 364-369.
[13] F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, D. G. Watson, *Acta Cryst. B* **1979**, *35*, 2331-2339.
[14] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[15] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227-235.
[16] The Novel Materials Discovery (NOMAD) Repository, https://repository.nomad-coe.eu.
[17] http://oqmd.org.
[18] http://quantum-machine.org/datasets/.
[19] V. Stevanović, S. Lany, X. Zhang, A. Zunger, *Phys. Rev. B* **2012**, *85*, 115104.
[20] S. Lany, *Phys. Rev. B* **2013**, *87*, 085112.
[21] K. Winther, Max J. Hoffmann, Osman Mamun, Jacob R. Boes, Jens K. Nørskov, Michal Bajdich, and Thomas Bligaard. "Catalysis-hub. org: An Open Electronic Structure Database for Surface Reactions." ChemRxiv (2018).
[22] M. J. Eslamibidgoli, J. Huang, T. Kadyk, A. Malek, M. Eikerling, *Nano Energy* **2016**, *29*, 334-361.
[23] D. Axmark, M. Widenius, *Linux journal* **1999**, *1999*, 5.
[24] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[25] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810.
[26] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, *Nat. Rev. Mater.* **2018**, *3*, 5-20.
[27] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, *JOM* **2016**, *68*, 2045-2052.
[28] https://mpds.io/.
[29] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291-1307.
[30] W. McKinney in PyData Development Team." pandas: powerful Python data analysis toolkit, Release 0.18. 1" Dated May 3, **2016**.
[31] E. Jones, T. Oliphant, P. Peterson, **2014**.
[32] P. Alotto, M. Guarnieri, F. Moro, *Renewable Sustainable Energy Rev.* **2014**, *29*, 325-335.
[33] S. Ould Amrouche, D. Rekioua, T. Rekioua, S. Bacha, *Int. J. Hydrogen Energy* **2016**, *41*, 20914-20927.
[34] M. Aneke, M. Wang, *Appl. Energy* **2016**, *179*, 350-377.
[35] R. Seshadri, T. D. Sparks, *APL Mater.* **2016**, *4*, 053206.
[36] M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland, B. Meredig, *APL Mater.* **2016**, *4*, 053213.
[37] E. Chemali, P. J. Kollmeyer, M. Preindl, A. Emadi, *J. Power Sources* **2018**, *400*, 242-255.
[38] W. Zhang, Y. Hu, L. Ma, G. Zhu, Y. Wang, X. Xue, R. Chen, S. Yang, Z. Jin, *Adv. Sci.* **2018**, *5*, 1700275.
[39] J. Qiao, Y. Liu, J. Zhang, Electrochemical reduction of carbon dioxide: fundamentals and technologies, CRC Press, **2016**.
[40] W. Li, R. Jacobs, D. Morgan, *Comput. Mater. Sci.* **2018**, *150*, 454-463.
[41] Q. Xu, Z. Li, M. Liu, W.-J. Yin, *J. Phys. Chem. Lett.* **2018**, *9*, 6948-6954.
[42] J. L. DiMeglio, J. Rosenthal, *J. Am. Chem. Soc.* **2013**, *135*, 8798-8801.
[43] J. D. Shakun, P. U. Clark, F. He, S. A. Marcott, A. C. Mix, Z. Liu, B. Otto-Bliesner, A. Schmittner, E. Bard, *Nature* **2012**, *484*, 49.
[44] M. R. Singh, J. D. Goodpaster, A. Z. Weber, M. Head-Gordon, A. T. Bell, *Proc. Natl. Acad. Sci. U.S.A* **2017**, *114*, E8812-E8821.
[45] Y.-A. Zhu, D. Chen, X.-G. Zhou, W.-K. Yuan, *Catal. Today* **2009**, *148*, 260-267.
[46] A. A. Peterson, F. Abild-Pedersen, F. Studt, J. Rossmeisl, J. K. Nørskov, *Energy Environ. Sci.* **2010**, *3*, 1311-1315.
[47] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nat. Commun.* **2017**, *8*, 872.
[48] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, *Sci. Adv.* **2017**, *3*, e1603015.
[49] J. Wang, H. Cao, J. Z. Zhang, Y. Qi, *Sci. Rep.* **2018**, *8*, 6349.
[50] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams in *Convolutional networks on graphs for learning molecular fingerprints, Vol.*, Advances in neural information processing systems, **2015**, pp.2224-2232.
[51] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360-365.
[52] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268-276.
[53] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *APL Mater.* **2013**, *1*, 011002.
[54] S. K. Jha, J. Bilalovic, A. Jha, N. Patel, H. Zhang, *Renewable Sustainable Energy Rev.* **2017**, *77*, 297-317.
[55] B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, C. A. Sutton, *AIChE-Journal* **2018**, *64*, 2311-2323.
[56] J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard, H. Jonsson, *J. Phys. Chem. B* **2004**, *108*, 17886-17892.
[57] B. Hammer, J. Nørskov, *Surf. Sci.* **1995**, *343*, 211-220.
[58] B. Hammer, J. K. Nørskov in *Theoretical surface science and catalysis—calculations and concepts, Vol. 45*, Elsevier, **2000**, pp.71-129.
[59] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, *Nat. Chem.* **2009**, *1*, 37.
[60] J. Greeley, I. E. L. Stephens, A. S. Bondarenko, T. P. Johansson, H. A. Hansen, T. F. Jaramillo, J. Rossmeisl, I. Chorkendorff, J. K. Nørskov, *Nat. Chem.* **2009**, *1*, 552.
[61] F. Calle-Vallejo, M. T. Koper, *Electrochim. Acta* **2012**, *84*, 3-11.
[62] M. J. Eslamibidgoli, M. H. Eikerling, *Curr. Opin. Electrochem.* **2018**, *9*, 189-197.
[63] O. M. Magnussen, A. Groß, *J. Am. Chem. Soc.* **2019**.
[64] A. Groß, S. Sakong, *Curr. Opin. Electrochem.* **2018**.
[65] I. M. Georgescu, S. Ashhab, F. Nori, *Rev. Mod. Phys.* **2014**, *86*, 153.
[66] I. Kassal, S. P. Jordan, P. J. Love, M. Mohseni, A. Aspuru-Guzik, *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105*, 18681-18686.
[67] T. Jebara in Machine learning: discriminative and generative, vol. 755, Springer Science & Business Media, **2012**.
[68] Z. Ghahramani, *Nature* **2015**, *521*, 452.
[69] L. Bottou in *Large-scale machine learning with stochastic gradient descent*, Springer, **2010**, pp.177-186.
[70] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *89*, 094104.
[71] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, *Annu. Rev. Mater. Res.* **2015**, *45*, 195-216.
[72] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241-2251.

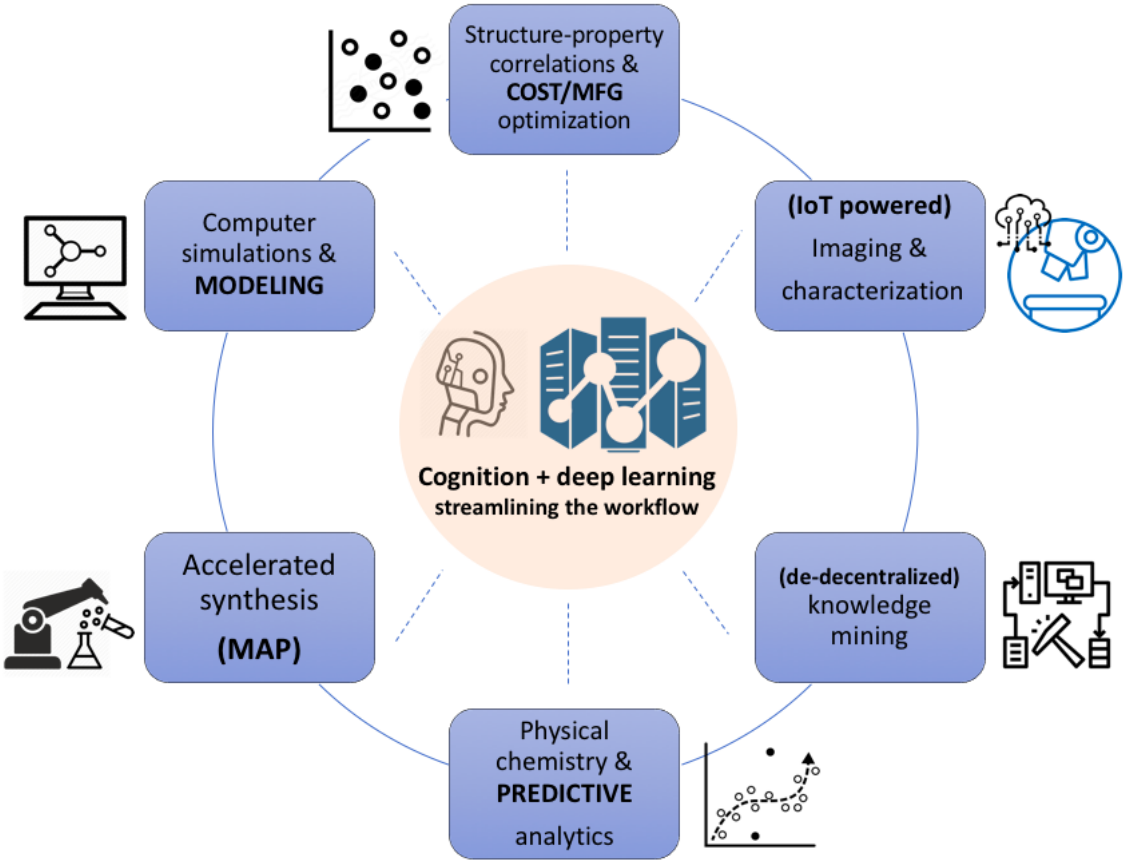

Figure 1. Schematics of current approaches in AI-driven materials design and discovery. [Adopted from A 3-part webinar from the Material Research Society, presented by academic material scientists that use AI].

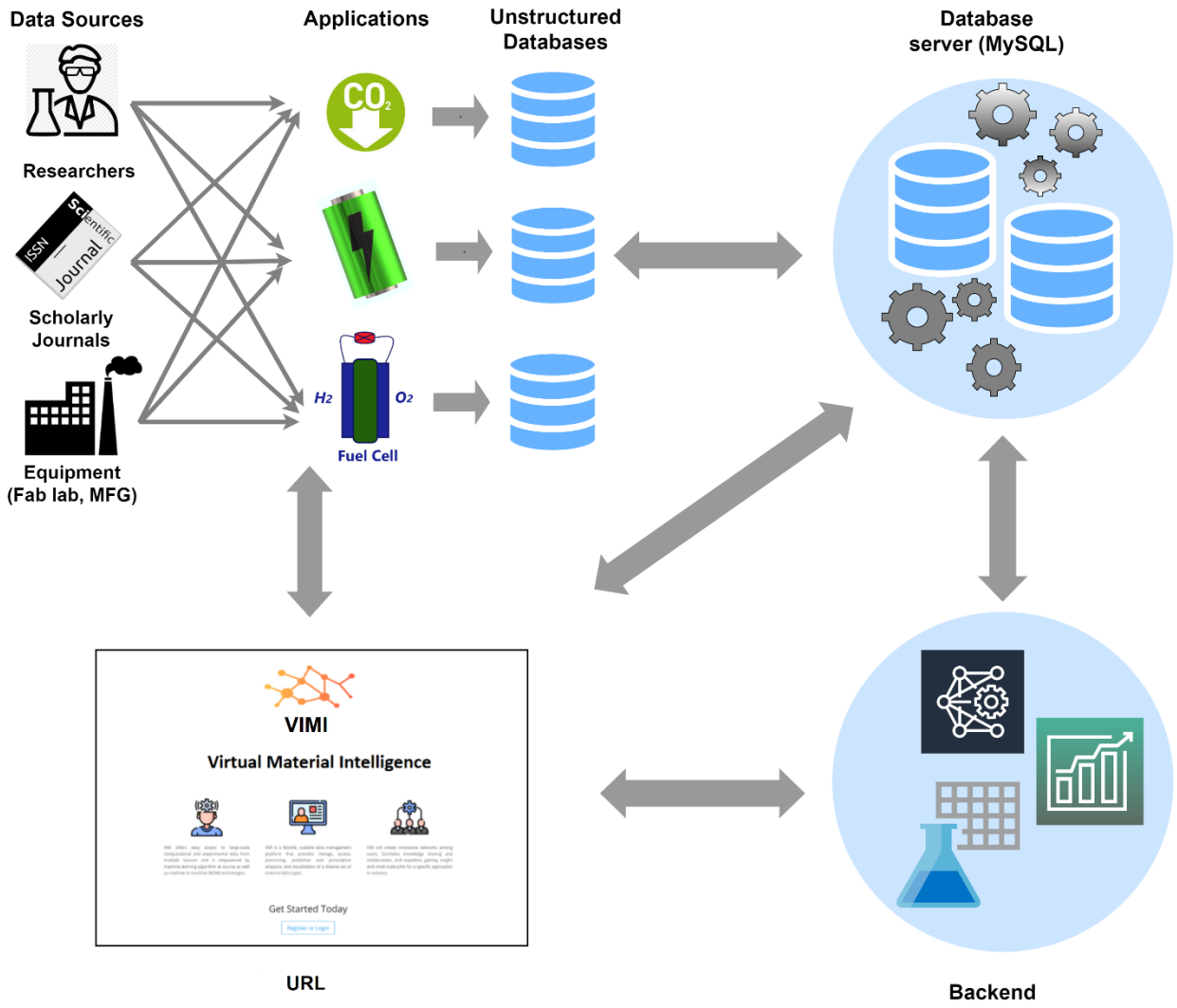

Figure 2. Workflow of VIMI platform, showing the relation between database for applications, database server, backend and frontend webpage.

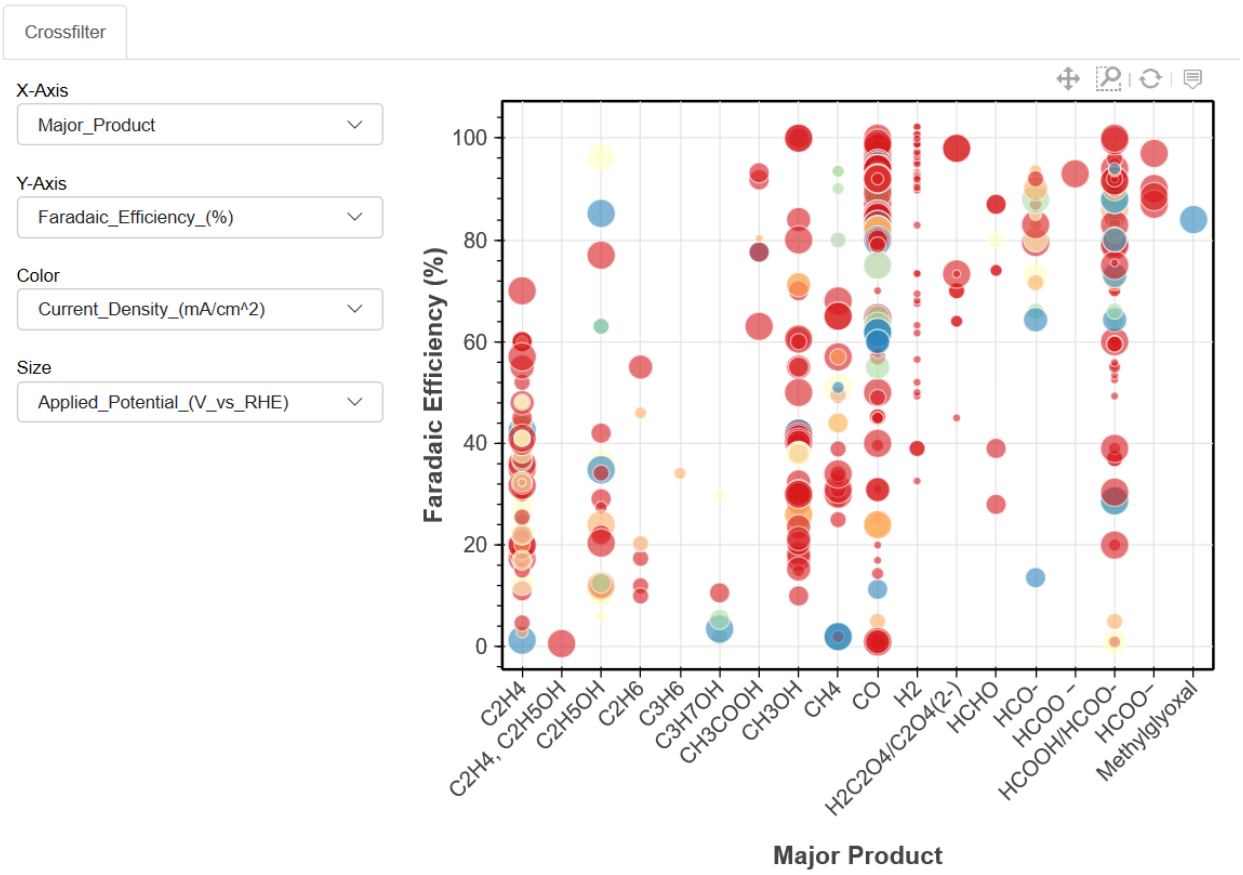

Figure 3. Screenshot of web-based visualization tool, that permits the simultaneous visualization of four parameters among various list of measurements. Several variables can be chosen as abscissa and ordinate.

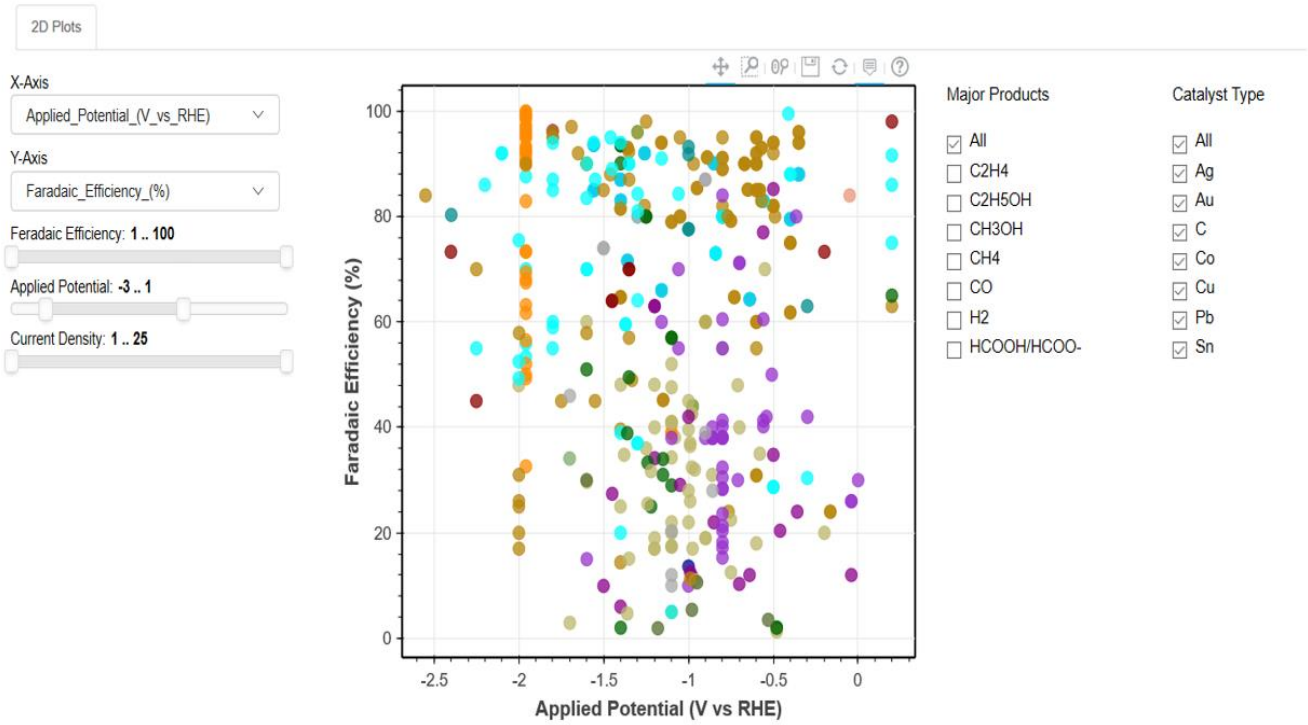

Figure 4. Screenshot of web-based visualization tool, that permits the simultaneous visualization of two parameters. The range of values for x-axis and y-axis can be changed via sliders. Users may select the type of products and catalyst.

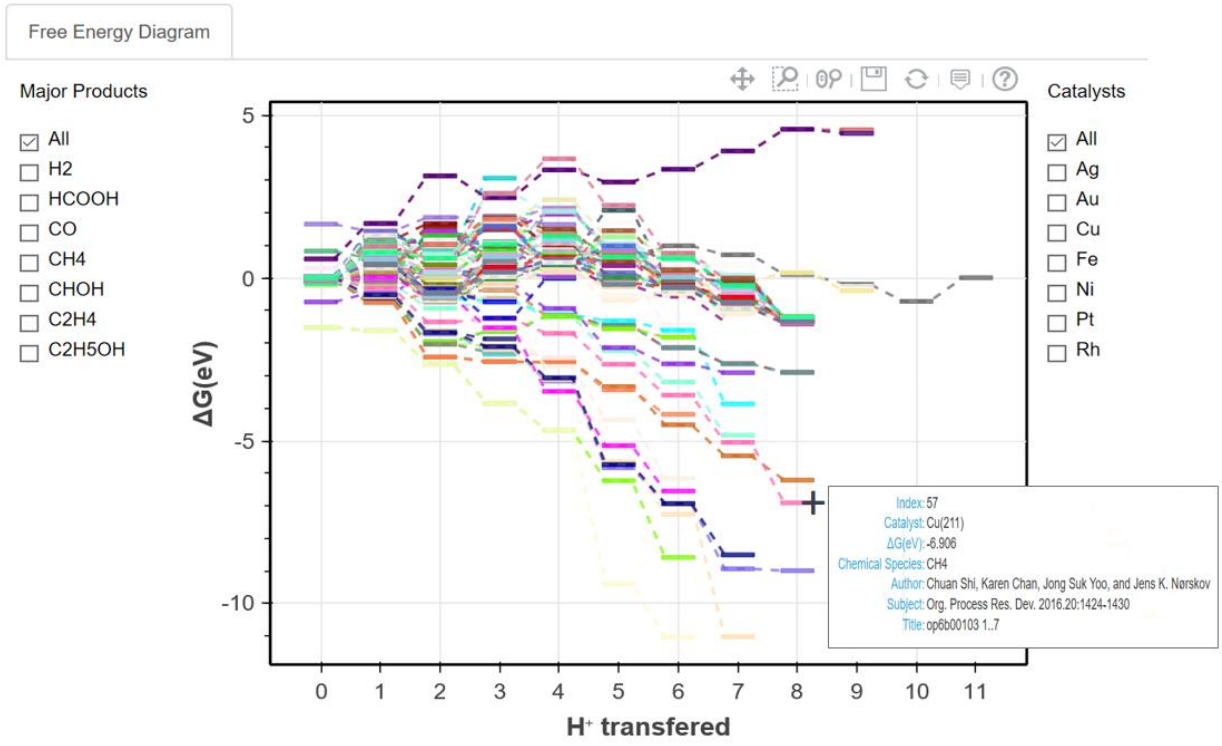

Figure 5. Screenshot of Free energy diagrams of CO2 conversion to different products

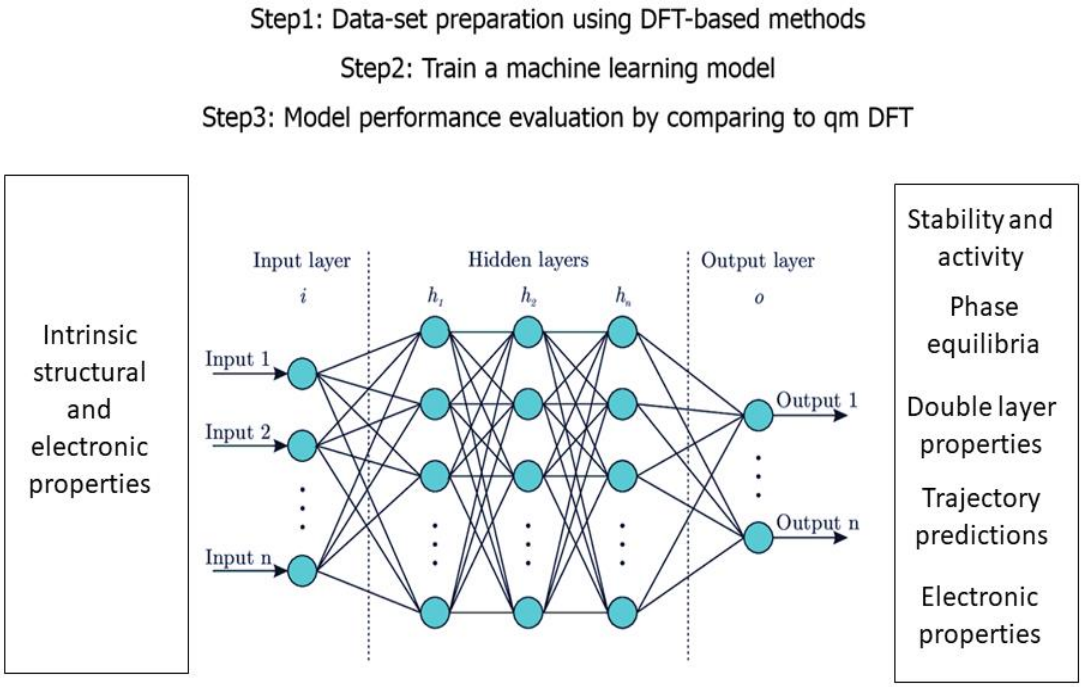

Figure 6. General schematic for DFT-based machine learning methodology in electrocatalysis