doi:10.2533/chimia.2022.346 Chimia 76 (2022) 346–349

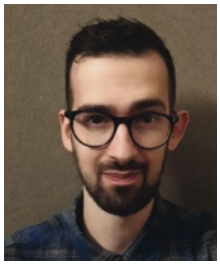
HTE and Data Analysis for Discovery and Molecular-level Understanding of Catalysts

Jordan De Jesus Silva[§*] and Christophe Copéret

[§]SCS-DSM award for best virtual poster in Organic Chemistry

Abstract: The combination of high-throughput experimentation (HTE) and data analysis is a valuable methodology for mechanistic interrogation and rational development of catalysts. In this article, we point out the general structure of HTE-data analysis workflow and illustrate how it can be applied with examples of olefin metathesis and cyanation reactions.

Keywords: Catalysis · Cyanation · Data analysis ·High-throughput experimentation · Olefin metathesis



Jordan De Jesus Silva was awarded a Bachelor's and Master's degree in Chemistry at ETH Zurich in 2016 and 2018, respectively. Recipient of a AFR Individual PhD Grant of the Luxembourg National Research Fund, he is currently pursuing a PhD degree in the group of Prof. Christophe Copéret at ETH Zurich, focusing on the combination of high-throughput experimentation and data analysis to investigate known and develop new reactions.

1. Introduction

Catalysis is at the heart of efficient chemical processes and is directly associated with sustainable development, by lowering energy consumption while optimizing resources. Furthermore, it will also provide a way to transition from fossil energy and chemical resources to renewables.[1] In industrial settings, heterogeneous catalysts are essential as they allow process intensification (decrease of energy intensive steps, typically associated with separation and regeneration), but they suffer from their intrinsic complexity. Hence, they are developed mostly *via* empirical approaches. In some cases, they can be replaced by homogenous catalysts, which are powerful alternatives due to their often higher selectivity, lower operational temperatures and easier rational developments. For the latter, the operations (recycling) are far more complex and often require years of development. Overall, catalysis, whether homogenous or heterogeneous, requires tedious optimization due to the large parameter space (concentration, temperature, pressure, additives…) that influences catalytic performance (activity, selectivity, and stability). In that context, one approach to speed up developments, that embraces the complexity of catalysis, is high-throughput experimentation (HTE),[2] which has emerged in the late 1990s and has gained momentum more recently, in particular with the emergence of improved data analysis and machine learning-based approaches.[3] For instance, laboratory robotics simultaneously perform multiple tasks that enable time-efficient screening of catalysts[4] in a broad range of well-defined conditions, thus generating large and reliable catalytic data sets. However, finding the underlying rationalization behind the success of a particular catalyst formulation remains a formidable challenge. At the opposite end, computational chemistry, in particular, based on density functional theory (DFT), has demonstrated its power to describe reaction pathways and to rationalize reactivity and selectivity patterns but at the expense of long and tedious work. More recently, structure–activity studies based on multivariate linear regression analyses have demonstrated their efficiency in identifying – in a more timely and cost-effective way – promising correlations with predictive power.[5] In this article, we describe how combining the efficiency of HTE methods with data analysis *via* multivariate linear regression fitting of catalytic results and computational rationalization of data allows for computer-guided prediction of catalysis research and rational design. Furthermore, we discuss how, in this context, the emergence of machine learning is yet offering new possibilities, which could revolutionize catalyst design and process implementation.

2. Methodology

The general HTE-data analysis workflow is shown in Fig. 1. The approach described in this article is best for the development of catalytic systems where organic ligands and additives are used to modulate the catalytic performance and one is looking to discover and optimize the catalyst structure or formulation, *e.g.* well-defined and ill-defined molecular and supported catalysts including nanoparticles where the organic ligands can play a major role.

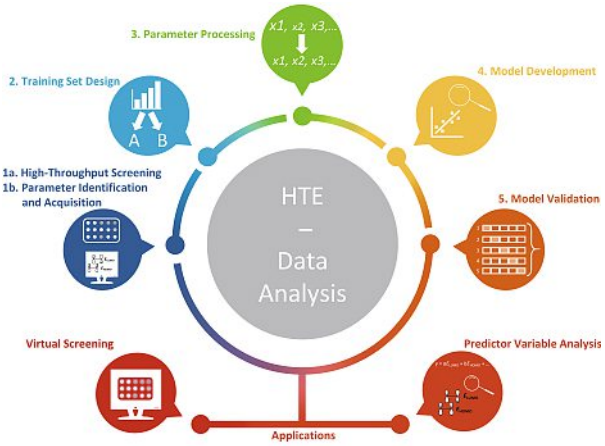


Fig. 1. HTE-data analysis workflow.

*Correspondence: J. De Jesus Silva, E-mail: jordan.silva94@live.fr
Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1-5, CH–8093 Zurich, Switzerland

Once the system under study is defined, high-throughput screening should be used to perform catalyst evaluation in a time efficient and reproducible way (step 1a). Based on analysis of proposed mechanisms and catalyst structure, suitable ligand descriptors should be identified and validated (step 1b). For the construction of generalizable, unbiased models, which are aimed at making accurate predictions for a wide range of different molecules, the gathered experimental data (TOF, TON, yield, *etc.*) should be divided in a training set, used for model construction, and an external validation set, necessary for the verification of the models (step 2). The descriptors have to be normalized to possess the same scale and deviation, so that the coefficient in future models reflects the variance of each parameter (step 3). First univariate correlations are done to see which ligand descriptors are most relevant for catalysis and to identify possible data subsets of structurally related ligands. Consequently, preliminary multivariate models can be constructed, *e.g.* by least-squares linear regression by forward feature selection, effectively evaluating the change in statistics caused by addition/removal of each parameter and incorporation of the most important term in each step (step 4). The generated models have to be validated by (internal) cross-validation like Q2 or k-fold means, or by external validation with empirical results that are known before model development (step 5). The goal of this workflow is to gain a better understanding of reaction mechanisms through analysis of interactions that are described by ligand parameters, but also to predict new active catalysts *via* extrapolation in concert with virtual screening. The application of this workflow will be discussed in two case studies. In the first study, the methodology was applied to a well-studied reaction, olefin metathesis, in order to investigate the key descriptors that drive the catalysis for both homogeneous and heterogeneous systems. The second example focuses on the development of a new cyanation protocol and investigation of the optimal ligand properties for the design of improved palladium cross-coupling catalysts.

3. Case Studies

3.1 *Olefin Metathesis*

Olefin metathesis is a prototypical example of an (atom) efficient reaction catalyzed by group 6 metals, with a broad industrial interest, ranging from petrochemicals, polymers to the fine chemical industry. Olefin metathesis, a Nobel-prize-winning technology,[6] is used to produce propene, an essential component of polymers, *via* the OCT process (WO_3/SiO_2), long chain olefins *via* the SHOP (MoO_3/Al_2O_3), biomass-derived oils or complex pheromones and drugs (Molecular Mo or W Schrock-type catalysts).[7] Over the last decades, a multitude of catalysts have been synthesized by a serendipity driven approach[8] and their reactivity was rationalized by computational studies.[9] Further understanding of the key parameters driving alkene metathesis and the deactivation pathways still remains a challenging task. Towards this goal, libraries of homogeneous and heterogeneous Schrock-type olefin metathesis catalysts were efficiently synthesized using high-throughput experimentation, specifically by using bis-pyrrolide type Mo alkylidene molecular complexes as ideal candidates due to their ease of synthesis and modularity.[10] In fact, the pyrrolido ligand can readily be exchanged *via* protonolysis with a XH molecule *e.g.* X = aryloxides or even silica as a support, so that over 200 formulations were readily prepared from 35 selected phenols and with/without silica partially dehydroxylated at 700 °C (SiO_{2-700}).[11] In parallel, density functional theory (DFT) calculations on the phenolic ligands were used to acquire simple steric and electronic molecular descriptors to correlate to the anticipated reaction outputs (Fig. 2A, left). Testing the *in situ* generated complexes in the homometathesis of 1-nonene in a robotized way enabled the monitoring of the reaction progress at different time points by retrieval and successive gas chromatography analysis of reaction aliquots. Analysis of the raw data allowed for extraction of conversions, product selectivities, as well as respective turnover numbers (TON) and turnover frequencies (TOF) as catalytic output descriptors (Fig. 2A, right). After parameter processing, subsequent identification of univariate correlations highlighted the non-anticipated importance of splitting the phenolic data set in two subgroups differing by the presence/absence of aryl substituents in *ortho* positions of the respective phenol ligands, drastically improving individual univariate correlations with several computed descriptors (Fig. 2B). Multivariate linear regression analysis was then utilized to obtain internally validated, predictive models that portray the impact of the interplay of stereo-electronic effects of the ligands on TOF and TON responses for both groups (Fig. 2C). The resulting models captured the well-established importance of the σ–donation ability of the ligand in modulating the activity of the catalysts,[12] which increased the confidence in the meaningfulness of the analysis. More importantly though, the models uncovered the influence of non-covalent interactions in tuning activity and performance, in particular for aryl-arm bearing phenolic ligands, hence providing a new lever that may be exploited for the future design of improved d^0 metathesis catalysts.

3.2 *Cyanation*

Nitriles are important structural motifs in pharmaceuticals and natural products[13] and their cyano moiety serves as a valuable precursor for numerous functional group interconversions.[14] Ample efforts have been put in developing catalytic cyanation protocols, for instance employing nucleophilic or electrophilic cyano sources.[15] While a major concern remains the toxicity of the employed cyanation reagents, the choice of the optimal reaction conditions, and in particular of the ligand to catalyze the reaction, is often not clear. In this regard, the HTE-data analysis methodology was applied to develop a novel palladium-catalyzed electrophilic cyanation protocol, opting for classic Suzuki-Miyaura cross-coupling conditions, using aryl boronic acids and *N*-cyano succinimide as cyanating agent (Fig. 3).[16] Accelerated investigation of the ligand effect was automated using a liquid handling robot to screen 90 ligands belonging to either monophosphine, bisphosphine or the miscellaneous subgroup. All tests were hereby performed in triplicate to assess the reproducibility of the results, yielding 288 formulations to analyze via gas chromatography (Fig. 3A). Similar to what was described in Section 3.1, the workflow involved calculating DFT-derived ligand descriptors to relate the electronic and steric properties of the ligands to the experimentally determined yield for the mono- and bisphosphine subsets. For the bisphosphine subset, however, ligand parameters specific to the $PdCl_2$ adduct were assessed additionally to describe the bidentate nature of the ligands. Relying on univariate and multivariate linear regression analysis, structurally-responsive ligand behavior[17] was identified as the main characteristic required in an optimal ligand, displaying the ability to stabilize the metal in their bisligated state while their hemilability is able to open up a coordination site that may be needed to enable catalysis (Fig. 3B). XantPhos turned out to excel in this regard and was further used as ligand to investigate the protocol for different substrates, demonstrating excellent functional group tolerance in particular with electron-withdrawing moiety bearing aryl boronic acids (Fig. 3C).

4. Outlook: Beyond Multivariate Linear Regression

In recent years, tremendous progress in the area of machine learning (ML) and artificial intelligence has facilitated the implementation of algorithms for non-specialists.[18] The multidimensionality of chemical space complicates the use of such algorithms for the synthetic community, given the requirements for large amount of data to efficiently navigate that space. HTE has been of outmost importance in unlocking the potential of ML in

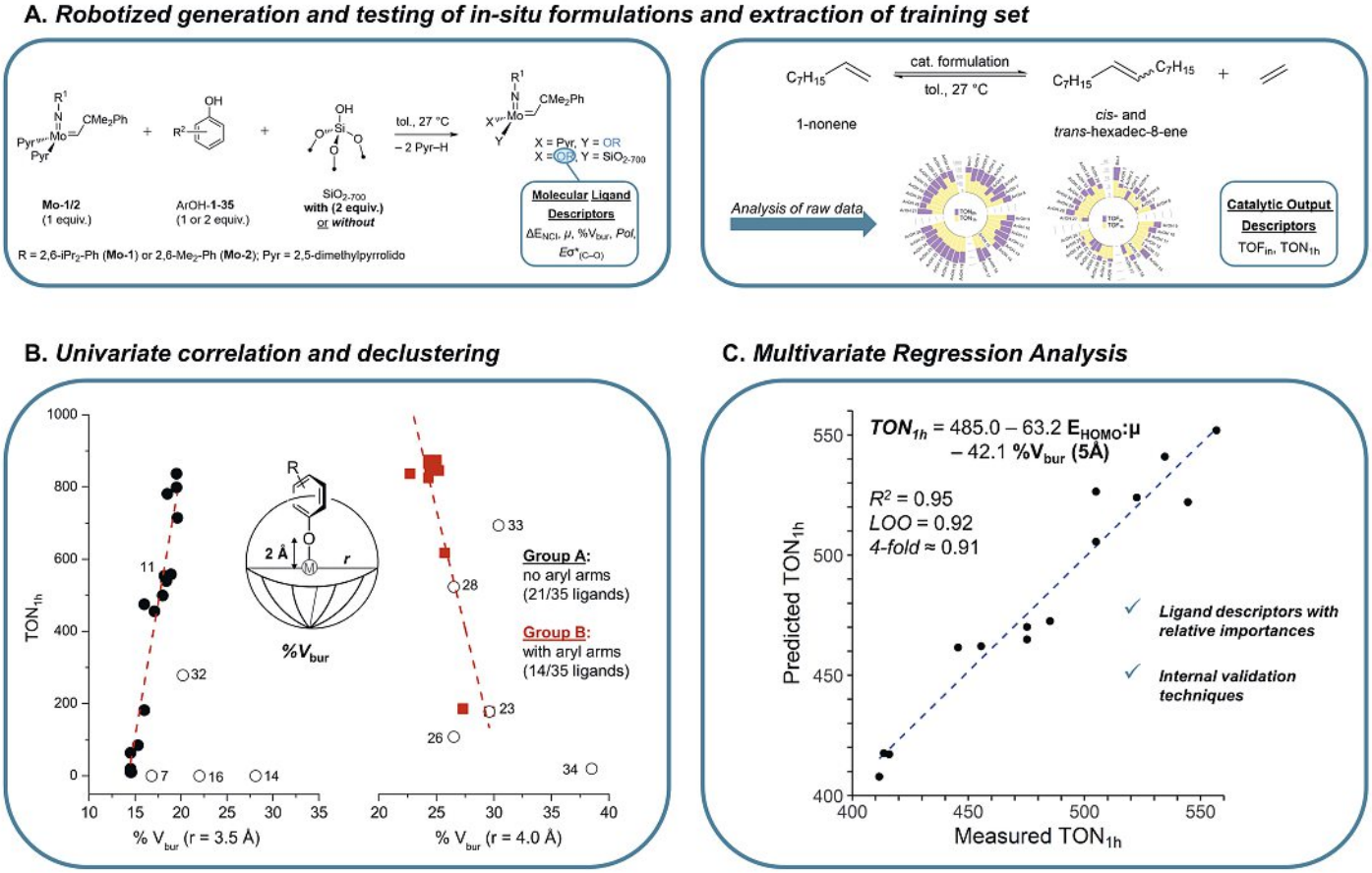


Fig. 2. Olefin metathesis case study. Investigation of ligand effects for homogeneous and surface organometallic chemistry derived d0 olefin metathesis catalyst enabled by the HTE-data analysis approach.

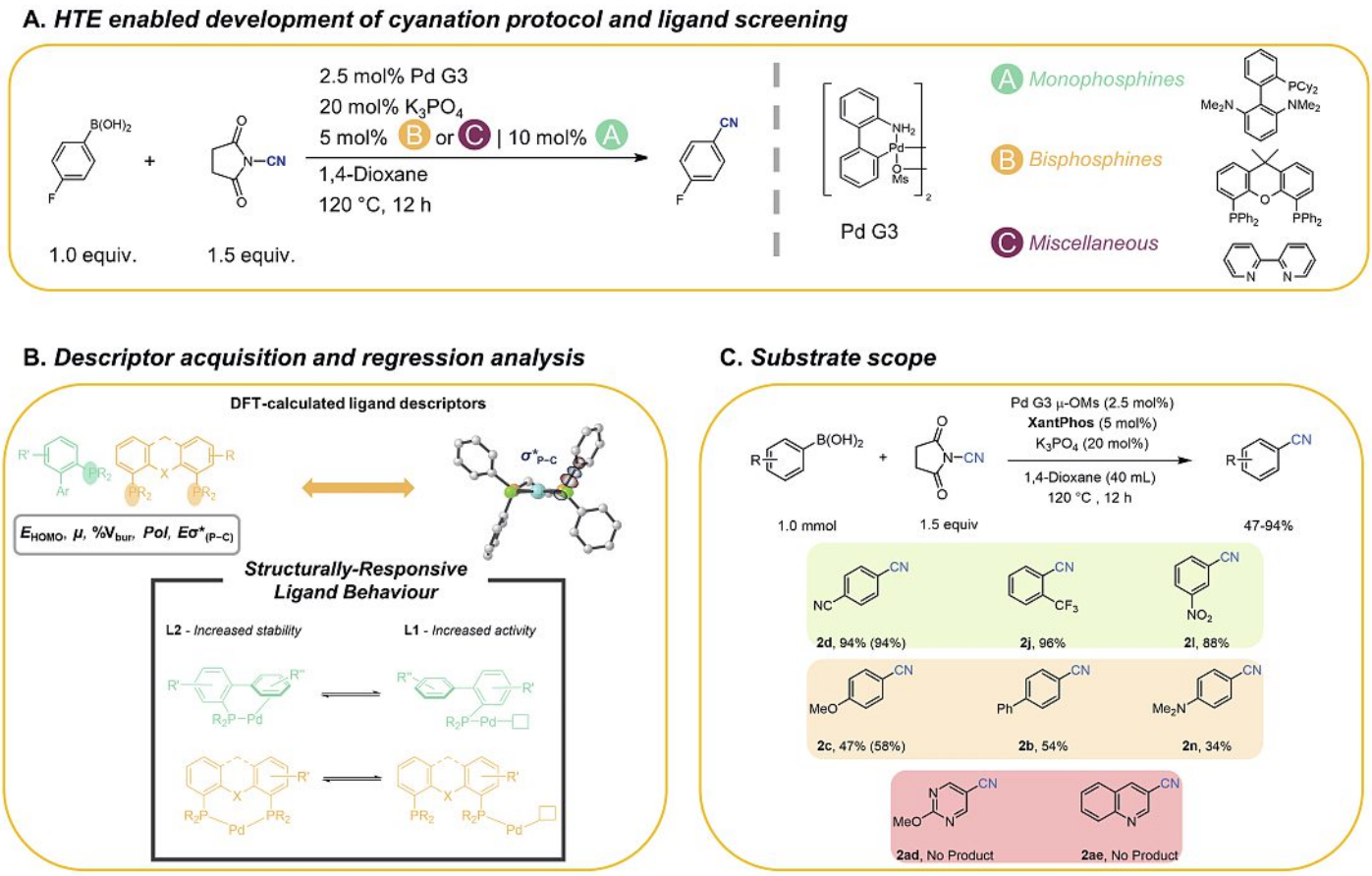


Fig. 3. Cyanation case study. HTE-enabled development of cyanation protocol and investigation of optimal ligand properties.

Chemistry. The strive of data-driven strategies to the chemical sciences has since known a clear uptick in applications[19] in both homogeneous and heterogeneous catalysis.[20] The power of machine learning resides in its pattern recognition and self-learning without explicit tailored programming. In this regard, the combination of ML with HTE opens up new avenues towards self-driven laboratories by autonomously planning, executing and processing reactions.[21]

5. Conclusion

In summary, the use of high-throughput experimentation in concert with data analysis was demonstrated to be effective towards mechanistic interrogation and accelerated reaction development. While molecularly well-defined systems were explored here, this approach is applicable to a broader range of catalyst classes, such as supported or unsupported nanoparticles, where ligand additives can play a major role. Such approaches are currently under investigation in our group. While HTE ensures time-efficient execution of synthetic steps and reproducibility of the performed reactions, data analysis, and in particular, simple multivariate linear regression models appeal through their ease of use and interpretability. Exciting developments of ML-based methods are yet offering new possibilities to take this methodology to the next level.

Acknowledgements

J.D.J.S. was supported by the National Research Fund, Luxembourg (AFR Individual Ph.D. Grant 12516655). We wish to thank Marco A. B. Ferreira, Matthew S. Sigman, Antonio Togni, as well as Niccolò Bartalucci, Alexey Fedorov, and Benson Jelier, for many discussions and their involvement in some of the original works (see references below).

Received: January 31, 2022

[1] J. Gong, R. Luque, *Chem. Soc. Rev*. **2014**, *43*, 7466, https://doi.org/10.1039/C4CS90084G.
[2] A. Gordillo, S. Titlbach, C. Futter, M. L. Lejkowski, E. Prasetyo, L. T. Alvarado Rupflin, T. Emmert, S. A. Schunk, 'High-Throughput Experimentation in Catalysis and Materials Science', Wiley-VCH, Weinheim, **2014**.
[3] C. M. Bishop, 'Pattern recognition and machine learning', Springer New York, New York, **2006**.
[4] D. R. Romer, V. J. Sussman, K. Burdett, Y. Chen, K. J. Miller, *ACS Comb. Sci*. **2014**, *16*, 551, https://doi.org/10.1021/co500087b.
[5] C. B. Santiago, J. Y. Guo, M. S. Sigman, *Chem. Sci*. **2018**, *9*, 2398, https://doi.org/10.1039/C7SC04679K.
[6] a) Y. Chauvin, *Angew. Chem. Int. Ed*. **2006**, *45*, 3740, https://doi.org/10.1002/anie.200601234; b) R. H. Grubbs, *Angew. Chem. Int. Ed*. **2006**, *45*, 3760, https://doi.org/10.1002/anie.200600680; c) R. R. Schrock, *Angew. Chem. Int. Ed.* **2006**, *45*, 3748, https://doi.org/10.1002/anie.200600085.
[7] K. J. Ivin, J. C. Mol, 'Olefin Metathesis and Metathesis Polymerization', Elsevier Science, Oxford, **2014**.
[8] M. J. Benedikter, F. Ziegler, J. Groos, P. M. Hauser, R. Schowner, M. R. Buchmeiser, *Coord. Chem. Rev*. **2020**, *415*, 213315, https://doi.org/10.1016/j.ccr.2020.213315.
[9] C. Copéret, Z. J. Berkson, K. W. Chan, J. de Jesus Silva, C. P. Gordon, M. Pucino, P. A. Zhizhko, *Chem. Sci*. **2021**, *12*, 3092, https://doi.org/10.1039/D0SC06880B.
[10] A. S. Hock, R. R. Schrock, A. H. Hoveyda, *J. Am. Chem. Soc*. **2006**, *128*, 16373, https://doi.org/10.1021/ja0665904.
[11] a) M. A. B. Ferreira, J. De Jesus Silva, S. Grosslight, A. Fedorov, M. S. Sigman, C. Copéret, *J. Am. Chem. Soc.* **2019**, *141*, 10788, https://doi.org/10.1021/jacs.9b04367; b) J. De Jesus Silva, M. A. B. Ferreira, A. Fedorov, M. S. Sigman, C. Copéret, *Chem. Sci*. **2020**, *11*, 6717, https://doi.org/10.1039/D0SC02594A.
[12] X. Solans-Monfort, C. Copéret, O. Eisenstein, *Organometallics* **2012**, *31*, 6812, https://doi.org/10.1021/om300576r.
[13] F. F. Fleming, L. Yao, P. C. Ravikumar, L. Funk, B. C. Shook, *J. Med. Chem*. **2010**, *53*, 7902, https://doi.org/10.1021/jm100762r.
[14] R. C. Larock, 'Comprehensive organic transformations: a guide to functional group preparations', Wiley-VCH, New York; Chichester, **1997**.
[15] a) Y. J. Z. L. Yan Guobing, *Chinese J. Org. Chem*. **2012**, *32*, 294, https://doi.org/10.6023/cjoc201802007; b) G. Yan, Y. Zhang, J. Wang, *Adv. Synth. Catal.* **2017**, *359*, 4068, https://doi.org/10.1002/adsc.201700875; c) S. Pimparkar, A. Koodan, S. Maiti, N. S. Ahmed, M. M. M. Mostafa, D. Maiti, *ChemComm* **2021**, *57*, 2210, https://doi.org/10.1039/D0CC07783F.
[16] J. De Jesus Silva, N. Bartalucci, B. Jelier, S. Grosslight, T. Gensch, C. Schünemann, B. Müller, P. C. J. Kamer, C. Copéret, M. S. Sigman, A. Togni, *Helv. Chim. Acta* **2021**, *104*, e2100200, https://doi.org/10.1002/hlca.202100200.
[17] J. M. Blacquiere, *ACS Catalysis* **2021**, 5416, https://doi.org/10.1021/acscatal.1c00613.
[18] a) G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem*. **2017**, *38*, 1291, https://doi.org/10.1002/jcc.24764; b) X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, *Chem. Rev*. **2019**, *119*, 10520, https://doi.org/10.1021/acs.chemrev.8b00728.
[19] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Cent. Sci*. **2021**, *7*, 1622, https://doi.org/10.1021/acscentsci.1c00535.
[20] a) T. Ahneman Derek, G. Estrada Jesús, S. Lin, D. Dreher Spencer, G. Doyle Abigail, *Science* **2018**, *360*, 186, https://doi.org/10.1126/science.aar5169; b) J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang, S. E. Denmark, *J. Am. Chem. Soc*. **2020**, *142*, 11578, https://doi.org/10.1021/jacs.0c04715; c) X. Li, P. M. Maffettone, Y. Che, T. Liu, L. Chen, A. I. Cooper, *Chem. Sci*. **2021**, *12*, 10742, https://doi.org/10.1039/D1SC02150H; d) K. McCullough, T. Williams, K. Mingle, P. Jamshidi, J. Lauterbach, *Phys. Chem. Chem. Phys*. **2020**, *22*, 11174, https://doi.org/10.1039/D0CP00972E.
[21] a) M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, *Commun. Chem*. **2021**, *4*, 112, https://doi.org/10.1038/s42004-021-00550-x; b) N. S. Eyke, B. A. Koscher, K. F. Jensen, *Trends Chem.* **2021**, *3*, 120, https://doi.org/10.1016/j.trechm.2020.12.001.