Tobias Gensch

Digitale Katalyse

Digitale Katalyse ist die Idee, computergenerierte Hypothesen zum Design, Optimierung und Verständnis katalytischer Reaktionen einzusetzen, indem Experimente datengetrieben sowohl ausgewertet als auch vorgeschlagen werden. Die Ziele sind dabei einerseits, schneller und effizienter zu optimalen Katalysatoren oder Reaktionsbedingungen zu kommen, und andererseits, ganz neue Katalysatorstrukturen oder chemische Reaktionen zu finden. Gleichermaßen ermöglicht die Interpretation datengestützter Modelle komplementäre Einblicke in Reaktionsmechanismen und die fundamentalen Beziehungen zwischen Struktur und Reaktivität. Die grundlegende Annahme ist, dass statistische Modelle die sehr hochdimensionalen Faktoren, die den Ausgang chemischer Reaktionen bestimmen, mit geeigneten Daten lernen können. Diese Forschung findet an der Schnittstelle von experimenteller Chemie, Computerchemie und Data Science statt, parallel zur Weiterentwicklung datengetriebener Methoden in anderen Bereichen der Chemie, wie zum Beispiel der medizinischen Chemie oder den Materialwissenschaften. Während es in den letztgenannten vor allem um das Design von Molekülen oder Materialien mit bestimmten biologischen oder physikalischen Eigenschaften geht, steht in der Katalyse das gezielte Design chemischer Reaktivität im Vordergrund. Solche datengetriebenen Methoden werden sowohl für die heterogene, als auch die homogene Katalyse entwickelt, doch sollen im Folgenden vor allem Beispiele und Anwendungen der molekularen Katalyse beschrieben werden.

Die Optimierung katalytischer Reaktionen ist eine komplexe Aufgabe durch die schiere Anzahl an Reaktionsparametern, Zusatzstoffen und Katalysatorstrukturen. Dementsprechend erfordert dies traditionell viel Zeit und Aufwand, wobei Fortschritt im Wesentlichen auf empirische trial-and-error Prozesse sowie Literaturkenntnis und chemische Intuition angewiesen ist. Zur Suche einiger Reaktionsparameter wie Temperatur, Reaktionsdauer oder Konzentration sind teilweise auch numerische Methoden etabliert, wie zum Beispiel das Design-of-Experiments (DoE).[1] Hinter diesem Begriff stehen Verfahren, die dem Anwender ganz konkrete Experimente vorschlagen, um zunächst den Einfluss der genannten Parameter auf die Reaktion zu ergründen und durch numerische Interpolation schließlich zu optimieren. Das ist möglich, da diese Reaktionsparameter numerische Werte haben, die einfach und kontinuierlich verändert werden können: zum Beispiel kann der Einfluss der Temperatur auf eine Reaktion sehr einfach auch ohne komplizierte Hilfsmittel ergründet werden, indem man in einer Testreihe den relevanten Temperaturbereich in Schritten von 10 °C abtastet, und es ist auch ohne weiteres möglich, zwischen diesen Intervallen jede beliebige Temperatur einzustellen, wenn es sich als günstig herausstellt. Das offensichtliche Problem ist, dass die Optimierung chemischer Strukturen, wie zum Beispiel Katalysatoren oder Liganden, so nicht ohne weiteres stattfinden kann, da Moleküle weder Zahlen sind, noch kontinuierlich verändert werden können. Sobald man also in der Lage ist, Änderungen an Katalysatorstrukturen quantitativ zu erfassen, dann ist eine gezielte und effizientere Strukturoptimierung möglich, die von allen Daten Gebrauch machen kann, die zu einer Reaktion vorliegen.

Demnach gibt es in der Digitalen Katalyseforschung einen allgemeinen Workflow mit verschiedenen zentralen Aufgaben: Zunächst müssen Moleküle in numerische oder andere maschinenlesbare Darstellungen übersetzt werden. Daraufhin müssen geeignete Algorithmen gefunden und trainiert werden, um mit chemischen Daten Modelle zu erhalten. Mithilfe dieser können dann schließlich neue Experimente vorgeschlagen werden. Idealerweise sollte abschließend auch durch die Interpretation der Modelle weitere Validierung sowie Wissensgewinn stattfinden.

Ursprünge: Lineare freie Enthalpie-Beziehungen, molekulare Deskriptoren

Den Zusammenhang von Molekülstruktur und chemischer Reaktivität zu quantifizieren ist eines der Kernthemen der physikalischen organischen Chemie. In diesem Sinne gehen viele Grundlagen des hier Beschriebenen bereits auf lineare freie Enthalpie-Zusammenhänge (linear free-energy relationship, LFER) wie die Hammettgleichung zurück.[2] Solche LFER beschreiben, dass manche elektronische oder sterische Einflüsse auf die relativen freien Reaktionsenthalpien ähnlicher Moleküle, die nach dem gleichen Mechanismus reagieren, mit linearen Zusammenhängen beschrieben werden können. Dabei wird der Einfluss auf die relative Reaktivität durch empirische Parameter ausgedrückt, zum Beispiel die Hammett σ-Parameter für den elektronischen Effekt substituierter Aromaten. Ist eine solche freie Energie-Beziehung bekannt, kann also bereits die Reaktivität anderer Moleküle durch ihre entsprechenden Parameter vorhergesagt werden. Generell werden LFER aber vor allem zum Verständnis von Reaktionsmechanismen eingesetzt und nicht zur Optimierung chemischer Reaktionen. Neben diesen empirischen Reaktivitätsparametern ist es auch schon lange etabliert, Moleküle durch (experimentell bestimmte) Eigenschaften zu beschreiben und sie so numerischen Analysen zugänglich zu machen. Im Bereich der Übergangsmetallkatalyse sind die von Tolman beschriebenen Kegelwinkel (Tolman

Dr. Tobias Gensch
Institut für Chemie, Technische Universität Berlin
Straße des 17. Juni 115, D-10623 Berlin
Tobias.Gensch@tu-berlin.de

DOI: 10.26125/pz07-7e55

cone angle) und elektronischen Parameter (Tolman electronic parameter) exemplarisch, die auch über die ursprünglich betrachteten Phosphinliganden hinaus weiterhin absolute Standarddeskriptoren für Ligandeneffekte in Metallkomplexen sind.[3] In letzter Zeit werden statt gemessener Größen überwiegend Moleküleigenschaften, die mittels quantenchemischer Simulation zugänglich sind, zur Quantifizierung chemischer Reaktivität eingesetzt.[4] So können auch Vorhersagen über bisher unsynthetisierte Moleküle getätigt werden. Modelle, die chemische Reaktivität mit konkreten Moleküleigenschaften verknüpfen, haben oft den Vorteil einer hohen Interpretierbarkeit, da Zusammenhänge zum Reaktionsmechanismus über die im Modell vertretenen Eigenschaften gelernt werden können. Dabei finden neben komplexeren Modelltypen weiterhin auch (multivariate) lineare Modelle Anwendung.[5]

Aktuelle Entwicklungen

Eine Konsequenz der diskreten, nicht-numerischen Natur von Molekülen ist es, dass die Beziehungen verschiedener Moleküle zueinander für Menschen bestenfalls intuitiv zugänglich sind. Das betrifft Einschätzungen zur Ähnlichkeit bzw. Trends innerhalb einer Gruppe von Molekülen oder auch die Auswahl von Versuchsreihen, die eine relevante Spanne physikalisch/chemischer Eigenschaften möglichst vollständig und gleichmäßig umfassen sollen, wie es zum Beispiel beim Katalysatorscreening der Fall ist. Eine Anwendung der digitalen Katalysatorforschung, die in dieser Richtung helfen kann, ist die näherungsweise Abbildung des „chemischen Raumes“, der durch eine Molekülgruppe beschrieben wird [6] wie zum Beispiel „Phosphor-basierte Liganden“. Dabei werden zunächst repräsentative Vertreter dieser Gruppe in eine numerische Darstellung übersetzt (zum Beispiel als ihre physikalisch-chemischen Eigenschaften). Im Anschluss können mit geeigneten Verfahren wie etwa der Hauptkomponentenanalyse (principal component analysis, PCA) niederdimensionale Abbildungen, quasi „Landkarten des chemischen Raumes“, erstellt werden, die ein sehr einfaches, intuitives Hilfsmittel für die angesprochenen Anwendungen sein können. So können etwa zur Versuchsplanung möglichst unterschiedliche Punkte auf dieser Karte ausgewählt werden, um ausgewogene Testreihen zum Ligandenscreening zu erhalten. Alternativ können nach einem „Hit“ eines erfolgreichen Liganden auch ähnliche Vertreter aus der räumlichen Nähe in dieser Darstellung vorgeschlagen werden. Mit geeigneten Machine Learning Methoden kann die Navigation eines solchen chemischen Raumes auch in höheren Dimensionen stattfinden. Dies ermöglicht unter anderem die Behandlung von Liganden als quasi-kontinuierliche Variable in numerischen Optimierungen. In einer aktuellen Anwendung wurde eine stereoselektive, palladiumkatalysierte Kreuzkupplung von einem Syntheseroboter selbstständig optimiert, wobei fünf Reaktionsparameter inklusive des Liganden simultan von einem auf Bayesian optimization basierten Algorithmus variiert wurden.[7]

Das Design neuer Molekülstrukturen, wie zum Beispiel von Liganden für eine neue katalytische Reaktion, ist bisher immer auf das Potential der Ideen begrenzt, auf die daran Beteiligte eben kommen können. So kann es sein, dass optimale Katalysatoren nicht gefunden wurden, nur weil an einem bestimmten Strukturmotiv festgehalten wurde oder eher seltene Strukturen nicht in Erwägung gezogen wurden. Eine besondere Vision der digitalen Katalyseforschung ist es, durch entsprechende Modelle auch Vorschläge für neue Katalysatoren zu generieren, die eben durch reine Intuition nicht gefunden worden wären.[8] Zuvor wurde die Problematik angesprochen, die inhärent diskreten Moleküle in kontinuierliche Zahlenräume zu übersetzen. Doch nach solch einer Übersetzung steht man vor dem umgekehrten Problem: wie findet man konkrete Molekülstrukturen, die laut Modell eine bessere katalytische Reaktion ermöglichen? Da die allermeisten Molekülrepräsentationen lediglich numerische Beschreibungen von Molekülen sind, die noch nicht einmal eindeutig sind, gibt es keinen direkten Weg die Zahlen wieder in ein Molekül zu übersetzen. Um also wirklich neue Katalysatorstrukturen zu erhalten, werden generative Modelle benötigt, also solche, die gelernt haben aus einer kontinuierlichen Molekülrepräsentation wieder ganz konkrete Molekülstrukturen mit den gewünschten physikalischen und chemischen Eigenschaften zu erzeugen. Solche Modelle werden aktuell mit verschiedenen Ansätzen entwickelt, die teilweise auf künstlichen neuronalen Netzwerken oder genetischen Algorithmen basieren. Momentan sind diese Methoden hauptsächlich im Kontext des Medikamenten- und Materialdesigns in Entwicklung, aber die Anwendung zum Ligandendesign ist absehbar und vielversprechend.[9]

Zukunft: Umfassende prädiktive Modelle ganzer Reaktionsklassen

In Zukunft wird es möglich sein, statt einzelner Reaktionen ganze Reaktionsklassen auf einmal zu betrachten, und so potentiellen Anwendern einer Reaktion für die speziell benötigen Substrate gezielt optimale Katalysatoren und Reaktionsbedingungen vorschlagen zu können. Zum Beispiel ist es realistisch, dass dies zunächst für Transformationen wie die palladiumkatalysierte Kreuzkupplung geschieht, zu denen sehr viele Veröffentlichungen mit diversen Substraten und Katalysatoren vorliegen. Die langfristige Vision ist es, dass entsprechend starke Modelle irgendwann aus dem gesamten Schatz bisheriger Veröffentlichungen so viel über Chemie gelernt haben, dass sie auch optimale Katalysatorstrukturen für seltene Reaktionen vorschlagen können oder sogar für die Vorhersage komplett unbekannter Reaktionen genutzt werden können. So könnten idealerweise aufwändige Verfahren, wie die Prozessoptimierung oder die Entwicklung von Leadmolekülen, deutlich beschleunigt und Ressourcen anderweitig verwendet werden. In der Realität kommt bei der Vorhersage solcher Modelle jedoch nicht „der eine“ optimale Vorschlag heraus, sondern im Gegenteil potentiell eine unüberschaubar große Menge von Vorschlägen. Das heißt, dass chemische Expertise weiterhin benötigt wird, um aus dieser Vielzahl sinnvolle Vorschläge zu filtern und diese in konkrete chemische Experimente zu übersetzen. Die Hoffnung hier ist also, einen neuen Weg zur Generierung von Hypothesen zu eröffnen, der aus einer größeren Vielzahl von Möglichkeiten schöpfen kann, als es einem Menschen möglich ist, und so vorher übersehene „weiße Flecken“ auf der Karte des chemischen Raumes auch zu berücksichtigen.

Dies wird nur möglich sein, wenn alle in der chemischen Literatur vorhandenen Informationen vereinheitlicht werden, damit

möglichst umfassende Datensätze entstehen, in denen auch subtile Substrat-Katalysator-Interaktionen oder seltene Effekte erfasst werden können. Das kann umso besser funktionieren, je vollständiger und einheitlicher die Datensätze publiziert werden, die in chemischen Studien anfallen. Dies gilt insbesondere für Experimente, die subjektiv als „schlechte" Ergebnisse wahrgenommen werden, also lediglich geringe Ausbeuten oder Selektivitäten erzielten oder sogar gar kein gewünschtes Produkt erzeugten. Solche Experimente sind für quantitative Analysen absolut unabdingbar, da es physikalisch-chemisch natürlich Gründe gibt, warum eine gewisse Kombination von Reaktionsparametern oder eine Katalysatorstruktur nur in einer geringen Ausbeute resultierte. Dies erfordert das Bewusstsein der Praktizierenden, dass „Datenwissenschaft" mit dem Erzeugen der Daten beginnt, und solche „schlechten" Ergebnisse gleichwertiger Bestandteil des gesamten Datensatzes sind. Es muss natürlich auch erwähnt werden, dass die Reproduzierbarkeit und Genauigkeit chemischer Resultate inhärent auch die Genauigkeit der damit entwickelten Modelle bestimmen. Das ist vor allem in der Chemie ein Problem, da den Praktizierenden sehr oft gar nicht alle Einflussfaktoren auf das Ergebnis einer Reaktion bekannt sind, wie zum Beispiel der Einfluss des genutzten Glasgerätes oder Rührers oder Spuren von Verunreinigungen. Auch hier kann eine erhöhte Wahrnehmung der quantitativen Rolle chemischer Daten dazu beitragen, zumindest durch möglichst vollständige Dokumentation für möglichst gute Reproduzierbarkeit zu sorgen.

Referenzen

[1] Paul M. Murray, Simon N. G. Tyler, Jonathan D. Moseley: Beyond the Numbers: Charting Chemical Reaction Space, *Accounts of Chemical Research* 2016 **49**, 1292 – 1301.

[2] Corwin Hansch, Albert Leo, Robert W. Taft: A Survey of Hammett Substituent Constants and Resonance and Field Parameters, *Chemical Reviews* 1991 **2**, 165 – 195.

[3] Chadwick A. Tolman: Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis, *Chemical Reviews* 1977 **3**, 313 – 348.

[4] Derek J. Durand, Natalie Fey: Computational Ligand Descriptors for Catalyst Design, *Chemical Reviews* 2019 **11**, 6561 – 6594.

[5] Matthew S. Sigman, Kaid C. Harper, Elisabeth N. Bess, Anat Milo: The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond, *Accounts of Chemical Research* 2016 **49**, 1292 – 1301.

[6] Natalie Fey: Lost in chemical space? Maps to support organometallic catalysis, *Chemistry Central Journal* 2015 **9**, 38.

[7] Melodie Christensen, Lars Yunker, Folarin Adedeji, Florian Häse, Loic Roch, Tobias Gensch, Gabriel dos Passos Gomes, Tara Zepel, Matthew Sigman, Alan Aspuru-Guzik, Jason Hein: Data-science driven autonomous process optimization, *ChemRxiv* 2020, 10.26434/chemrxiv.13146404.v2.

[8] Benjamin Sanchez-Lengeling, Alán Aspuru-Guzik: Inverse molecular design using machine learning: Generative models for matter engineering, *Science* 2018 **6400**, 360 – 365.

[9] Daniel Schwalbe-Koda, Rafael Gómez-Bombarelli: Generative Models for Automatic Chemical Design, In: Kristof T. Schütt, Stefan Chmiela, O. Anatole von Lilienfeld, Alexandre Tkatchenko, Koji Tsuda, Klaus-Robert Müller (eds) Machine Learning Meets Quantum Physics. Lecture Notes in Physics, vol 968. Springer, Cham 2020.

Dr. Tobias Gensch

Tobias Gensch leitet seit 2020 eine Nachwuchsgruppe an der TU Berlin mit Unterstützung des Exzellenzclusters UniSysCat und eines Liebigstipendiums des FCI. Seine Vision ist es, mittels Computermodellen neue Katalysatoren und neue chemische Reaktionen zu finden. Dazu vereint er Laborchemie, Computerchemie und Data Science zu einem Programm, in dem alle drei Aspekte zusammenarbeiten, um neue Experimente zu designen und die Modellentwicklung und -validierung zu ermöglichen. Insbesondere interessiert ihn dabei die Katalyse mit Metallcarbenkomplexen als besonders vielseitige Reaktionsplattformen. Dieses Forschungsprogramm ist die natürliche Weiterentwicklung und Vereinigung seiner Interessen an Organometallchemie, Entdeckung katalytischer Reaktionen, Computerchemie, Machine Learning und Reaktionsmechanismen.

Er studierte zunächst in seiner Heimatstadt Dresden Chemie. Dort erkannte er die Vorzüge der gleichzeitigen Anwendung von Labor- und Computerchemie zur Untersuchung von Reaktionsmechanismen bereits während seiner Bachelor- und Masterarbeit in der Gruppe von Prof. Hans-Joachim Knölker, die er zur Synthese und Reaktivität von Arylpalladiumkomplexen anfertigte. Seine Doktorarbeit fertigte er mit Prof. Frank Glorius an der WWU Münster an und forschte da zu neuen Methoden zur C-H Aktivierung mit Rhodium- und Cobaltkomplexen und entwickelte dabei sein Interesse am quantitativen Charakter chemischer Daten, zum Beispiel durch Vergleich von Reaktionsbedingungen mittels dem Robustness Screening. Schließlich ging er als Leopoldina Postdoktorand nach Salt Lake City an die University of Utah, um mit Prof. Matthew Sigman die statistische Analyse chemischer Daten zum Verständnis und zur Vorhersage katalytischer Reaktionen zu untersuchen. Insbesondere interessierten ihn dabei Phosphin-basierte Katalysatoren sowie die Untersuchung nichtkovalenter Wechselwirkungen.