

Using Data Science to Rationalize and Design Catalysts from Small Experimental Data Sets

Dr. Luigi Cavallo

King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Jueves, 29 de octubre 2026

12:00h

Sala de grados - Edificio A - Facultad Ciencias



CICLO CONFERENCIAS ISQCH 2026



Instituto de Síntesis Química y Catálisis Homogénea



MARCO DE
AUTOEVALUACIÓN DE LA
EXCELENCIA



CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



Universidad
Zaragoza



Facultad de Ciencias
Universidad Zaragoza

Facultad de Ciencias, Universidad de Zaragoza - CSIC
C/ Pedro Cerbuna, 12. Zaragoza 50009. Spain

Using Data Science to Rationalize and Design Catalysts from Small Experimental Data Sets

Luigi Cavallo

*King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
email: luigi.cavallo@kaust.edu.sa*

The development of new coordination catalysts increasingly relies on the integration of computational chemistry and data-science methodologies. While machine learning has shown significant promise in accelerating catalyst discovery, most existing approaches rely on large and highly curated datasets that are rarely available in experimental catalysis. In practice, catalyst optimization campaigns typically produce fewer than 50 experimentally characterized catalysts, often with heterogeneous conditions and incomplete structure–activity relationships. Extracting reliable predictive models from such limited data remains a major challenge.

In this presentation, we explore data-science strategies specifically designed for small experimental datasets in homogeneous catalysis. Our approach combines computational descriptors derived from electronic structure calculations with statistical learning methods that remain robust in the low-data regime. Rather than aiming for purely predictive models, we focus on interpretable structure–activity relationships that can rationalize observed experimental trends and guide catalyst design.

Starting from experimentally reported catalytic systems, typically consisting of fewer than 50 data points, we construct descriptor sets capturing steric, electronic, and geometric features of coordination complexes. These descriptors are then analyzed using multivariate linear regression and other low-complexity machine-learning models to avoid overfitting while maintaining interpretability. Particular attention is given to feature selection, uncertainty estimation, and cross-validation strategies appropriate for extremely small datasets.

This framework allows us to identify the molecular features most strongly correlated with catalytic performance and to propose new catalyst candidates that can be experimentally tested. In some case studies, the models provide mechanistic insights into the role of ligand environment and metal–ligand interactions in determining catalytic activity and selectivity.

Concluding, this work highlights how data-science tools can be adapted to the realities of experimental catalysis, where data are scarce but chemically rich. By integrating computational descriptors with interpretable machine learning, we show that meaningful catalyst design strategies can emerge even from very limited datasets.



Webpage: <https://www.kaust.edu.sa/en/study/faculty/luigi-cavallo>

Prof. Luigi Cavallo is a Professor of Chemistry at **King Abdullah University of Science and Technology (KAUST)** and Director of the Computational Chemistry Laboratory. His research focuses on computational chemistry and catalysis, with particular expertise in organometallic chemistry, olefin polymerization, olefin metathesis, N-heterocyclic carbenes (NHCs), and photoredox catalysis.

Prof. Cavallo earned his Ph.D. in Chemistry from the University of Naples (Italy) in 1991 and further developed his expertise in density functional theory (DFT) through collaborations with Prof. Tom Ziegler at the University of Calgary. Before joining KAUST, he held academic positions at the University of Naples and the University of Salerno.

Internationally recognized for his contributions to computational catalysis, he has authored more than 500 scientific publications with over 60000 citations (being a Highly Cited Researcher), and developed widely used molecular descriptors and computational tools for catalyst design, including influential methodologies such as steric maps and %V_{Bur} index. His work has significantly advanced the understanding of catalyst structure-reactivity relationships and computer-assisted catalyst development. Through his research, mentorship, and collaborations, Prof. Cavallo got several prestigious prizes and has become one of the leading figures in computational organometallic chemistry and catalysis, helping bridge theoretical chemistry, machine learning, and experimental catalyst design.