doi:10.2533/chimia.2023.48 Chimia 77 (2023) 48–55

Can AI Help Improve Water Quality? Towards the Prediction of Degradation of Micropollutants in Wastewater

Hiroko Satoh[a], Jasmin Hafner[b], Jürg Hutter[a], and Kathrin Fenner*[ab]

Abstract: Micropollutants have become a serious environmental problem by threatening ecosystems and the quality of drinking water. This account investigates if advanced AI can be used to find solutions for this problem. We review background, the challenges involved, and the current state-of-the-art of quantitative structure–biodegradation relationships (QSBR). We report on recent progress combining experiment, quantum chemistry (QC) and chemoinformatics, and provide a perspective on potential future uses of AI technology to help improve water quality.

Keywords: Chemoinformatics · Degradation of micropollutants in wastewater · Quantitative structure–biodegradation relationships (QSBR) · Quantum chemistry (QC)

Hiroko Satoh received her PhD degree in chemistry from Ochanomizu University in 1996. After a Postdoctoral Fellow at RIKEN, she was appointed as assistant professor in 2000 and promoted to associate professor in 2002 at National Institute of Informatics (NII), Japan. She conducted her project on data-driven chemical reaction prediction under a PRESTO program of Japan Science Technology Agency (JST) in 1998–2001. In 2015 she moved to University of Zurich and was concurrently appointed as associate professor at Research Organization of Information and Systems (ROIS), Japan. Her research interests cover a broad range of development and applications of computational and data-driven chemistry.

Jasmin Hafner studied biology at the University of Lausanne and received her PhD degree from the Swiss Federal Institute of Technology in Lausanne (EPFL) in 2020. She is now a post-doc at the Environmental Chemistry Department of the Swiss Federal Institute of Aquatic Science and Technology (Eawag). Her research is focused on computational modelling of biodegradation processes of micropollutants.

Jürg Hutter studied Molecular Sciences at ETH Zurich and received his Doctoral degree from University of Zurich in 1988. In 1990 he spent a year as Postdoctoral Fellow at Ecole Polytéchnique in Paris. Until 1993 he worked at the Center for Supercomputing of ETH Zurich. He was a Research Staff Member first at the IBM Research Laboratory in Rüschlikon and then at the Max-Planck Institute for Solid State Research in Stuttgart. In 2000 he became assistant professor at University of Zurich. In 2004 he was appointed to associate professor and in 2009 promoted to full professor for Physical Chemistry at University of Zurich.

Kathrin Fenner studied chemistry at the University of Zurich and received her Doctoral degree from ETH Zurich in 2001. Nowadays, she is Professor for Environmental Chemistry at the Chemistry Department of the University of Zurich and Group Leader at the Environmental Chemistry Department of the Swiss Federal Institute of Aquatic Science and Technology (Eawag). In her research, she is interested in the principles of microbial biotransformation of chemicals in the environment, with the goal to predict and minimize environmental persistence of chemicals. Kathrin chairs the Section Chemistry and the Environment of the Swiss Chemical Society.

1. Introduction

A general process cycle for developing new molecules includes molecular design, synthetic design, reaction prediction, synthesis (experiment), structure determination (experiment) and assessment of functions suggesting new molecular design for the next cycle (Fig. 1). The prime target is to design molecular structures that have desired functions, such as pharmacological or pesticidal activity. Synthetic accessibility is another important evaluation criteria. However, to design an optimal molecule, one should consider not only these *positive* properties, but also *negative* properties of those molecules, like toxicity and persistency in the environment. Since people have become aware of the serious environmental problems caused by micropollutants, demand to consider the burden on the environment in chemical development has steeply increased.

In the last decade, artificial intelligence (AI) and machine learning (ML) technology has rapidly conquered a wide variety of domains including natural science. Chemistry's history working with AI and ML indeed goes back to the 1960s,[1–3] *i.e.*, to the

*Correspondence: Prof. Dr. K. Fenner[ab], E-mail: Kathrin.Fenner@eawag.ch
[a]Department of Chemistry, University of Zurich; [b]Department of Environmental Chemistry, Eawag, Überlandstrasse 133, CH-8600 Dübendorf

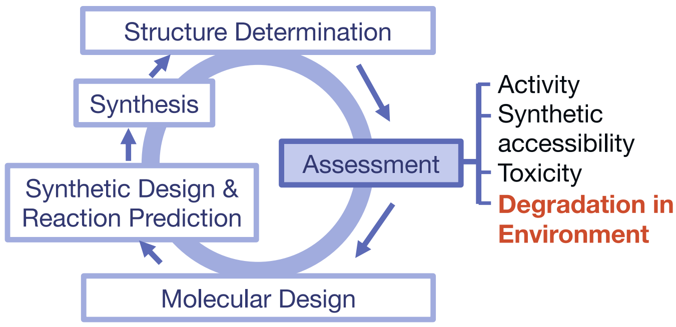


Fig. 1. A general process cycle for developing new molecules. Molecular design needs to consider not only positive functions but also negative ones, including persistency in environment.

research fields called chemometrics and chemoinformatics. This is largely due to the characteristics of chemistry, being a discipline profoundly depending on data (*e.g.*, molecular, reaction and spectral data) in the development of chemical knowledge and theory.

Environmental chemistry has been one of the important targets of chemometrics and chemoinformatics. However, environmental chemistry data has different characteristics from, *e.g.*, chemical reaction data in organic chemistry, which makes ML applications more challenging.

One of the recent trends in ML applications for chemistry is using quantum chemical (QC) methods for data acquisition.[4–12] QC electronic structure can provide more detailed and precise data going beyond simple parameter-based analysis.

In this article, we will discuss possible applications using AI and ML technology to help improve water quality, especially focusing on the degradation of chemical compounds in wastewater. We first describe the problems of micropollutants in wastewater and the differences in availability and quality of environmental chemistry data compared to lab organic chemistry data. We then highlight several ML applications for prediction of degradations in wastewater, including using *ab initio* QC together with experimental data, and discuss the challenges involved. Finally, we give a perspective on potential future applications of AI technology to improve water quality.

2. Micropollutants in Wastewater

Today, there are over 100'000 chemicals on the market in Europe,[13] likely more than 300'000 globally,[14] and numbers are increasing.[15] It is easily perceivable and has been widely documented through monitoring in different environmental compartments that many of these chemicals may be released to some extent into the environment during manufacturing, use, recycling or disposal (*e.g.*, refs [16–18]). Chemicals used in different industries (*e.g.*, solvents, reagents, adjuvants, lubricants, antioxidants, biocides, *etc.*) and in domestic applications (*e.g.*, in human medicine, washing agents, personal care products) may be released into industrial and domestic wastewaters, which are then treated in either specialized industrial, or, in most cases, domestic wastewater treatment plants (WWTPs).

During wastewater treatment, different processes might lead to removal of the chemicals from the water stream, including absorption unto sludge, degradation by the activated sludge microbial community, and volatilization. While sorption to sludge might play a role for rather hydrophobic compounds (*i.e.*, neutral chemicals with log $K_{ow} > 3$), any removal of more polar chemicals is mainly driven by microbial biodegradation. Yet, WWTPs are designed to remove general nutrients (*i.e.*, C, N and P), and hence are not optimized for the removal of the wide variety of chemicals potentially entering them through wastewater. As a consequence, many chemicals are not or only partially removed during biological wastewater treatment, and are thus released into surface water bodies like lakes and rivers.[19] There, they are typically found as highly complex mixtures of several hundreds to thousands of chemicals present at rather low concentrations (*i.e.*, in the low μg/L to ng/L range), which is why they are often termed 'micropollutants'.

To avoid negative impacts of these chemical releases on aquatic ecosystems in the densely populated areas of Switzerland, Switzerland has updated its Water Protection Act in 2014 to require technical measures on selected municipal WWTPs to reduce micropollutant loads to surface waters.[19] The most widely adopted measures include additional polishing steps of the biologically treated wastewater through either reaction with ozone (*i.e.*, ozonation) or adsorption to powdered or granular activated carbon. While adsorption to activated carbon removes the pollutants from the wastewater, degradation by microbial activity or ozonation does not always lead to complete mineralization – and hence removal – of the compounds, but may lead to the accumulation of potentially problematic transformation products. Hence, to be able to understand or even improve the degradation of micropollutants in wastewater treatment, both degradation by activated sludge microbial communities as well as by ozonation should ideally be predictable based on chemical structure.

3. Key Differences between Environmental Degradation Information and Reaction Data from Bench Chemistry

Degradation prediction is related to the wide field of reaction prediction, which is one of the most basic tasks in chemistry. Since an embryonic idea of data-driven reaction prediction started in 1980s,[20] various systems have been developed. After a first modest peak of development in 1990s,[21,22] recent advances in ML technologies have pushed automatic reaction prediction to a practically usable level.[23–25]

These systems have been targeting general organic chemical reactions carried out at the lab bench. In principle, similar methods can also be applied to predicting reactions of chemicals, *i.e.*, chemical degradation in wastewater. However, if one wants to address the full complexity of chemical reactions in wastewater, one needs to adjust the basic strategy. The major problem with degradation processes in wastewater comes from their much lower degree of controllability and traceability, leading to lower quality and quantity of chemical degradation information from wastewater.

3.1 *Controllability and Traceability*

Chemical reactions performed at the lab bench or industry scale are complex and it is still challenging to predict their outcome precisely. Chemical reactions are determined by a complicated interplay of many factors, such as electronic and geometrical properties of reactants, their changes during the chemical reactions, solvent, temperature, concentration, pressure, reaction time and presence of catalysts. The outcome of reactions varies depending on the degree of contribution of these factors. Therefore, predicting reactions corresponds to solving the relationships of data in multi-dimensional space. Even for simplified reactions in isolated conditions as often assumed in QC calculations, one has to deal with a huge 'chemical reaction space'.

Chemical reactions in a wastewater environment do not only have a higher complexity – given that they are mostly catalyzed by enzymes in a highly complex and diverse microbial community – but the available reaction information has a number of additional uncertainties and knowledge gaps.

Table 1 summarizes the differences between general (organic) chemical reactions run and analyzed at lab bench (RxnTyp1), those for model systems simulating the wastewater environment at lab bench (RxnTyp2) and those monitored at full scale in an actual wastewater treatment environment (RxnTyp3).

Table 1. Property differences for general (organic) chemical reactions at lab bench (RxnTyp1), chemical reactions in a model system at lab bench that simulates the environment (RxnTyp2) and chemical reactions in environmental wastewater (RxnTyp3).

	RxnTyp1 General Lab Bench	RxnTyp2 Env. Model Lab Bench	RxnTyp3 Env. Wastewater
Controllability and Monitoring			
Reactants	Yes	Yes	No
Quantity	Yes	Yes	No
Timing	Yes	Yes	No
Reproducible?	Yes, in principle	Difficult	Difficult
Time Scale	Hours ~ a few days	Hours ~ days	Hours ~ Years
Reaction Condition			
Solvent	Wide variety	Water	Water
Temperature	Wide range	0 ~ 40 °C[a]	0 ~ 40 °C[a]
pH	Wide range	Small range	Small range
Pressure	Wide range	Atmospheric pressure	Atmospheric pressure
Traceability			
Structure Identification	Yes	Possible, with uncertainty	Possible, with uncertainty
Product/Mass Balance	Yes	Possible, but challenging	Possible, but challenging
Reaction Mechanism	Sometimes	Possible	Possible, but challenging

[a]Assuming standard environment.

A major problem of RxnTyp3 is that reaction conditions cannot be controlled, but only be (partially) observed. Therefore, the range of conditions sampled in different monitoring campaigns is extremely wide and measurement results, *i.e.*, extent and products of degradation, accordingly vary based on when and where those samples were obtained. Furthermore, the time scale of RxnTyp3 may be long (*i.e.*, months to year) making it difficult to accurately assess differences between slowly degrading chemicals based on monitoring information. Finally, reaction kinetics cannot be directly observed, but only removal during wastewater treatment. To estimate reaction kinetics from such information, the relevant part of the wastewater treatment system must be fully parameterized and explicitly modeled, leading to additional uncertainty in the final kinetic estimate.

Therefore, it is very important to design appropriate lab model systems to simulate degradation in a wastewater environment (RxnTyp2). Such experiments are typically carried out with complex microbial communities directly sourced from the treatment steps of the WWTP, typically from the activated sludge basin (*i.e.*, aerated, suspended biomass). Compared to RxnTyp1, there are a limited variety of choices of types of reactants and reaction conditions that can be explored. Many reactants and reaction conditions for RxnTyp2 are limited to those available under realistic wastewater treatment conditions. Another limitation of RxnTyp2 is that full mass balances and hence product ratios are difficult to achieve since tracing and quantification of reaction products is challenging in the complex wastewater matrix, unless radioactively labelled compounds are used. The latter is very costly and such data is only (publicly) available for a very limited number of compounds. With appropriate efforts in product analysis, *e.g.*, using high-resolution mass spectrometry and MS/MS interpretation, enzymatic reaction types can be assessed in terms of general categories but since sequences of enzymatic reactions may have generated observed products, such interpretation is much more challenging than in the clean, single reaction systems typical for RxnTyp1.

3.2 *Quality and Quantity of Data*

The differences in RxnTyp3 affect data quality and quantity. As shown in Fig. 2, chemical reaction schemes for RxnTyp2-3 are often not complete. Together with the uncertain nature of the RxnTyp2-3 reactions as mentioned above, it is difficult to obtain a high data quality for environmental reactions.

Most of the databases of chemical reactions and compounds are from bench chemistry, RxnTyp1, and many databases are available.[12,26–30] The world's largest organic chemical reaction databases are CAS reactions[27] and Beilstein,[28] which contain approximately 150 million and 10 million reactions, respectively. The Beilstein database can be currently accessed from Reaxys.[29] Use of many commercially available databases of RxnTyp1 for data-driven research is restricted, but some datasets are accessible without such restrictions, for example, datasets mostly collected from patent data.[30] They are used for data-driven applications, such as Molecular Transformer.[24] QC-based reaction data, including detailed reaction mechanisms, are also available.[12]

In contrast, hardly any sizable, well-curated databases are available for RxnTyp2-3 for degradation in wastewater. With the exception of the Eawag-Sludge package in enviPath,[31] data collections related to degradation in wastewater are mostly spread throughout the literature in the form of lists of data that are mostly not electronically available nor curated according to any standard formats (*e.g.*, refs. [32,33]). The poor quality and quantity of data for RxnTyp2-3 makes data-driven approaches to train predictive models extremely difficult.

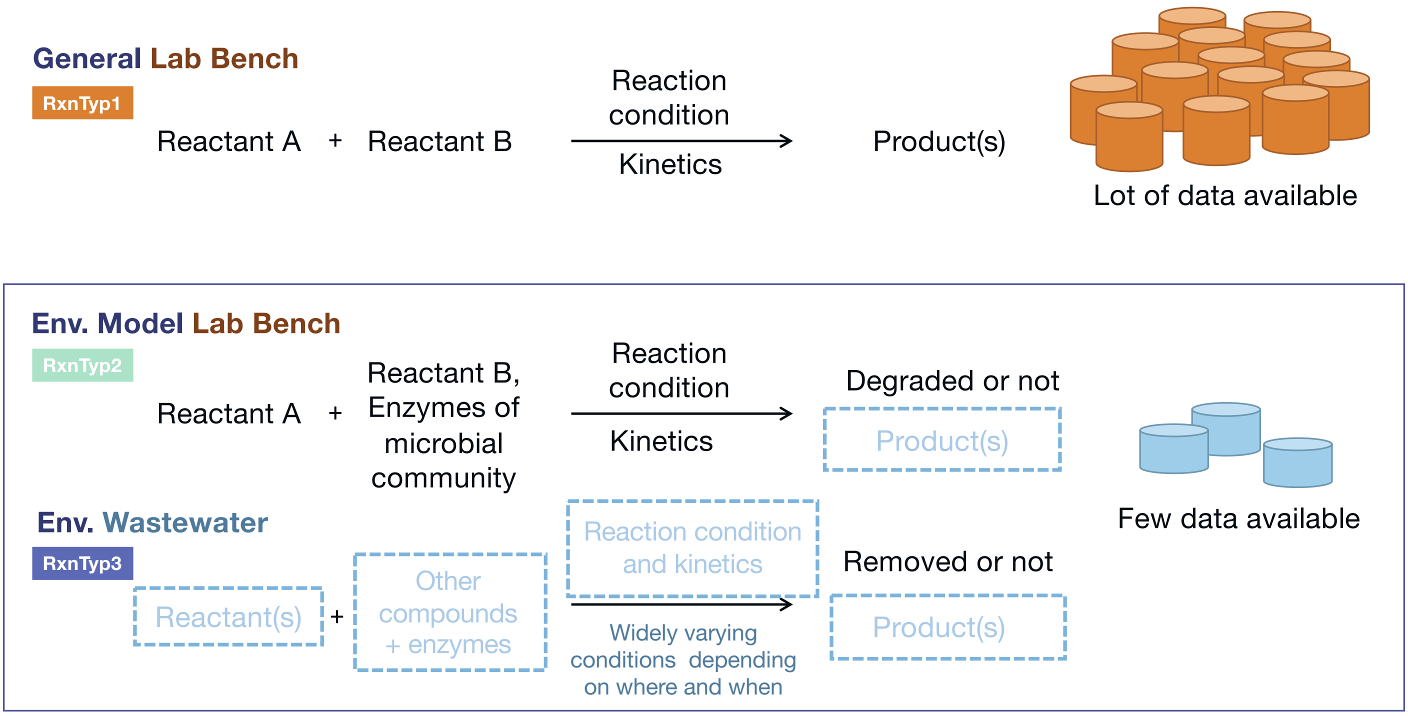


Fig. 2. Information in chemical reactions of environmental chemistry. It makes data-driven prediction more difficult.

4. Predicting Degradation in Wastewater – Challenges and Potential Application of *Ab initio* QC Methods

4.1 *Small Data Sets Challenge Application of ML Approaches*

Sizes of data sets for micropollutant degradation kinetics (*i.e.*, half-lives or rate constants) under somewhat standardized testing conditions for processes relevant for wastewater treatment are typically on the order of a few hundred data points at most.[34] At envipath.org, we currently host the most up-to-date, well curated and annotated sets of microbial degradation half-life data for activated sludge and soil. These currently contain 6259 half-lives for 895 chemicals in soil, and 563 rate constants or half-lives for 59 chemicals in activated sludge, respectively.

For prediction of degradation half-lives during wastewater treatment, so-called quantitative structure-(bio-)degradation relationships (QSBR models) are most typically developed and used. These mostly employ rather classical multivariate regression modeling and standard ML approaches (*e.g.*, support vector machines (SVM), random forest). In terms of descriptors, different chemical fingerprints and/or standard cheminformatics descriptors are often used.[35]

The most recent example of a QSBR developed for predicting biodegradation half-lives in WWTPs during activated sludge treatment was trained on data for 51 compounds ($R^2 = 0.69$) and validated with data for 18 compounds ($R^2 = 0.5$).[36] While the authors claim that their final root-mean-square error (RMSE) of 0.37 is reasonable, it is actually rather large relative to the dynamic range of the data (0.1–3.9 d^{-1}, *i.e.*, about 1.6 log units). Moreover, with only 18 validation chemicals, it remains hard to judge how well the model really generalizes. The probably most comprehensive recent attempt to develop QSBR models for half-life prediction are for degradation in water, sediment and soil. These models were trained on a data set of semi-quantitative half-life data for about 200 chemicals. Good results for these chemicals are reported both in training and cross-validation (*i.e.*, R^2 typically above 0.8 and a RMSE for logarithmic half-lives around 0.3 relative to a range of four orders of magnitude in half-lives).

However, when applied to an external validation set of 39 structurally diverse pharmaceuticals, which would be typical wastewater contaminants, the model clearly indicates a mismatch with the applicability domain and indeed does not provide meaningful prediction outcomes ($R^2 < 0$) (own evaluations). Due to limited coverage of the chemical space in the training set, which focuses on simple or halogenated hydrocarbons, these models can therefore not be applied to more complex substance classes such as pharmaceuticals.

While enviPath contains biodegradation data for more chemicals than the previously used data sets for developing models for half-life prediction, results from initial trials with model development are not very satisfactory. Preliminary model evaluations yield RMSE ~0.5, but corresponding R^2 values < 0.3 indicate that the model descriptors only capture a fraction of the variance in the half-life data (in this case for pesticide degradation in soil). Compared to the good superior model performance reported by Lombardo *et al.*[37] on a smaller data set containing more homogenous structures, these preliminary model evaluations suggest that the high structural variability in modern pesticides and pharmaceuticals is hard to capture with current QSBR approaches, despite the larger data sets in our hands.

To obtain predictive models with large applicability domains that include complex structures, possible solutions are currently being explored. For example, models can be trained on joint data sets containing half-lives observed in different environmental systems to increase the data volume and increase its structural diversity (combined learning). Furthermore, the molecular descriptors could be tailored to the problem by choosing descriptors that indicate the presence of biochemical reactive sites relevant to biodegradation. In the end, however, radically new ideas may be needed to substantially improve the predictive power of biodegradation models.

4.2 *QC Application for Prediction of Chemical Reactions*

Ab initio QC predictions of key chemical properties determining environmental degradation processes could overcome the deficiencies of descriptor-based, statistical QSBR models for structurally complex molecules. To do so, the key properties relevant in the rate-determining steps of an environmental degradation reaction chain have to be predicted. However, this rate-determining step can be expected to vary between enzymes and is mostly unknown for the large number of enzymes potentially involved in micropollutant degradation. Hence, direct application of QC-based properties for degradation prediction is very likely not possible for the bulk of biodegradation reactions in natural microbial communities where thousands of enzymes are working in concert.

In contrast, ozonation, used as a polishing step in advanced wastewater treatment, is a purely chemical reaction between ozone (and other reactive oxygen species) and wastewater contaminants and could thus be a valid target for *ab initio* QC approaches. Indeed, it has been shown that the second-order rate constant of the reaction of ozone with chemicals in aqueous solution could be well predicted by QC-calculated orbital energies. Specifically, ozonation of aromatic compounds was predicted well by the highest occupied molecular orbital (HOMO) (or HOMO-n if the HOMO was not placed on the aromatic ring), and the ozonation of olefines and amines correlated well with the natural bond orbital of the carbon–carbon π bond in the case of olefins or the nitrogen lone-pair electrons in the case of amines, respectively.[38] The respective linear correlations were later implemented in prediction tools for ozonation reactions, along with a rule-based system to also predict the resulting ozonation products.[39]

One limitation of the suggested approach was that the correlations between the QC descriptors and the second-order rate constants for reaction with ozone were still compound-group specific and not generally applicable. For aromatic compounds, for instance, different relationships were obtained for phenols versus other benzene derivatives such as anilides or benzotriazoles. For more complex molecules, it will be very challenging to assign them to the right group of compounds. For future developments, the goal will be to find appropriate complementary descriptors that allow developing more broadly applicable predictive relationships.

4.3 *QC Application for Specific Relevant Enzyme Systems*

Laccases, also called multicopper oxidases, are a group of enzymes that are known as potent monooxygenases, particularly in the case of fungal laccases.[40] In metatranscriptomic data from activated sludge microbial communities, we found gene transcripts of bacterial laccases to be rather abundant and their abundances to align with oxidative transformations of micropollutants, *e.g.*, hydroxylations, or oxidative N- and O-dealkylations. Bacterial laccases typically have a lower reduction potential than fungal ones, and therefore can be expected to require a mediator compound for efficient catalysis of oxidation reactions.

When the respective laccase sequences extracted from wastewater were cloned and expressed in *E. coli*, the respective purified enzymes indeed showed increased activity in combination with a typical laccase mediator compound, *i.e.*, 2,2-azino-bis-3-ethylbenzthiazoline-6-sulfonic acid (ABTS).[41] It has been suggested that, in such laccase-mediator systems, the laccase itself oxidizes ABTS to the doubly charged cation $ABTS^{2+}$, which then in turn oxidizes other chemicals with lower oxidation potential than the mediator itself.[40] Since previous data suggest that, most likely, the latter reaction of the activated mediator with the chemical is the rate-limiting step in this process, the system reduces again to a purely chemical reaction system, which, in principle, should be predictable from suitable QC descriptors.

Therefore, we have performed an experimental study on the enzymatic activity of a specific bacterial laccase (also called multicopper oxidases (MCO)) for which abundant gene transcripts were found in activated sludge from WWTPs. We then compared the experimental outcomes of this study with a number of potentially relevant QC descriptors to see whether they can predict reactivity in the experimental system. In the experimental study,[41] degradation reactions were examined for 20 structurally diverse micropollutants (**1**–**20** of Scheme 1) with MCO and ABTS as a mediator (**21**, shown in a solid box of Scheme 1) in ammonium acetate buffer at pH 4–6 (Fig. 3). The progress of degradation was checked at defined timepoints: 0, 2, 6, 10 and 30 h. As a result, 8 compounds (**13**–**20**, those in the dotted box of Scheme 1) were found to be significantly degraded, while 12 compounds (**1**–**12**) were not significantly removed during the experimental periods. See Ref. [41] for more experimental details.

We have searched for indicators that could distinguish between the reactive and non-reactive compound groups. We investigated the following nine *ab initio* QC parameters: one-electron oxidation potential,[42] electronegativity,[43] ionization energy,[43] electron affinity,[43] chemical hardness,[43] HOMO energy, lowest unoccupied molecular orbital (LUMO) energy, HOMO-LUMO energy gap and polarizability. All of the calculations were performed at the M062X/6-311+G(2df,2p)//M06L/6-311+G(2df,2p) level with the SMD solvent model by using the Gaussian 16 program package.[44]

We found that plotting the one-electron oxidation potential against the HOMO-LUMO energy gap clearly separated reacted and not-reacted compounds into two groups (Fig. 4). A model trained using those two parameters with the linear SVM method, which is a ML method for regression and clustering, showed a good prediction capability.

It is notable that such simple QC descriptors were able to identify differences in degradability across a set of structurally highly variable micropollutants. This result indicates applicability of QC descriptors to a case where properties based on electronic structures play an important role in the rate-limiting step of the degradation processes. QC descriptors provide more precise and detailed view of electronic structure-based properties, which are different from simple structure or parameter-based descriptors. Those QC descriptors can be similar even for structures with different functional groups (*e.g.*, **14** and **19**), or they can be different for similar structures (*e.g.*, **1** and **4**).

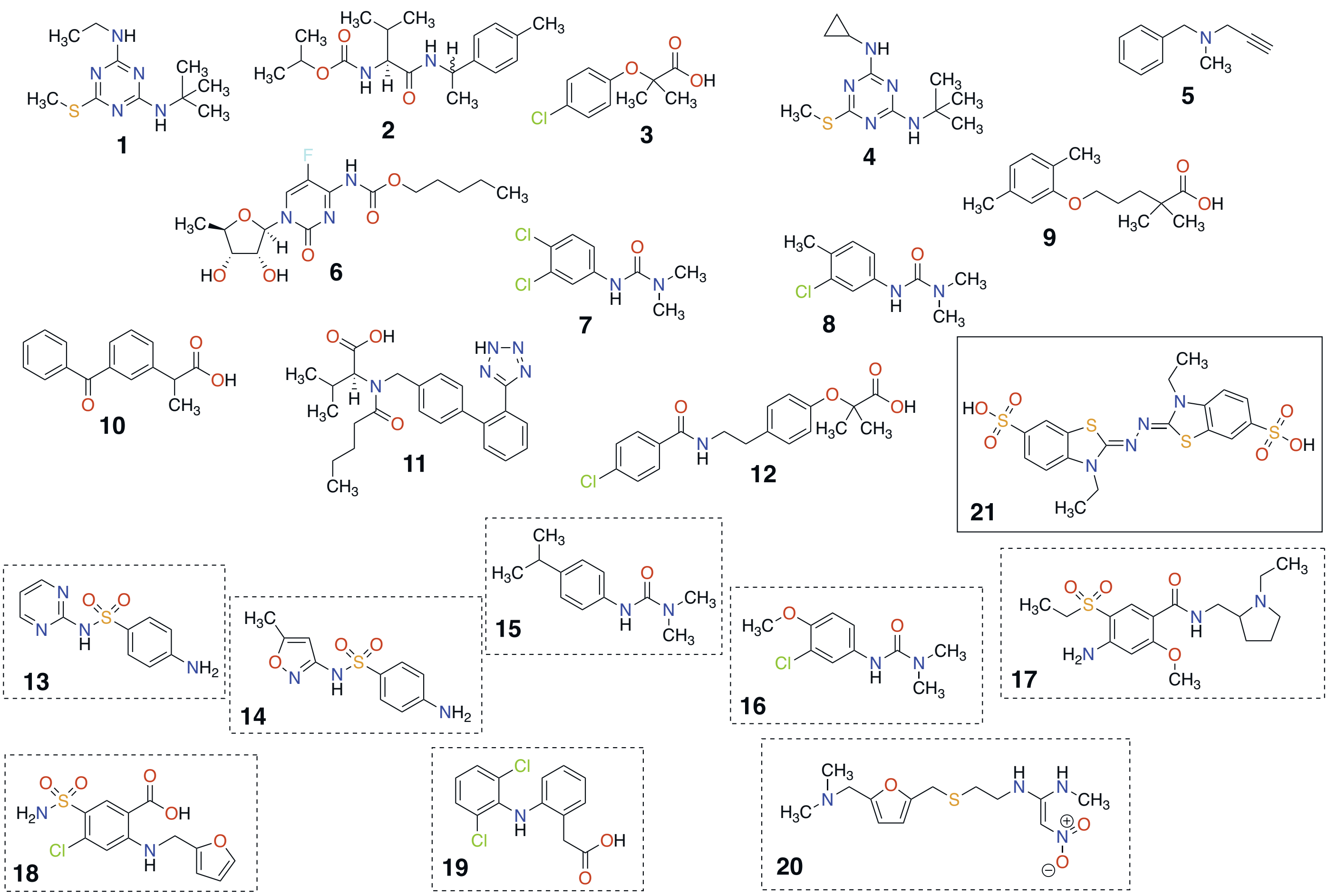
Based on these promising first results, we are further investigating a larger set of micropollutants with QC and ML methods as well as laccase-based degradation experiments.

5. Outlook

5.1 *QC for Data Acquisition and for Precise Descriptors*

As described above, QC descriptors have a high potential to overcome the deficiencies of descriptor-based statistical QSBR models for structurally complex molecules. QC descriptors are usually obtained by time-consuming calculations. However, the situation has greatly improved thanks to advances in computer technology and method development. For example, ML-based force fields, initiated by Behler and Parrinello,[45,46] paved the way for enormously accelerating computations of structures while keeping the same level of accuracy. Use of AI methods has been more common also to generate descriptors with higher accuracy from lower level calculations. Furthermore, advanced methods for exploration of potential energy surfaces allow to automatically calculate not only equilibrium structures but also transition states as well as reaction pathways.[47] These impressive advancements encourage using QC methods for data acquisition[4–12] as mentioned in the Introduction. The same methods can be used to enrich the quality and quantity of descriptors for predicting degradation in wastewater.

Despite enormous efforts, calculations of large molecular systems, like enzymes, are still challenging. This is especially true when chemical reaction events are involved. Combining these new advanced methods will allow to calculate QC descriptors for



Scheme 1. Compounds examined experimentally.[41] Compounds **13**–**20** were degraded, while **1**–**12** were not. Those in the dotted box were degraded. **21** is a mediator.

much larger systems than is possible today. That in turn will help build more accurate models, look into mechanistic details of degradation and maybe even design artificial enzymes having the capability to degrade micropollutants that are persistent in the environment and difficult to degrade with natural enzymes.

5.2 AI Technology for Efficient Experiments

There are several possibilities how AI technology can support more efficient experiments, *i.e.*, to improve quality and quantity of experimental data for both of RxnTyp2 and RxnTyp3 (Fig. 5).

One possibility is measurement informatics (MI), where AI technology is applied to various measurement techniques, *e.g.*, spectroscopy.[48–50] The aims of MI include extracting more in-

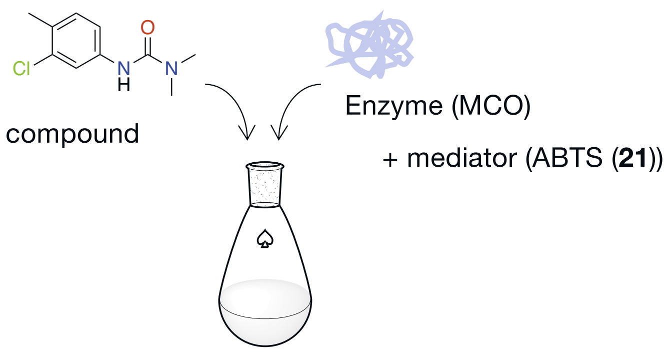


Fig. 3. Experimental study on degradation reactions of 20 compounds with multi-copper oxidase (MCO) and a mediator (ABTS (**21**)). As a result, 8 compounds were degraded, while 12 compounds were not.

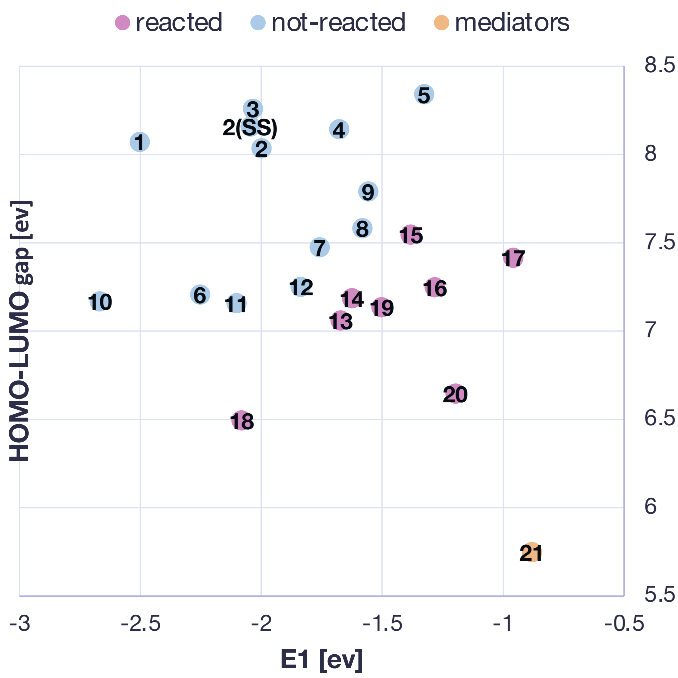


Fig. 4. Plot of one-electron oxidation potential (E1) against HOMO-LUMO gap. Reactive and non-reactive compounds are clearly clustered. A linear SVM model trained with these descriptors shows a good prediction capability.

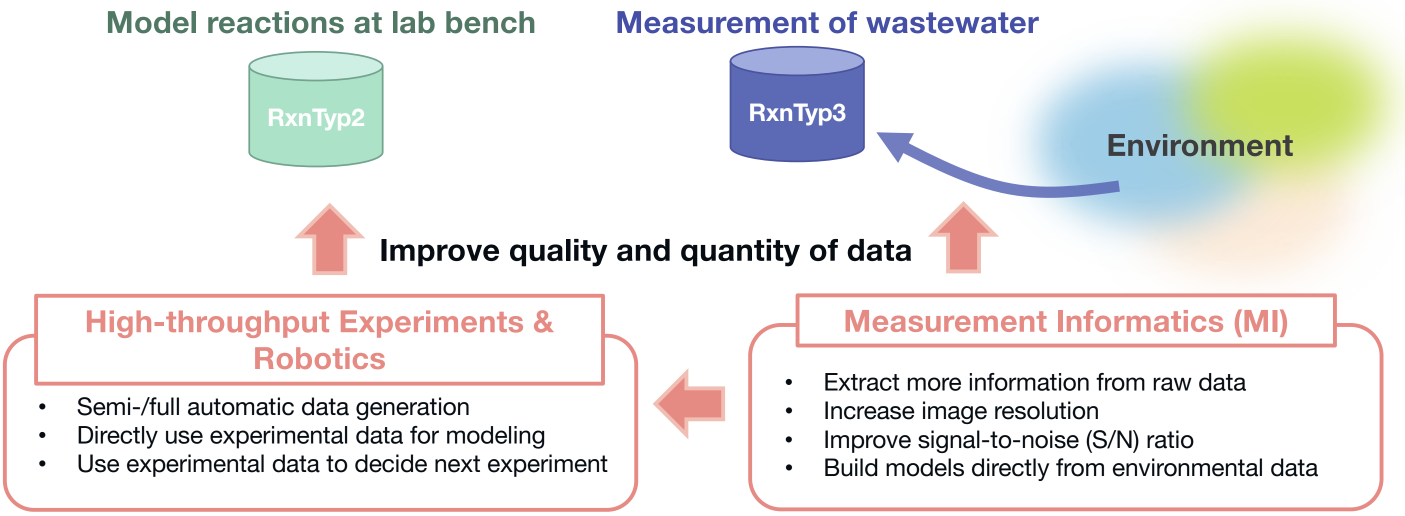


Fig. 5. How AI technology can help improve quality and quantity of experimental data for wastewater.

formation from raw data, increasing image resolution, improve signal-to-noise (S/N) ratio, and building models directly from raw data. Several types of AI methods can be used, including convolutional neural network, data assimilation, and Bayesian statistics.

For RxnTyp2 and RxnTyp3 data, MI methods could help extract more and higher quality information from high-resolution mass spectrometry data obtained from direct monitoring at WWTPs or from lab bench experiments with environmental samples, such as activated sludge microbial communities. These samples are typically highly complex, *i.e.,* they contain a very diverse mixture of hundreds if not thousands of natural and anthropogenic small molecules as well as larger biomolecules. Extracting information on micropollutants against this matrix background is challenging and could be supported by MI.

AI and MI methods are also useful for model experiments at lab bench (RxnTyp2) by supporting the application of high-throughput experimentation and robotics. This combination would make it possible to use experimental data, partly or fully automatically generated from high-throughput experiments, to build ML models. Similar approaches have been applied for catalytic design[51] and materials science.[52–54] It was demonstrated that such methods efficiently produce more capable ML models to predict properties and to design desired molecules or materials.

6. Conclusions

We provide a short review on the challenges associated with applying AI for the prediction of micropollutant degradation in wastewater, mostly associated with limitations in available environmental data. However, as outlined in this article, there are ways to cope with these problems by combining AI and MI technology. Also, QC methods might provide direct access to relevant properties for reactions where the rate-determining step is mostly driven by purely chemical interactions. Hence, combinations of these methods might help overcome the limited amount of data available for training statistical modeling approaches. We hope this account can inspire others to contribute to the development of methods and technologies that support improving water quality and designing degradable new molecules.

Received: November 15, 2022

[1] S. Sasaki, H. Abe, T. Ouki, M. Sakamoto, S. Ochiai, *Anal. Chem.* **1968**, *40*, 2220, https://doi.org/10.1021/ac50158a061.
[2] A. M. Duffield, A. V. Rovertson, C. Djerassi, B. G. Buchanan, G. L. Shtherland, E. A. Feigenbaum, J. Lederberg, *J. Am. Chem. Soc.* **1969**, *91*, 2977, https://doi.org/10.1021/ja01039a026.
[3] S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito, T. Yamasaki, *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 4, 211, https://doi.org/10.1021/ci60016a007.
[4] H. Satoh, S. Itono, K. Funatsu, K. Takano, T. Nakata, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 671, https://doi.org/10.1021/ci9801567.
[5] H. Satoh, K. Funatsu, K. Takano, T. Nakata, *Bull. Chem. Soc. Jpn.* **2000**, *73*, 1955, https://doi.org/10.1246/bcsj.73.1955.
[6] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003, https://doi.org/10.1088/1367-2630/15/9/095003.
[7] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Sci. Data* **2014**, *1*, 140022, https://doi.org/10.1038/sdata.2014.22.
[8] R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. von Lilienfeld, *J. Chem. Phys.* **2015**, *143*, 084111, https://doi.org/10.1063/1.4928757.
[9] M. Nakata, T. Shimazaki, *J. Chem. Inf. Model.* **2017**, *57*, 1300, https://doi.org/10.1021/acs.jcim.7b00083.
[10] H. Satoh, T. Oda, K. Nakakoji, T. Uno, S. Iwata, K. Ohno, *J. Comput. Chem., Jpn.* **2015**, *14*, 77, https://doi.org/10.2477/jccj.2015-0048.
[11] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nat. Commun.* **2017**, *8*, 872, https://doi.org/10.1038/s41467-017-00839-3.
[12] H. Satoh, T. Oda, K. Nakakoji, T. Uno, S. Iwata, K. Ohno, 'RMapDB: chemical reaction route map data for quantum mechanical-based data chemistry', *Materials Cloud Archive*, **2020**, *2020.138*, https://archive.materialscloud.org/record/2020.138.
[13] H. P. H. Arp, S. E. Hale, *ACS Environ. Au* **2022**, https://doi.org/10.1021/acsenvironau.2c00024.
[14] Z. Wang, G. W. Walker, D. C. G. Muir, K. Nagatani-Yoshida, *Environ. Sci. Technol.* **2020**, *54*, 5, 2575, https://doi.org/10.1021/acs.est.9b06379.
[15] E. S. Bernhardt, E. Rosi, M. O. Gessner, *Front. Ecol. Environ.* **2017**, *15*, 2, 84, https://doi.org/10.1002/fee.1450.
[16] J. L. Wilkinson, A. B. A. Boxall, D. W. Kolpin, C. Teta, *Proc. Natl. Acad. Sci USA* **2022**, *119*, 8, e2113947119, https://doi.org/10.1073/pnas.2113947119.
[17] S. Anliker, S. Santiago, K. Fenner, H. Singer, *Water Res.* **2022**, *215*, 118221, https://doi.org/10.1016/j.watres.2022.118221.
[18] E. L. Schymanski, H. P. Singer, P. Longrée, M. Loos, M. Ruff, M. A. Stravs, C. R. Vidal, J. Hollender, *Environ. Sci. Technol.* **2014**, *48*, 3, 1811, https://doi.org/10.1021/es4044374.
[19] R. I. L. Eggen, J. Hollender, A. Joss, M. Schärer, C. Stamm, *Environ. Sci. Technol*. **2014**, *48*, 14, 7683, https://doi.org/10.1021/es500907n.
[20] K. Funatsu, S. Sasaki, *Tetrahedron Comput. Method.* **1988**, *1*, 27, https://doi.org/10.1016/0898-5529(88)90006-1.
[21] P. Röse, J. Gasteiger, *Anal. Chem. Acta* **1990**, *235*, 163, https://doi.org/10.1016/S0003-2670(00)82071-1.
[22] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 34, https://doi.org/10.1021/ci00023a005.
[23] M. H. S. Segler, M. P. Waller, *Chemistry* **2017**, *23*, 5966, https://doi.org/10.1002/chem.201605499.
[24] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572, https://doi.org/10.1021/acscentsci.9b00576.
[25] C. W. Coley, W. Jin, L. Roggers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370, https://doi.org/10.1039/C8SC04228D.

[26] E. Zass, in 'Handbook of Chemoinformatics: From Data to Knowledge', Ed. J. Gasteiger, Wiley-VCH, **2003**, p. 667, https://onlinelibrary.wiley.com/doi/book/10.1002/9783527618279.
[27] https://www.cas.org/cas-data/cas-reactions
[28] A. J. Lawson, in 'Handbook of Chemoinformatics: From Data to Knowledge', Ed. J. Gasteiger, Wiley-VCH, **2003**, p. 608, https://onlinelibrary.wiley.com/doi/book/10.1002/9783527618279.
[29] https://www.reaxys.com/#/search/quick.
[30] D.M. Lowe, 'Extraction of Chemical Structures and Reactions from the Literature', PhD Thesis, University of Cambridge, **2012**.
[31] https://envipath.org/package/4a3cd0f4-4d2b-4f00-b3e6-a29e721f7038.
[32] R. Gulde, U. Meier, E. L. Schymanski, H-P. E. Kohler, D. E. Helbling, S. Derrer, D. Rentsch, K. Fenner, *Environ. Sci. Technol.* **2016**, *50*, 6, 2908, https://doi.org/10.1021/acs.est.5b05186.
[33] K. Voigt, in 'Handbook of Chemoinformatics: From Data to Knowledge', Ed. J. Gasteiger, Wiley-VCH, **2003**, p. 722, https://onlinelibrary.wiley.com/doi/book/10.1002/9783527618279.
[34] D. Awfa, M. Ateia, D. Mendoza, C. Yoshimura, *ACS EST Water* **2021**, *1*, 3, 498, https://dx.doi.org/10.1021/acsestwater.0c00206.
[35] C. Rücker, K. Kümmerer, *Green Chem*. **2012**, *14*, 875, https://doi.org/10.1039/c2gc16267a.
[36] T. M. Nolte, G. Chen, C. S. Van Schayk, K. Pinto-Gil, A. J. Hendriks, W. J. G. M. Peijnenburg, A. M. J. Ragas, *Sci. Total Environ*. **2020**, *708*, 133863, https://doi.org/10.1016/j.scitotenv.2019.133863.
[37] A. Lombardo, A. Manganaro, J. Arning, E. Benfenati, *Sci. Total Environ.* **2022**, *838*, 156004, https://doi.org/10.1016/j.scitotenv.2022.156004
[38] M. Lee, S. G. Zimmermann-Steffens, J. S. Arey, K. Fenner, U. Von Gunten, *Environ. Sci. Technol*. **2015**, *49*, 16, 9925, https://doi.org/10.1021/acs.est.5b00902.
[39] M. Lee, L. C. Blum, E. Schmid, K. Fenner, U. Von Gunten, *Environ. Sci.: Processes* **2017**, *19*, 465, https://doi.org/10.1039/c6em00584e.
[40] J. Margot, P.-J. Copin, U. Von Gunten, D. A. Barry, C. Hollinger, *Biochem. Eng. J.* **2015**, *103*, 47, https://doi.org/10.1016/j.bej.2015.06.008.
[41] A. Athanasakoglou, K. Fenner, *Environ. Sci. Technol.* **2022**, *56*, 313, https://doi.org/10.1021/acs.est.1c05803.
[42] W. A. Arnold, *Environ. Sci. Process Impacts* **2014**, *16*, 832, https://doi.org/10.1039/C3EM00479A.
[43] S. A. Grimmel, M. Reiher, *CHIMIA* **2021**, *75*, 311, https://doi.org/10.2533/chimia.2021.311.
[44] Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian, Inc., Wallingford CT, **2016**.
[45] J. Behler, M. Parrinello, *J. Chem. Phys.* **2007**, *127*, 1, 07B603, https://doi.org/10.1063/1.2746232.
[46] J. Behler, *Chem. Rev*. **2021**, *121*, 10037, https://doi.org/10.1021/acs.chemrev.0c00868.
[47] K. Ohno, H. Satoh, 'Exploration on Quantum Chemical Potential Energy Surfaces. Towards the Discovery of New Chemistry', Royal Society of Chemistry, **2022**, https://pubs.rsc.org/en/content/ebook/978-1-83916-490-3.
[48] T. Ueno, H. Iwasawa, *Synchrotron Radiat. News* **2022**, https://doi.org/10.1080/08940886.2022.2112497.
[49] P. Taechawattananant, K. Yoshii, Y. Ishihama, *J. Proteome Res.* **2021**, *20*, 5, 2291, https://doi.org/10.1021/acs.jproteome.0c00819.
[50] C.-H. Chang, D. Yeung, V. Spicer, K. Ogata, O. Krokhin, Y. Ishihama, *J. Proteome Res.* **2021**, *20*, 7, 3600, https://doi.org/10.1021/acs.jproteome.1c00185.
[51] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura, T. Taniike, *ChemCatChem*, **2019**, *11*, 1, https://doi.org/10.1002/cctc.201801956.
[52] H. Y. Zang, A. R. De La Oliva, H. N. Miras, D-L. Long, R. T. McBurney, L. Cronin, *Nat. Commun.* 2014, 5, 3715, https://doi.org/10.1038/ncomms4715.
[53] Y. Iwasaki, M. Ishida, M. Shirane, *Sci. Technol. Adv Mater.* **2020**, *21:1*, 25, https://doi.org/10.1080/14686996.2019.1707111.
[54] Y. Jiang, D. Salley, A. Sharma, G. Keenan, M. Mullin, L. Cronin, *Sci. Adv*. **2022**, *8*, eabo2626, https://doi.org/10.1126/sciadv.abo2626.