

Paul P. Martin¹
 Nicole Graulich¹
 Benjamin Münch²
 Oliver Tepner²
 Christian Schumburg³
 Andreas Nehring³
 Jenna Koenen⁴
 Marvin Rost⁴

¹Justus-Liebig-Universität Gießen
²Universität Regensburg
³Leibniz Universität Hannover
⁴Technische Universität München

Maschinelles Lernen in der chemiedidaktischen Forschung: Ist Lernen in der Chemie mit maschinellem Lernen messbar wie Temperatur?

Beginn einer epistemologischen Reise

Die Erfindung des Thermometers markierte einen Wendepunkt im wissenschaftlichen Verständnis von Temperatur. Durch die Möglichkeit, ein zuvor immaterielles Phänomen zu quantifizieren, erlaubte das Thermometer die Umwandlung qualitativer Empfindungen in messbare Daten. Wie bei den meisten wissenschaftlichen Innovationen war diese Entwicklung jedoch weder unmittelbar noch geradlinig. Wie Hasok Chang (2004) in seinem Buch „Inventing Temperature“ zeigt, markierte der Einsatz von Thermometern den Beginn einer „epistemologischen Reise“; geprägt von Unsicherheit darüber, was gemessen wurde, wie numerische Werte zu interpretieren sind und ob diese Messungen natürliche Phänomene abbilden können.

Die Bildungsforschung steht heute vor vergleichbaren epistemologischen und methodologischen Herausforderungen. Fortschritte im maschinellen Lernen (ML) ermöglichen es, Daten aus gleichbleibenden Quellen – etwa aus qualitativen Erhebungen, standardisierten Tests oder Logfiles – neu zu analysieren und skalierbar zu interpretieren, wodurch Rückschlüsse auf zuvor nur schwer zugängliche Lernprozesse gezogen werden können (Martin & Graulich, 2023; Zhai, Yin et al., 2020). Dabei verspricht die Anwendung von ML-Methoden neue Analysemöglichkeiten in der Bildungsforschung: Durch die Erkennung von Mustern und die Bewertung ihrer Reliabilität in großen, unstrukturierten Datensätzen ermöglichen sie detaillierte, prozessorientierte Analysen komplexer kognitiver und affektiver Prozesse (Martin et al., 2024, 2025; Rosenberg & Krist, 2021; Roski et al., 2024; Rost et al., 2025; Tschisgale et al., 2023; Zeller & Riese, 2025). Angesichts dieser Potenziale argumentieren verschiedene Wissenschaftler:innen, dass ML-Methoden ein transformatives Potenzial für die Bildungsforschung besitzen (Bauer et al., 2025; Kubsch et al., 2023; Zhai, Haudek et al., 2020). Latente Aspekte des Lernens, die zuvor schwer beobachtbar waren, können nun quantifiziert oder klassifiziert werden, etwa durch sogenannte Word-Embeddings.

Bei der Anwendung von ML in den Naturwissenschaftsdidaktiken bestehen dennoch grundlegende Herausforderungen (Martin & Graulich, 2023; Nehring et al., 2025): Viele der Konstrukte, die ML auswerten soll, sind unscharf operationalisiert und nur lose mit etablierten bildungstheoretischen Konzepten verknüpft, sodass sich Fragen zur theoretischen Fundierung der ML-basierten Erkenntnisse ergeben (Kubsch et al., 2023). Zudem fehlt es derzeit an einem Konsens darüber, wie ML-Modelle in Bildungskontexten entwickelt, angewendet und interpretiert werden sollen (Krist & Kubsch, 2023). Weiterhin existieren kaum allgemein akzeptierte Richtlinien zur Interpretation der von ML erzeugten Klassifizierungen oder Quantifizierungen. Dies führt zu Problemen hinsichtlich der Validität von Forschungsergebnissen mit ML (Kapoor et al., 2024). Angesichts dieser Herausforderungen fehlt eine konzeptuelle

Grundlage dafür, wie Lernen sinnvoll mit ML-Methoden klassifiziert oder quantifiziert werden kann, um theoretisch fundierte Erkenntnisse zu Lernprozessen zu gewinnen. Während sukzessiv Standards für Datenannotation, Modellkalibrierung und Interpretierbarkeit entstehen, befinden sich die empirisch und theoretisch fundierten Weiterentwicklungen von ML-Methoden in der Bildungsforschung noch in einem frühen Stadium (Kubsch et al., 2023). In diesem Stadium ähnelt der Einsatz von ML in der Bildungsforschung der historischen Entwicklung der Thermometrie: Der Einsatz birgt transformatives Potenzial, erfordert jedoch sorgfältige Bemühungen zur Definition, Abstimmung, Standardisierung und kontinuierlichen Weiterentwicklung der involvierten Forschungspraktiken (Erduran & Levrini, 2024). Wir argumentieren, dass der Einsatz von ML in der Bildungsforschung nicht nur eine technische, sondern vor allem auch eine epistemologische Aufgabe ist. Um diese epistemologische Reise zu begleiten, schlagen wir ein Rahmenmodell zur Integration von ML in die Bildungsforschung vor, das auf der Analogie zwischen der historischen Entwicklung der Thermometrie und den aktuellen Anwendungen von ML basiert. Dabei integrieren wir Überlegungen aus dem Evidence-Centered Design zur Entwicklung von Assessments (Mislevy et al., 2003) sowie allgemeine Konzepte wissenschaftlicher Forschung im Sinne der wissenschaftlichen Methode, beispielsweise geleitet vom naturwissenschaftlichen Gang der Erkenntnisgewinnung. Diese Vernetzung erlaubt die vorgeschlagene Analogie mit erkenntnistheoretischen Verfahren der Wissenschaft zu verbinden. Das Rahmenmodell umfasst sieben Dimensionen, die eine strukturierte Integration von ML ermöglichen (Tabelle 1).

Tabelle 1. Leitfragen zur Integration von Methoden des maschinellen Lernens (ML) in die Bildungsforschung

Dimension	Leitfrage
Problemdefinition	Welche Konstrukte lassen sich mit ML-Methoden messen und wie werden sie – etwa in Programmiersprachen – operationalisiert?
Instrumentierung und Messung	Wie können ML-Modelle entwickelt und kalibriert werden, um Lernprozesse valide, reliabel und objektiv zu erfassen?
Testung und evidenzbasiertes Folgern	Wie können ML-Methoden systematisch in der Bildungsforschung erprobt werden, um durch Klassifizierungen oder Quantifizierungen einen zusätzlichen Informationsgewinn zu generieren, der ohne den Einsatz von ML nicht möglich wäre?
Vergleich und Replikation	Wie können ML-Methoden genutzt werden, um Forschungsergebnisse über Forschungssettings hinweg zu replizieren?
Konsensbildung	Welchen Konsens kann Bildungsforschung über Richtlinien – z. B. für Modelltraining, Benchmark-Datensätze oder Evaluationsprotokolle – entwickeln, damit ML-Methoden mit anerkannten Standards übereinstimmen?
Implementierung und deren Konsequenzen	Wie beeinflussen ML-basierte Forschung sowie deren Implementierungen und Erkenntnisse Lernende, Lehrende und Forschende?
Kontinuierliche Weiterentwicklung	Inwiefern müssen ML-Methoden weiterentwickelt werden, wenn sich das empirisch fundierte Verständnis ihrer Einsatzmöglichkeiten und der zugrunde liegenden Lernprozesse vertieft?

Einordnung der Vorträge anhand der Dimensionen der epistemologischen Reise

Im Folgenden greifen wir die sieben Dimensionen des epistemologischen Rahmenmodells auf, um die Beiträge der im Symposium vorgestellten Studien zu diesen Dimensionen zu skizzieren.

Im ersten Vortrag des Symposiums präsentierten Martin et al. potenzielle Richtlinien zur Untersuchung der Vertrauenswürdigkeit von ML-Modellen in der chemiedidaktischen Forschung. Diese Richtlinien können zur Entwicklung, Validierung und Berichterstattung von ML-Modellen mit Fokus auf Vertrauenswürdigkeit eingesetzt werden. Langfristig bieten sie das Potenzial der *Konsensbildung* für den Einsatz von ML-Modellen in der chemiedidaktischen Forschung. Eine solche Konsensbildung könnte die Entwicklung von ML-Modellen fördern, die belastbare *Vergleiche und Replikationen* von Forschungsergebnissen über verschiedene Forschungssettings hinweg ermöglichen. Insgesamt könnte dies dazu beitragen, die derzeit vorherrschende Heterogenität in der Methodenanwendung sowie in den damit verbundenen Forschungsdesigns zu strukturieren.

Im anschließenden Vortrag präsentierten Münch und Tepner den aktuellen Stand eines Projekts zum automatisierten Assessment schriftlicher Unterrichtsreflexionen von Chemie-Lehramtsstudierenden. Ziel ist die Entwicklung eines webbasierten Tools, das Studierende während eines Seminars zur Reflexionskompetenz durch Feedback unterstützt und gleichzeitig Lehrende entlastet. Eine klare *Problemdefinition* von Reflexionskompetenz, beispielsweise in Bewertungsrubriken, ist für eine solche Integration von ML in automatisierte Assessments notwendig. Dabei zeigt sich, dass die Qualität und Quantität der kodierten Reflexionen für das Training des Modells und nachfolgende Analysen relevant sind. Die praktische Nutzung des dabei entstehenden Tools fällt wiederum in die Dimension *Implementierung und deren Konsequenzen*.

Im dritten Beitrag des Symposiums präsentierten Schumburg et al. zwei Ansätze – Fine-Tuning und Prompting – zur Übertragbarkeit einer automatisierten Klassifikation von sogenannten „Talk Moves“ in den deutschsprachigen naturwissenschaftlichen Unterricht. Explorativ wurden beide Verfahren in einem experimentellen Design erprobt und im Hinblick auf die Übereinstimmung mit menschlichen Klassifikationen verglichen. Diese *Instrumentierung und Messung* diente dazu, ML-Methoden gezielt für fachspezifische Probleme zu validieren. Aufbauend auf dieser Validierung können Lernprozesse systematisch erfasst und quantifiziert werden. *Testung und evidenzbasiertes Folgern* sind wiederum von besonderer Relevanz, da die zugrunde liegenden Modelle weitgehend im Sinne einer „Black Box“ operieren. Zudem sind empirische Evidenzen darüber, welche Ansätze zur Weiterentwicklung von Sprachmodellen für die Unterrichtsanalyse zielführend sind, notwendig.

Koenen et al. untersuchten abschließend die Frage, ob naturwissenschaftliche Denk- und Arbeitsweisen im Chemielabor durch den Einsatz von Think-Aloud-Protokollen und modernen ML-Verfahren im Sinne der *Problemdefinition* abbildbar sind. Dazu wurden Transkripte von Studierenden ($N = 14$) aus jeweils drei problemorientierten Titrationsexperimenten re-analysiert. Ziel war es, ein ML-Modell mit Hilfe von Primärstudien-codes aus einer menschlichen qualitativen Inhaltsanalyse zu trainieren und zu evaluieren. Die Ergebnisse zeigten, dass ML-Ansätze zentrale Prozessschritte von Scientific Inquiry mit mittlerer bis hoher Übereinstimmung replizieren und zeitaufgelöst darstellen können, wenngleich die kategoriale Trennschärfe eine Herausforderung darstellt. Der Ausblick rahmte das Zusammenspiel von menschlicher und automatisierter Analyse als *Testung und evidenzbasiertes Folgern*, die die chemiedidaktische Forschung zukünftig ergänzen und erweitern.

Literatur

- Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K. & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, 37(2), Article 45.
- Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
- Erduran, S. & Levrini, O. (2024). The impact of artificial intelligence on scientific practices: An emergent area of research for science education. *International Journal of Science Education*, 46(18), 1982-1989.
- Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E. ... & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18), Article eadk3452.
- Krist, C. & Kubsch, M. (2023). Bias, bias everywhere: A response to Li et al. and Zhai and Nehm. *Journal of Research in Science Teaching*, 60(10), 2395-2399.
- Kubsch, M., Krist, C. & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, 60(2), 423-447.
- Martin, P. P. & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice*, 24(2), 407-427.
- Martin, P. P., Kranz, D. & Graulich, N. (2025). Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry. *International Journal of Artificial Intelligence in Education*, 35(3), 1465-1503.
- Martin, P. P., Kranz, D., Wulff, P. & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, 61(8), 1757-1792.
- Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003 (1), i-29.
- Nehring, A., Buschhüter, D., Kubsch, M., Ludwig, T., Wulff, P. & Neumann, K. (2025). Künstliche Intelligenz in den Naturwissenschaftsdidaktiken – gekommen, um zu bleiben: Potenziale, Desiderata, Herausforderungen. *Zeitschrift für Didaktik der Naturwissenschaften*, 31(1), Article 2.
- Rosenberg, J. M. & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, 30(2), 255-267.
- Roski, M., Sebastian, R., Ewerth, R., Hoppe, A. & Nehring, A. (2024). Learning analytics and the Universal Design for Learning (UDL): A clustering approach. *Computers & Education*, 214, Article 105028.
- Rost, M., Resch, K. & Lembens, A. (2025). Using computational grounded theory to analyze pre-service chemistry teachers' reflective practice regarding technology integration in classrooms within a service-learning-oriented seminar. *Journal of Science Education and Technology*, 1-22.
- Tschisgale, P., Wulff, P. & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, 19(2), Article 020123.
- Zeller, J. & Riese, J. (2025). Competency profiles of PCK using unsupervised learning: What implications for the structures of pPCK emerge from non-hierarchical analyses? *Journal of Research in Science Teaching*, 1-22.
- Zhai, X., C Haudek, K., Shi, L., H Nehm, R. & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, 57(9), 1430-1459.
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C. & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, 56(1), 111-151.