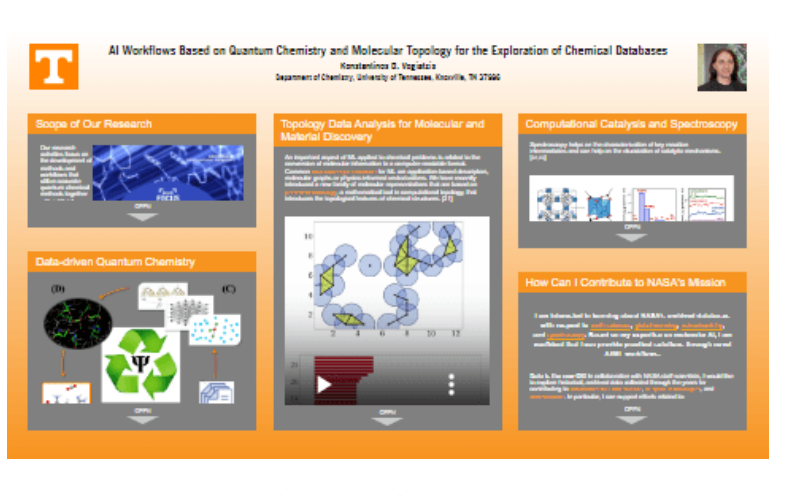
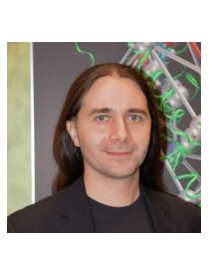
AI Workflows Based on Quantum Chemistry and Molecular Topology for the Exploration of Chemical Databases

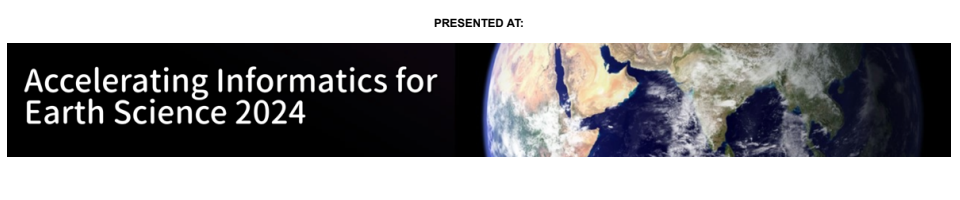

AI Workflows Based on Quantum Chemistry and Molecular Topology for the Exploration of Chemical Databases
Konstantinos D. Vogiatzis
Department of Chemistry, University of Tennessee, Knoxville, TN 37996
Scope of Our Research
Topology Data Analysis for Molecular and Material Discovery
Computational Catalysis and Spectroscopy
Data-driven Quantum Chemistry
How Can I Contribute to NASA's Mission

Konstantinos D. Vogiatzis

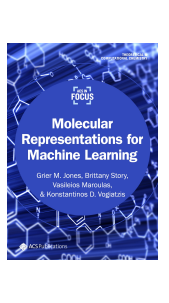
Department of Chemistry, University of Tennessee, Knoxville, TN 37996



PRESENTED AT:

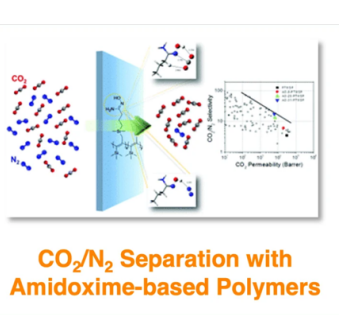

Accelerating Informatics for Earth Science 2024

SCOPE OF OUR RESEARCH

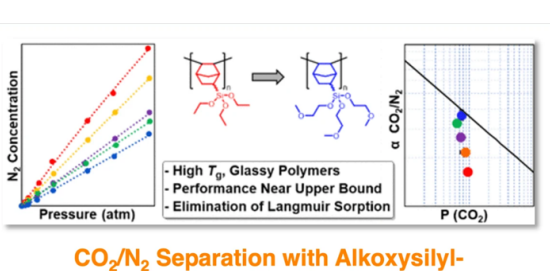

ACS in Focus
Molecular Representations for Machine Learning
Grier M. Jones, Brittany Story, Vasileios Maroulas, & Konstantinos D. Vogiatzis
ACS Publications

Our research activities focus on the development of methods and workflows that utilize accurate quantum chemical methods together with artificial intelligence and machine learning. Our overall objectives are to elucidate the fundamental physical principles underlying chemical reactivity and catalysis, separation processes, and heavy element chemistry, as well as to assist in the interpretation of experimental data. One of our aims is the development of effective methods that convert chemical information into computer-readable format (**molecular representations**). Representative examples from our research activities include:

1. Discovery of New Molecules with Stronger CO_2-affinity. We have contributed on the theoretical description of the noncovalent interactions of amine-based molecules with CO_2. [1-4] Our work has led to the design and synthesis of novel polymeric materials for diffusion-based CO_2 separation processes. We have developed novel machine learning workflows which led to the discovery of new molecules with enhanced CO_2 affinity. [1]

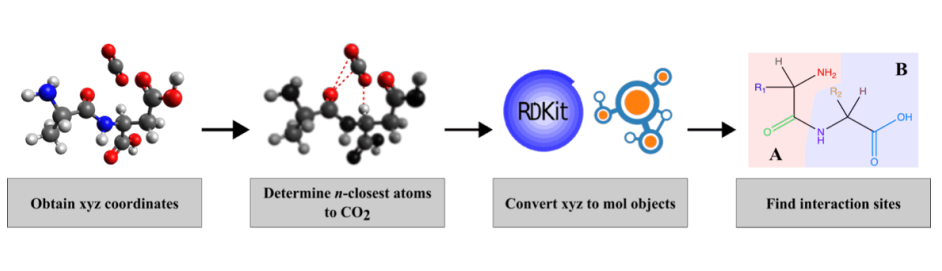


CO_2/N_2 Separation with Amidoxime-based Polymers

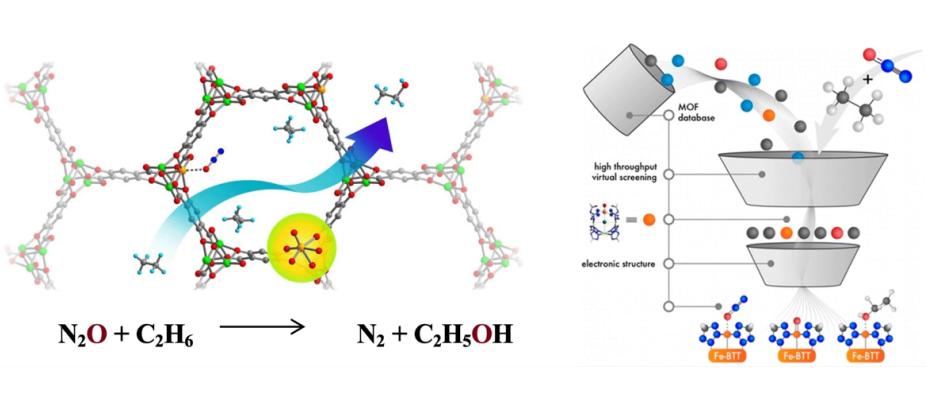


CO_2/N_2 Separation with Alkoxysilyl-substituted Vinyl-added Polynorbornenes

2. CO_2 Separation with Polymeric Membranes. Our group has published a series of articles on high-performing CO_2 separation membranes in collaboration with synthetic polymer chemists. [5-9] Our work has elucidated the atomistic effects between gas molecules and repeat units of polymeric materials that enhance CO_2/N_2 separation processes. We have also been active on the development of new machine learning models for polymer informatics. [10]



3. C-H Activation with Metal-Oxo Sites. We have elucidated the reaction mechanism of the oxidation of ethane to ethanol catalyzed by Fe(IV)-oxo sites, the first reported example of this reaction performed inside a metal-organic framework (MOF) [11]. That study was the basis for the first computational screening of MOF databases for oxidation catalysis. [12] We have explored the ligand field effects on the ground and excited states of reactive Fe-oxo and Cu-oxo species. [13,14] Recently, we published our first study on Fe(IV)-oxo catalyst optimization with novel machine learning tools developed by our group. [15]

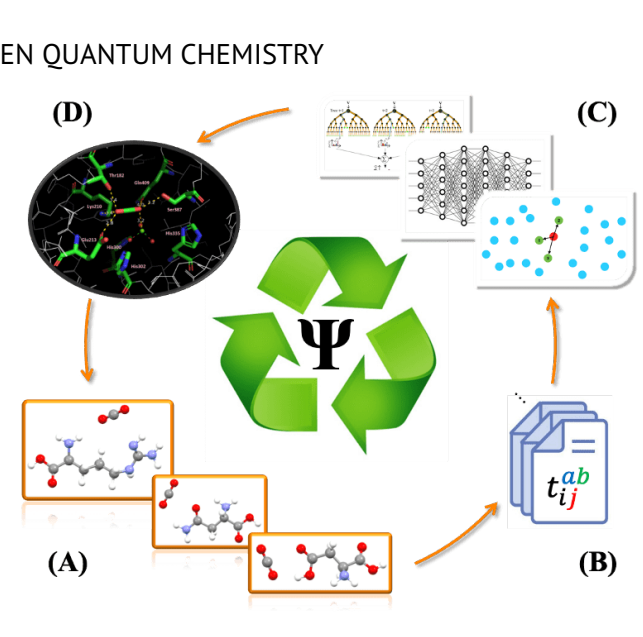

$N_2O + C_2H_6 \longrightarrow N_2 + C_2H_5OH$
MOF database
high throughput virtual screening
electronic structure

4. High Performance Computing for Electronic Structure Theory. We have contributed to the development of methods included in popular quantum chemical program packages (TURBOMOLE, NWChem). For example, our massively parallel implementation of the complete-active-space self-consistent-field (CASSCF) method maintains the record of the largest configuration interaction calculation with conventional methods. [16] Currently, our research group develops a novel hybrid scheme that combines conventional electronic structure theory methods with machine learning and provides high accurate results in significantly less computational effort. [17-20]

New Frontier:

Elevate accurate theoretical and computational chemistry from studies of a few molecules to the exploration of a larger chemical space

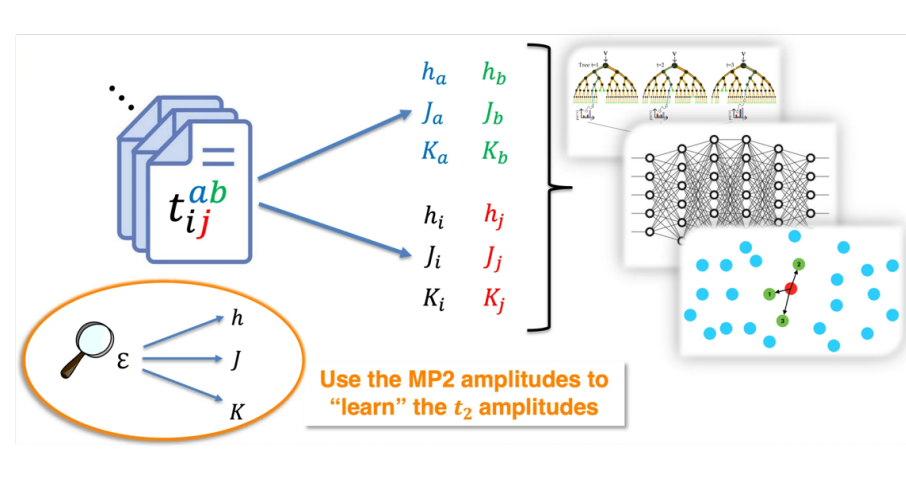
DATA-DRIVEN QUANTUM CHEMISTRY



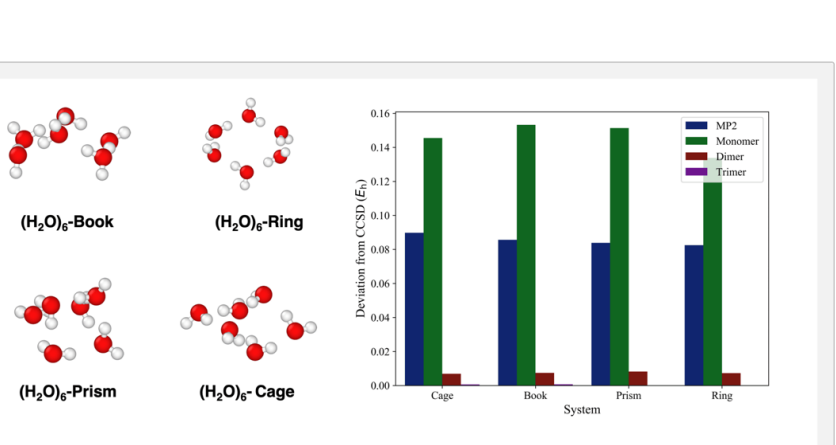
The **data-driven quantum chemistry** (DDQC) project aims on the development of methodologies that capitalize on recent progress of data sciences for expanding the applicability of accurate quantum chemical methods. Our approach is based on the "recycling" of molecular wave functions obtained at low computational cost to help train ML models which will provide fast and reliable energies and geometries of complex molecular systems without loss of accuracy. [17]

Coupled-cluster singles-and-doubles with perturbative triples excitations (CCSD(T)), the gold standard of quantum chemistry, is a method that provides high accuracy for a variety of chemical applications including chemical reactions, non-covalent interactions, and spectroscopy. An important feature of coupled-cluster is its systematic increase of accuracy with the increase of the basis sets size and the electron excitation manifold. However, the applicability of CCSD(T) is hindered by the system size. Quantum chemistry has greatly gained in the past years by the progress in data science, which has helped scientists to accelerate simulations and explore a large fraction of the chemical space in a reliable and systematic manner. In addition, data sciences have aided to surpass the size limitations of accurate methods such as CCSD(T) and increase their applicability.

The major goal of this project is the development of DDQC methods and, in particular, data-driven coupled-cluster (DDCC) approaches. These methodologies are based on electron correlation, a property that has a local, short-range character across all molecular species independent of their size. The goal of this project is to build transferable machine learning models that "learn" highly accurate CCSD(T) wave functions by utilizing data from low-cost methods such as Hartree-Fock (HF) and second-order perturbation theory (MP2).

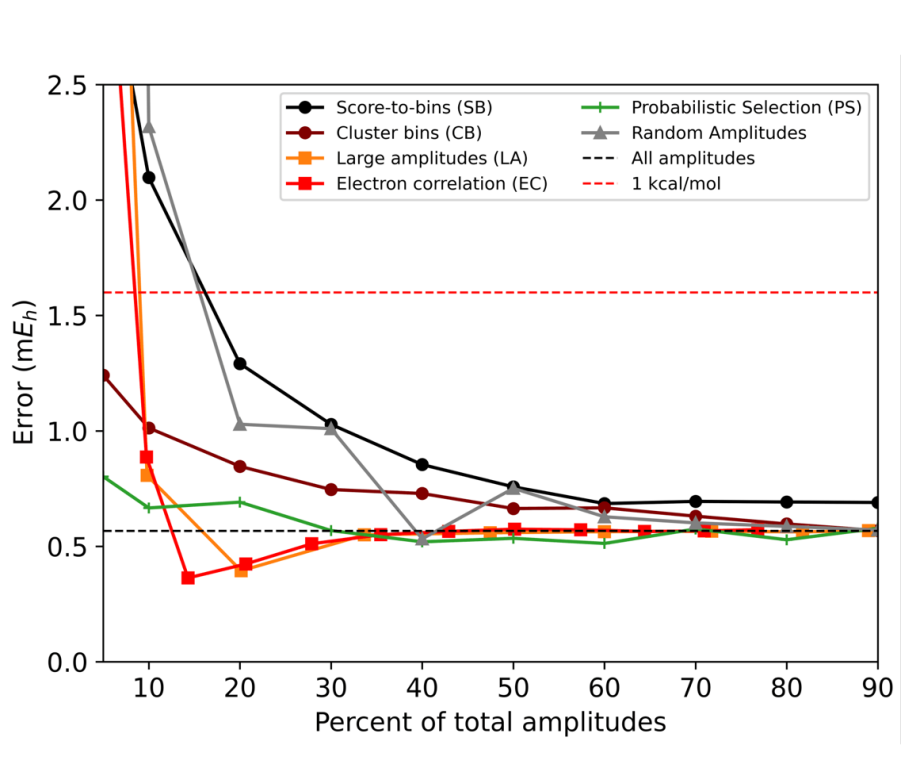


The DDCC scheme with singles-and-doubles excitations (DDCCSD) utilize data that describe a two-electron promotion to predict the two-electron amplitudes (t_2) that are needed for the computation of the CCSD energy. The method utilizes a physics-based molecular representation of the two-electron promotions. An important advantage of the proposed methodology is that DDCC is based on learning complex molecular wave functions. [18]

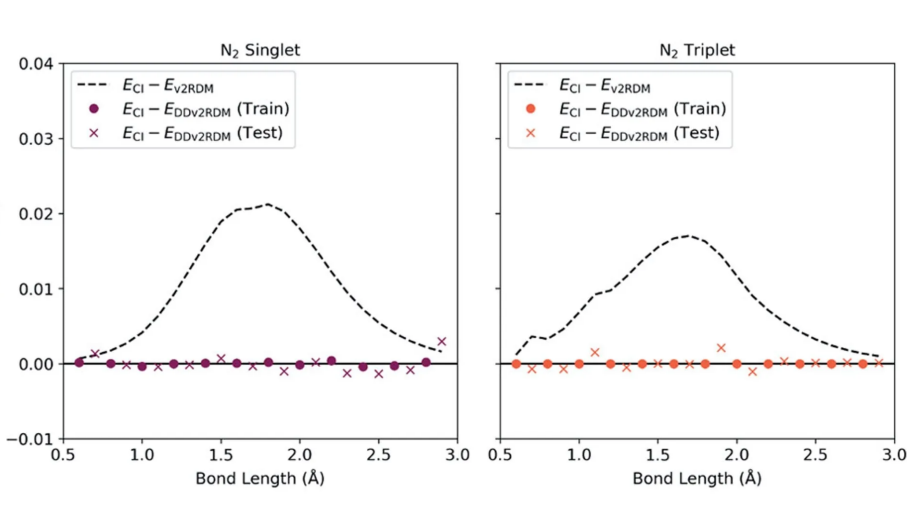


DDCC deviation for the four water hexamer structures trained with data from water monomers, dimers, and trimers.

One major limitation of the DDCC models is the size of training sets that increases exponentially with the system size. We have developed five different amplitude selection techniques (score-to-bins, cluster-to-bins, large amplitudes, electron correlation, and probabilistic selection) that reduce the amount of data used for training were evaluated, an approach that also prevents model overfitting and increases model's portability to more complex molecules or larger basis sets. [19]



Recently, we developed a similar data-driven scheme based on variational 2-reduced density matrix (v2RDM) theory for estimating the full configuration interaction (FCI) limit of small molecules. [20]



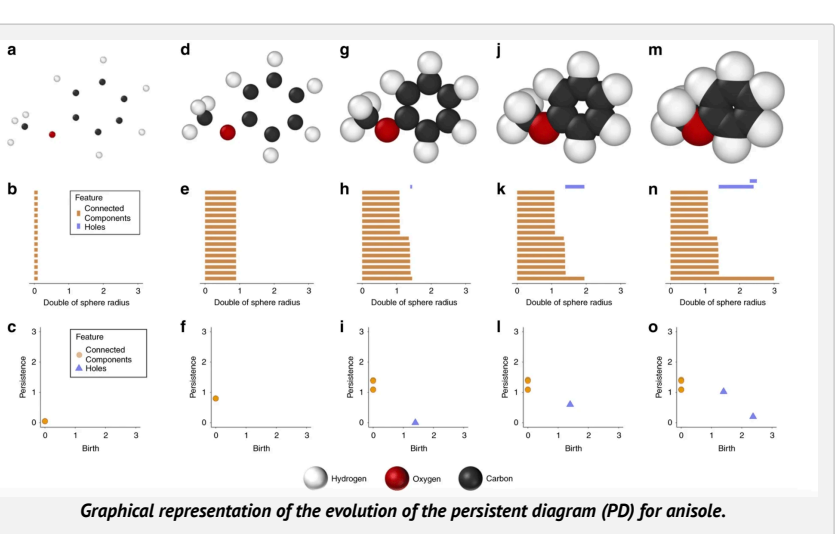
Currently, we are exploring more complex ML architectures, and we are expanding the DDQC applicabilty to properties and molecular geometries.

TOPOLOGY DATA ANALYSIS FOR MOLECULAR AND MATERIAL DISCOVERY

An important aspect of ML applied to chemical problems is related to the conversion of molecular information to a computer-readable format. Common **molecular representations** for ML are application-based descriptors, molecular graphs or physics-informed vectorizations. We have recently introduced a new family of molecular representations that are based on **persistent homology**, a mathematical tool in computational topology, that introduces the topological features of chemical structures. [21]

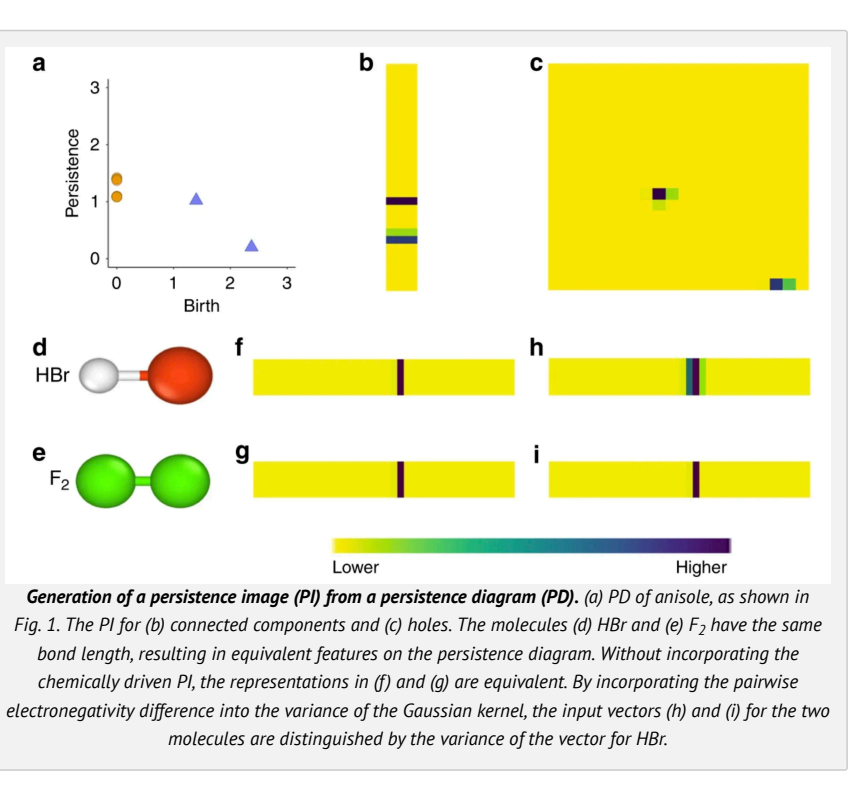
[VIDEO] https://res.cloudinary.com/amuze-interactive/video/upload/q_auto/v1716925342/tomandjerry/0F-C0-EF-D8-32-F5-D3-E8-94-88-B4-AC-8D-6A-73-AB/Video/Screen_Recording_2023-04-01_at_1.05.34_PM_y7lyp5.mp4

Specifically, the topological features (connected components or chemical *bonds*, *rings*, *cavities*) that persist across multiple scales are introduced into a diagram (**persistence diagram** or **PD**): [1]



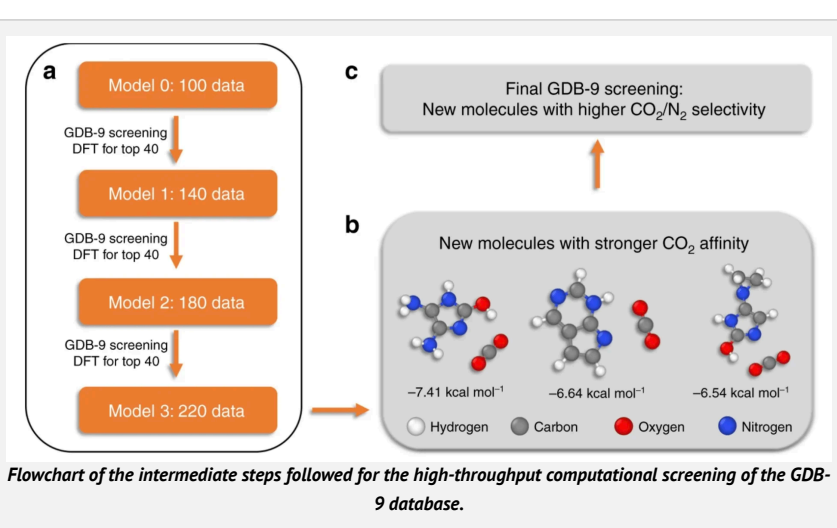
Graphical representation of the evolution of the persistent diagram (PD) for anisole.

The **persistence diagram** is then vectorized into a fixed-size image (**persistence image** or **PI**) that also consider the atom types of the molecular structures:



Generation of a persistence image (PI) from a persistence diagram (PD). *(a) PD of anisole, as shown in Fig. 1. The PI for (b) connected components and (c) holes. The molecules (d) HBr and (e) F_2 have the same bond length, resulting in equivalent features on the persistence diagram. Without incorporating the chemically driven PI, the representations in (f) and (g) are equivalent. By incorporating the pairwise electronegativity difference into the variance of the Gaussian kernel, the input vectors (h) and (i) for the two molecules are distinguished by the variance of the vector for HBr.*

We have utilized **PIs** as input to ML for the discovery of molecules with stronger CO_2 affinity. By computing the CO_2 interaction energies of only 220 molecules, we were able to accurately and reliably examine a dataset of ~133k molecular structures (GDB-9 database):

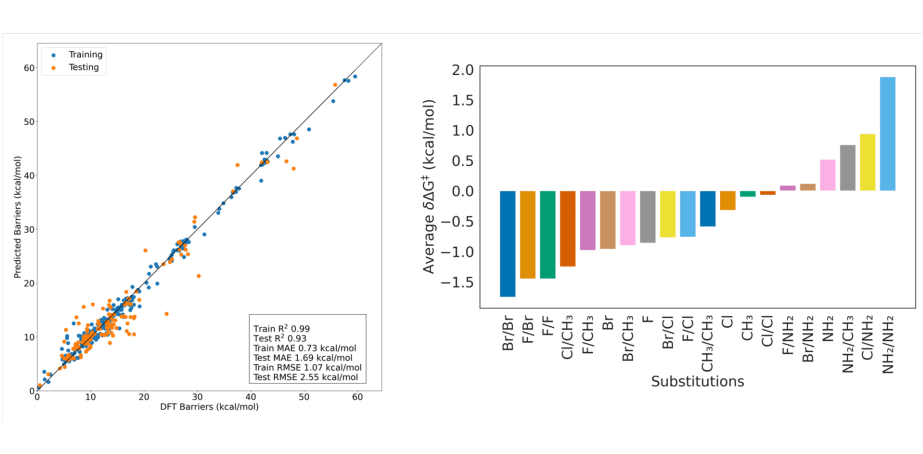


Flowchart of the intermediate steps followed for the high-throughput computational screening of the GDB-9 database.

We have also used PIs for encoding the topological information of organic ligands for catalyst design. [15] A database of 181,436 unique transition metal complexes bearing a Fe(IV)-oxo intermediate was compiled:

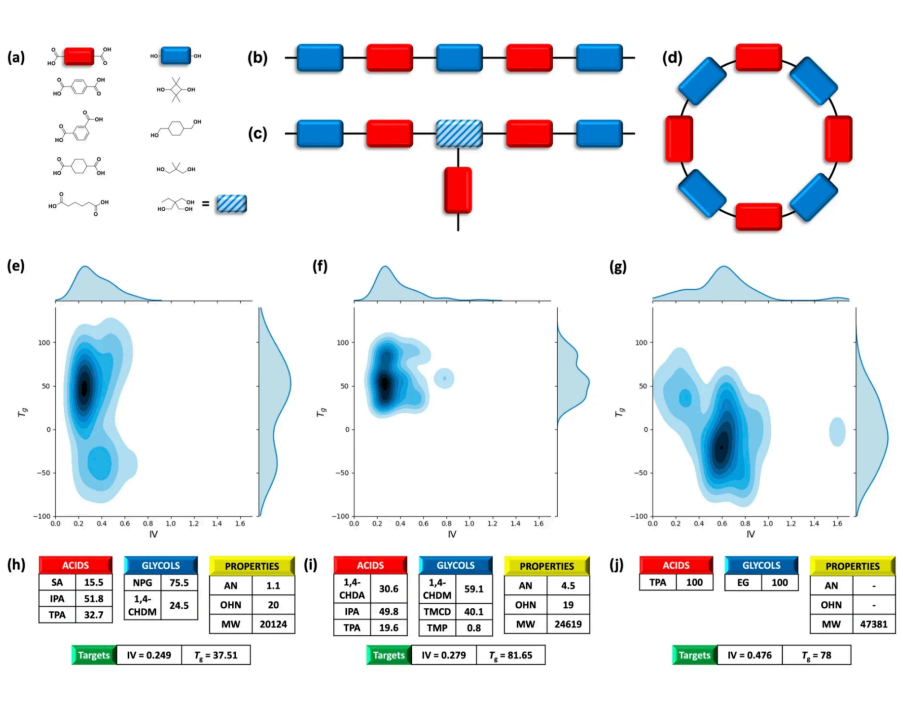
[VIDEO] https://res.cloudinary.com/amuze-interactive/image/upload/f_auto,q_auto/v1717030104/tomandjerry/0f-c0-ef-d8-32-f5-d3-e8-94-88-b4-ac-8d-6a-73-ab/image/picture1_zzrywm.mp4

We developed two ML models by utilizing data from only 500 transition metal complexes. The first was on spin classification (91.1% accuracy) and the second was accurately predicting the C-H activation barrier (mean absolute deviation of 1.69 kcal/mol):

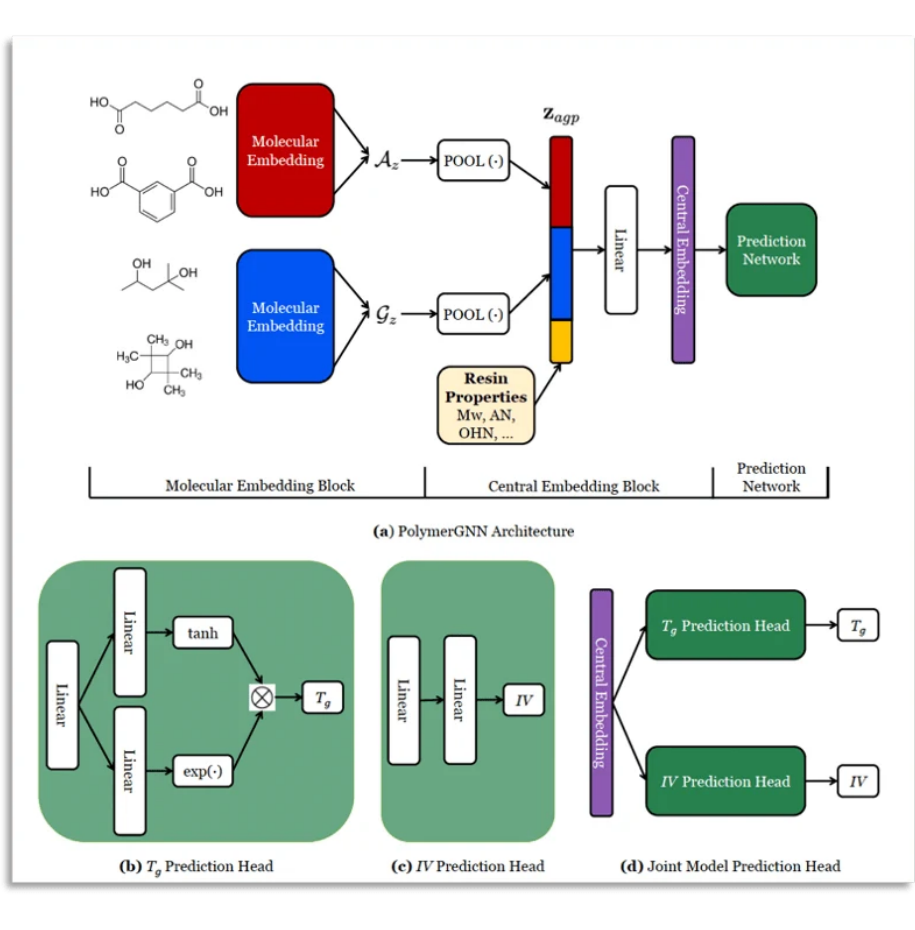


These two ML models helped us to develop new structure-catalytic functionality relationships for Fe(IV)-oxo complexes. Our group is currently utilizing the **PI** technology for **direct air capture**, ligand-based **actinide separations** and **drug discovery**.

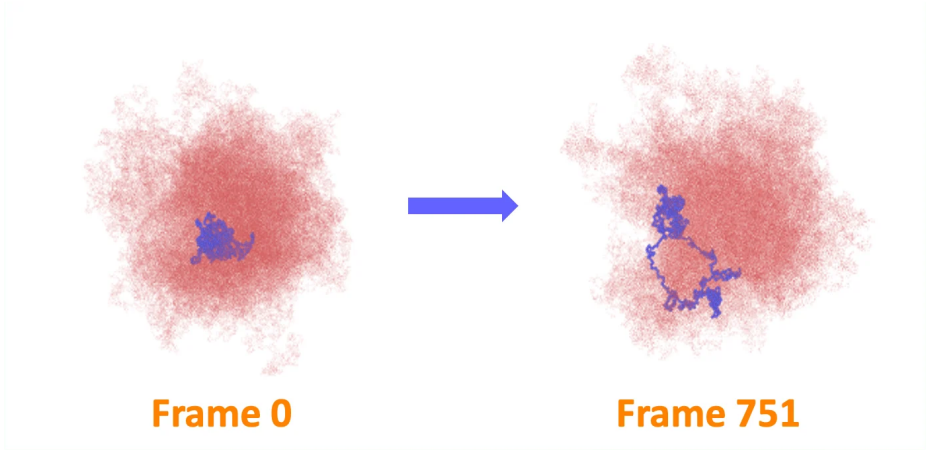
More recently, we developed **PolymerGNN**, a polymer graph neural network architecture for multitask property learning.[10] We have utilized experimentally-refined data by Eastman for a small dataset of polyesters (*glass transition temperature* and *intrinsic viscosity*):

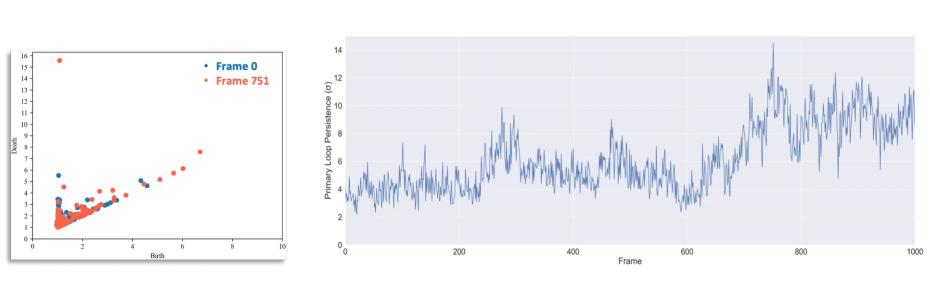


PolymerGNN allowed us to accurately estimate the viscosity of polymers by only considering the topology of the monomers composing each material:



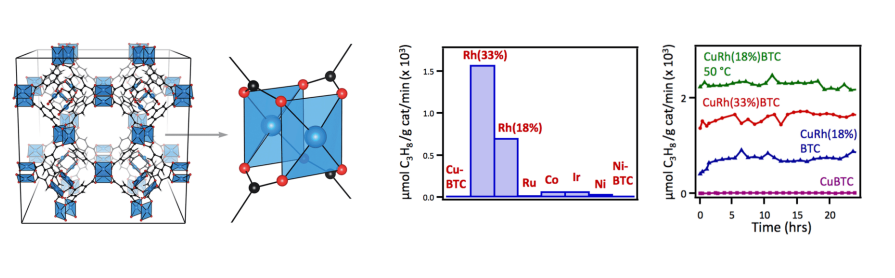
We have also applied **persistence homology** as a systematic **post-analysis tool of molecular simulations**. For example, we have analyzed data from ring polymers across different time frames and monitored the evolution of a ring chain from a globular to an expanded configuration.



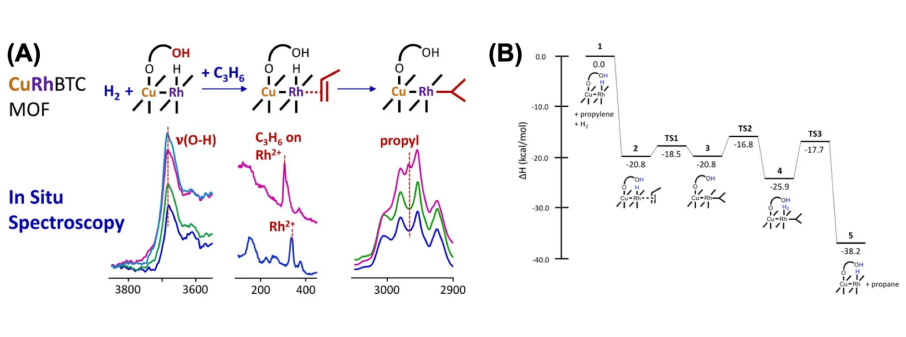


COMPUTATIONAL CATALYSIS AND SPECTROSCOPY

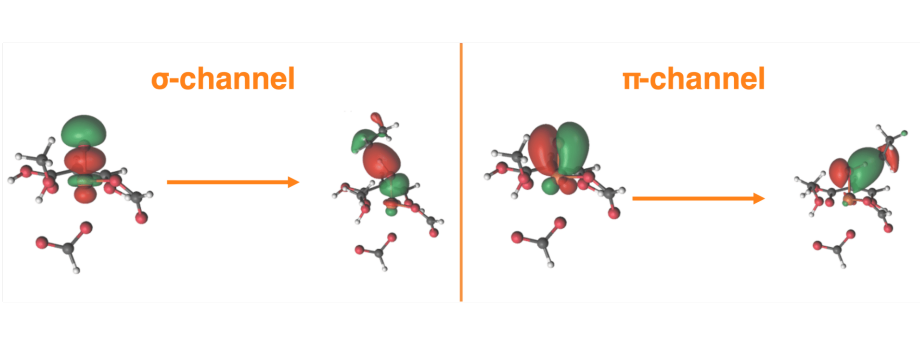
Spectroscopy helps on the characterization of key reaction intermediates and can help on the elucidation of catalytic mechanisms. [22,23]



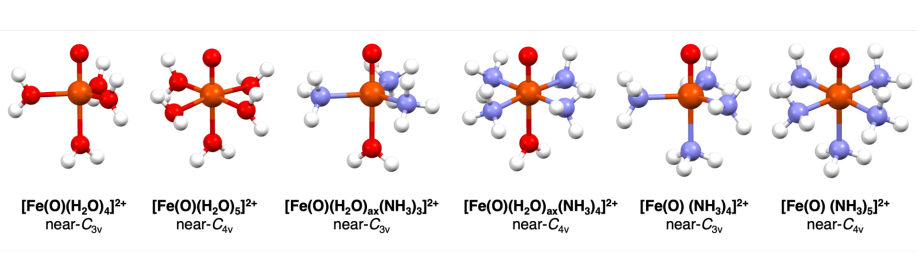
We have provided the theoretical examination on the first study of gas-phase catalytic hydrogenation of olefins in MOFs. [24,25] Our computational work proposed the decoordination of the MOF linker from the catalytically active metal center, which was later proved experimentally with spectroscopical methods. [26]



Excited electronic states can also lead to enhanced chemical reactivity via alternative reaction channels. Electronic spectroscopy has contributed on the elucidation of different reaction pathways for C-H activation via biomimetic Fe(IV)-oxo intermediates.



We have performed multiconfigurational calculations to elucidate the ligand effects on the chemical reactivity of the Fe(IV)-oxo cites. [14] Our results identified that an increasing strength of the ligand sphere decreases the accessibility to multiple reaction channels.


$[Fe(O)(H_2O)_4]^{2+}$ near-C_{3v}
$[Fe(O)(H_2O)_5]^{2+}$ near-C_{4v}
$[Fe(O)(H_2O)_{ax}(NH_3)_3]^{2+}$ near-C_{3v}
$[Fe(O)(H_2O)_{ax}(NH_3)_4]^{2+}$ near-C_{4v}
$[Fe(O)(NH_3)_4]^{2+}$ near-C_{3v}
$[Fe(O)(NH_3)_5]^{2+}$ near-C_{4v}

We have explored larger molecular datasets of Fe(IV)-oxo complexes for C-H activation with interpretable AI workflows that combine ligand topology, quantum chemical data, and kernel ridge regression models. [15] The model analysis allowed us to discover new structure-functionality relationships between molecular scaffolds and chemical reactivity. Ongoing work in our group extends these computational methodologies to **epoxidation** and **aziridination** reactions.

HOW CAN I CONTRIBUTE TO NASA'S MISSION

I am interested in learning about NASA's archived databases with respect to earth sciences, global warming, astrochemistry, and spectroscopy. Based on my expertise on molecular AI, I am confident that I can provide practical solutions through novel AI/ML workflows.

Data is the new Oil! In collaboration with NASA staff scientists, I would like to explore historical, archived data collected through the years for contributing to **informatics for Earth science**, **in-space technologies**, and **infrastructure**. In particular, I can support efforts related to:

- Real-time astrochemical data analysis from telescopes (infrared, mid-infrared) and other remote sensors.
- Climate studies on greenhouse gases and carbon dioxide distribution
- Air pollution
- Chemical analysis and characterization
- Detection of exotic chemical compounds

AUTHOR INFORMATION

Konstantinos "Kostas" Vogiatzis completed his bachelor's degree in chemistry at the University of Athens, Greece, in 2006. In 2008, he obtained his MSc in Applied Molecular Spectroscopy from the University of Crete, Greece, and he received his Ph.D. in 2012 from the Karlsruhe Institute of Technology, Germany. After an eight-month post-doctoral appointment at the Institute of Nanotechnology at the Karlsruhe Institute of Technology, he moved in 2014 at the University of Minnesota, where he performed post-doctoral research at the group of Prof. Laura Gagliardi. In 2016, Dr. Kostas Vogiatzis joined the University of Tennessee, Knoxville, as an Assistant Professor of theoretical and computational chemistry and in 2021 he was early tenured and promoted to Associate Professor. His research group is currently developing new computational methods based on electronic structure theory and machine learning for the theoretical examination of reactivity, catalysis, and separation processes. Kostas is the recipient of the ACS OpenEye Outstanding Junior Faculty Award for Spring 2021, the 2022 NSF CAREER award, the 2020 and 2022 Ffrancon Williams Endowed Faculty Award in Chemistry, and the 2023 Bodossaki Distinguished Young Scientist Award in Sciences.

Kostas Vogiatzis delivering his award seminar at the 2023 Bodossaki Distinguished Young Scientist Award in Sciences ceremony.

TRANSCRIPT

ABSTRACT

My research focuses on the development of **novel AI workflows based on modern machine learning, topological data analysis and computational chemistry**. I have experience working with computational data, as well as archived experimental data from chemical companies (e.g. *Eastman*). I can contribute to NASA's scope by offering the expertise that I have built on data exploration and interpretation, bridge chemistry with AI, as well as utilize databases that go beyond chemistry and material science.

I am interested to learn more about NASA's archived databases with respect to astrochemistry and spectroscopy, as well as how those can be explored. Based on my expertise on molecular AI, I am confident that I can provide practical solutions.

REFERENCES

1. J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas, K. D. Vogiatzis, Representation of Molecular Structures with Persistent Homology leads to the Discovery of Novel CO_2-philic Groups, *Nat. Commun.*, **2020**, *11*, 3230.
2. K. D. Vogiatzis, A. Mavrandonakis, W. Klopper, G. Froudakis, Ab initio Study of the Interactions between CO_2 and N-Containing Organic Heterocycles, *ChemPhysChem*, **2009**, *10*, 374.
3. A. G. Sylvanus, K. D. Vogiatzis, Accurate Interaction Energies of CO_2 with the 20 Naturally Occurring Amino Acids, *ChemPhysChem*, **2023**, *24*, e202300027.
4. A. G. Sylvanus, G. M. Jones, R. Custelcean, K. D. Vogiatzis, *In Silico* Screening of CO_2-Dipeptide Interactions for Biomimetic Carbon Capture, *Submitted*.
5. P.-F. Cao, B. Li, T. Hong, J. Townsend, Z. Qiang, K. Xing, K. D. Vogiatzis, Y. Wang, J. W. Mays, A. P. Sokolov, T. Saito, Super-Stretchable Polymeric Elastomers with Healable Gas Separation Capability, *Adv. Funct. Mater.*, **2018**, *28*, 1800741.
6. C. R. Maroon, J. Townsend, K. R. Gmernicki, D. J. Harrigan, B. J. Sundell, J. A. Lawrence III, S. M. Mahurin, K. D. Vogiatzis, B. K. Long, Elimination of CO_2/N_2 Langmuir Sorption and Promotion of "N2-phobicity" within High T_g, Glassy Membranes, *Macromolecules*, **2019**, *52*, 1589.
7. C. R. Maroon, J. Townsend, M. A. Higgins, D. J. Harrigan, B. J. Sundell, J. A. Lawrence III, J. T. O'Brien, D. O'Neal, K. D. Vogiatzis, B. K. Long, Addition-type Alkoxysilyl Substituted Polynorbornenes for Post-Combustion Carbon Dioxide Separations, *J. Membr. Sci.*, **2020**, *595*, 117532.
8. M. A. Higgins, C. R. Maroon, J. Townsend, K. D. Vogiatzis, B. K. Long, Evaluating the impact of functional groups on membrane-mediated CO_2/N_2 gas separations using a common polymer backbone, *J. Polym. Sci.*, **2020**, *58*, 2644.
9. X. Wang, T. J. Wilson, C. R. Maroon, J. Laub, S. Rheingold, K. D. Vogiatzis, B. K. Long, Fluorinated Vinyl-addition Polynorbornene Membranes for CO2/CH4 Separation, *ACS Appl. Polym. Mater.*, **2022**, *4*, 7976.
10. O. Queen, G. A. McCarver, S. Thatigotla, B. P. Abolins, C. L. Brown, V. Maroulas, K. D. Vogiatzis, Polymer Graph Neural Networks for Multitask Property Learning, *npj Comput. Mater.*, **2023**, *9*, 90.
11. P. Verma, K. D. Vogiatzis, N Planas, J. Borycz, D. J. Xiao, J. R. Long, L. Gagliardi, D. G. Truhlar, Mechanism of Oxidation of Ethane to Ethanol at Iron(IV)–Oxo Sites in Magnesium-Diluted Fe_2(dobdc), *J. Am. Chem. Soc.*, **2015**, *137*, 5770.
12. K. D. Vogiatzis, E. Haldoupis, D. J. Xiao, J. R. Long, J. I. Siepmann, L. Gagliardi, Accelerated Computational Analysis of Metal–Organic Frameworks for Oxidation Catalysis, *J. Phys. Chem. C*, **2016**, *120*, 18707.
13. K. D. Vogiatzis, G. Li, E. J. M. Hensen, L. Gagliardi, E. A. Pidko,* The Electronic Structure of the [Cu3(μ-O)3]2+ Cluster in Mordenite Zeolite and its Effects on the Methane to Methanol Oxidation, *J. Phys. Chem. C*, **2017**, *121*, 22295.
14. J. K. Kirkland, S. N. Khan, B. Casale, E. Miliordos, K. D. Vogiatzis, Ligand Field Effects on the Ground and Excited States of the Catalytically Active FeO2+ Species, *Phys. Chem. Chem. Phys.*, **2018**, *20*, 28786.
15. G. M. Jones, B. A. Smith, J. K. Kirkland, K. D. Vogiatzis, Data-Driven Ligand Field Exploration of Fe(IV)-oxo Sites for C-H Activation, *Inorg. Chem. Front.*, **2023**, *10*, 1062.
16. K. D. Vogiatzis, D. Ma, J. Olsen, L. Gagliardi, W. de Jong, Pushing Configuration-Interaction to the Limit: Towards Massively Parallel MCSCF Calculations, *J. Chem. Phys.*, **2017**, *147*, 184111.
17. G. M. Jones, P. D. V. S. Pathirage, K. D. Vogiatzis, Data–Driven Acceleration of Coupled–Cluster and Perturbation Theory Methods in *Quantum Chemistry in the Age of Machine Learning*, Elsevier. Edited by P. Dral, **2022**.
18. J. Townsend, K. D. Vogiatzis, Data-driven Acceleration of the CCSD Iterative Solver, *J. Phys. Chem. Lett.*, **2019**, *10*, 4129.
19. P. D. V. S. Pathirage, J. T. Phillips, K. D. Vogiatzis, Exploration of the Two-electron Excitation Space with Data-driven Coupled-cluster, *J. Phys. Chem. A*, **2024**, *128*, 1938.
20. G. M. Jones, R. R. Li, A. E. DePrince III, K. D. Vogiatzis, Data-driven Refinement of Electronic Energies from Two-electron Reduced-density-matrix Theory, *J. Phys. Chem. Lett.*, **2023**, *14*, 6377.
21. G. M. Jones, B. Story, V. Maroulas, K. D. Vogiatzis, Molecular Representations for Machine Learning, ACS In Focus, **2023**.
22. K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu, E. A. Pidko, Computational Approach to Catalysis by 3d Transition Metals: Challenges and Opportunities, *Chem. Rev.*, **2019**, *119*, 2453.
23. G. A. McCarver, T. Rajeshkumar, K. D. Vogiatzis, Computational Catalysis for Metal-Organic Frameworks: An Overview, *Coord. Chem. Rev.*, **2021**, *436*, 213777.
24. D. M. Shakya, O. Ejegbavwo, T. Rajeshkumar, S. D. Senanayake, A. J. Brandt, S. Farzandh, N. Acharya, A. Ebrahim, A. I. Frenkel, N. Rui, G. L. Tate, J. R. Monnier, K. D. Vogiatzis, N. B. Shustova, D. A. Chen, Selective Catalytic Chemistry at Isolated Rh(II) Nodes in Bimetallic Metal-Organic Frameworks, *Angew. Chemie Int. Ed.*, **2019**, *56*, 16533.
25. K. Metavarayuth, O. Ejegbavwo, G. A. McCarver, M. Myrick, T. Makris, K. D. Vogiatzis, S. Senanayake, O. Manley, A. Ebrahim, A. Frenkel, S. Hwang, T. Rajeshkumar, J. Jimenez, K. Chen, N. Shustova, D. Chen, Direct Identification of Mixed-Metal Centers in Metal-Organic Frameworks: $Cu_3(BTC)_2$ Transmetallated with Rh^{2+} Ions, *J. Phys. Chem. Lett.*, **2020**, *11*, 8138.
26. D. A. Chen, J. D. Jimenez, S. D. Senanayake, J. P. Stetzler, D. M. Shakya, G. A. McCarver, T. Rajeshkumar, K. D. Vogiatzis, A. Mathur, N. B. Shustova, M. L. Myrick, K. Metavarayuth, M. M. Royko, J. Lauterbach, G. L. Tate, J. R. Monnier, Mechanistic Investigations of Gas-Phase Catalytic Hydrogenation in Metal-Organic Frameworks: Cooperative Activity of the Metal and Linker Sites in $CuxRh_{3-x}(BTC)_2$, *J. Phys. Chem. C*, **2022**, *126*, 11553.