Chemie am Computer – Künstliche Intelligenz für Reaktionen

Datenerzeugung für das Maschinelle Lernen einer universellen Potentialenergiefläche

***O. T. Unke, K.-R. Müller**, Institut für Softwaretechnik und Theoretische Informatik, Maschinelles Lernen, Technische Universität Berlin*

Kurzgefasst

- Die quantenchemischen Eigenschaften von Molekülen können nur mit enormen Aufwand berechnet werden, wodurch in der Praxis nur Merkmale kleiner Moleküle genau bestimmt werden können
- Mit Methoden des Maschinellen Lernens ist es möglich diesen aufwendigen Rechenprozess zu umgehen und das Ergebnis vorherzusagen – vorausgesetzt es gibt geeignete Beispieldaten, aus denen gelernt werden kann
- Ein Datensatz, welcher alle denkbaren chemischen Verbindungen abdecket, ist von enormen Wert für die Wissenschaft

Vor etwa einhundert Jahren entwickelte Erwin Schrödinger die nach ihm benannte Schrödingergleichung. Sie enthält alle physikalischen Gesetze, die für die Beschreibung der Chemie notwendig sind. Das Lösen der Schrödingergleichung erlaubt es die Energie einer beliebigen Anordnung von Atomen zu bestimmen. Leider kann sie nur für besonders einfache Fälle, wie etwa ein einzelnes Wasserstoffatom, exakt gelöst werden. Für Moleküle aus mehreren Atomen müssen Näherungen angewendet und die Schrödingergleichung approximativ gelöst werden.

Doch selbst näherungsweise Lösungen sind extrem rechenintensiv: Je nach Anzahl der betrachteten Atome und gewünschter Genauigkeit kann es mehrere Stunden, Tage oder sogar Jahre dauern um die Energie eines Systems zu bestimmen. Das ist vorallem dann problematisch, wenn ein dynamischer Prozess, wie etwa eine chemische Reaktion, untersucht werden soll. Für solche Moleküldynamiksimulationen sind viele tausende Energieauswertungen notwendig. Um das aufwendige Lösungsverfahren der Schrödingergleichung zu umgehen werden für solche Anwendungen deshalb sogenannte Potentialenergieflächen (Abb. 1) eingesetzt. Dabei handelt es sich um mathematische Funktionen, die einer Anordnung von Atomen direkt eine bestimmte Energie zuordnen – ohne die Schrödingergleichung zu lösen.

Allerdings sind solche Funktionen äußerst komplex und es ist schwierig Potentialenergieflächen zu finden, welche genaue Vorhersagen ermöglichen. Methoden des Maschinellen Lernens, wie etwa künstliche neuronale Netzwerke, sind für derartige Problemstellungen besonders effektiv [1]. Sie können eine geeignete Funktion zur Darstellung der Potentialenergiefläche anhand von Beispielen eigenständig lernen. Die Vorraussetzung dafür ist, dass genügend Daten zur Verfügung stehen, mit denen das Netzwerk „trainiert“ werden kann.

Die benötigten Trainingsdaten werden typischerweise durch (näherungsweises) Lösen der Schrödingergleichung berechnet – ein aufwendiger Prozess. Zwar ist dieser Aufwand lediglich ein einziges Mal notwendig, allerdings liefert die resultierende Potentialenergiefläche auch nur für chemische Systeme genaue Vorhersagen, die in den Daten abgedeckt sind. In der Praxis bedeutet das, dass jedes Mal, wenn ein neues System untersucht werden soll, auch neue Daten generiert werden müssen. Mit einem Datensatz, der alle chemischen Systeme abdeckt, wäre es möglich eine Funktion für eine universell gültige Potentialenergiefläche zu lernen.

Da die Anzahl denkbarer chemischer Verbindungen prinzipiell unbegrenzt ist, scheint es auf den ersten Blick unmöglich einen solchen Datensatz zu erzeugen. Bei näherer Betrachtung des Problems lässt sich die tatsächlich benötigte Datenmenge aber erheblich einschränken. Die Wechselwirkung zwischen Atomen wird mit zunehmendem Abstand immer schwächer und kann ab einer gewissen Distanz vollständig vernachlässigt werden. Aus diesem Grund muss der Datensatz lediglich Verbindungen bis zu einer gewissen Maximalgröße abdecken. Außerdem liefern Atome in einer ähnlichen chemischen Umgebung ähnliche Beiträge zur Energie. So wurde zum Beispiel gezeigt, dass es anhand von Strukturen aus nur sieben Atomen (ohne Wasserstoffatome zu zählen) möglich ist, größere Moleküle genau vorherzusagen [2]. Da viele verschiedene Moleküle ähnliche „Substrukturen“ enthalten (Abb. 2) und für eine gegebene Maximalgröße nur eine begrenzte Anzahl an solchen Strukturen chemisch sinnvoll ist, lässt sich systematisch ein Datensatz aufbauen, der alle denkbaren Moleküle abdeckt.

Dass ein solches Vorhaben funktioniert wurde bereits für Proteine – eine bestimmte Gruppe von chemischen Verbindungen, die besonders für biologische Prozesse relevant sind – erfolgreich gezeigt [3]. Für Proteine, die aus den zwanzig natürlich vorkommenden Aminosäuren bestehen, sind bereits 2307 verschiedene Moleküle ausreichend um alle möglichen Substrukturen mit maximal acht Atomen (oh-

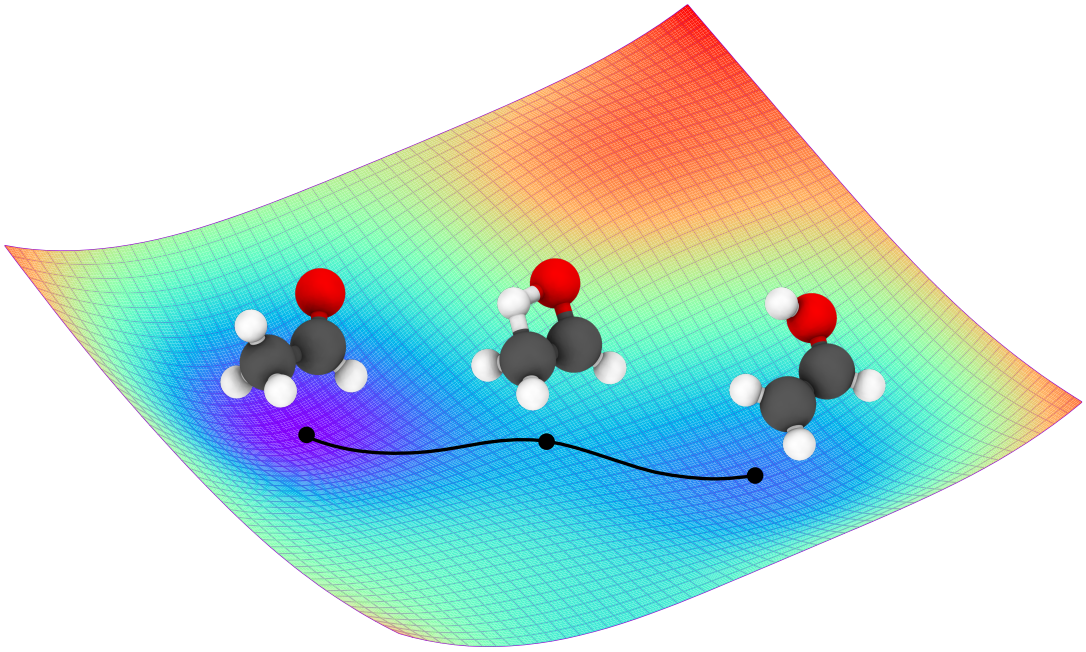

Abbildung 1: *Schematische Darstellung der Reaktion von Ethanal (links) zu Ethenol (rechts) auf der Potentialenergiefläche von C_2H_4O (violett: niedrige Energie, rot: hohe Energie). Die Dynamik der Atome wird durch die Form der Potentialenergiefläche bestimmt, so wie bei einer Kugel, die durch eine Hügellandschaft rollt. Anordnungen der Atome mit niedriger Energie – Moleküle – entsprechen „Tälern" auf der Potentialenergiefläche. Bei einer chemischen Reaktion „rollen" die Atome über einen „Bergsattel", den sogenannten Übergangszustand (mitte), von einem „Tal" ins nächste (schwarzer Pfad).*

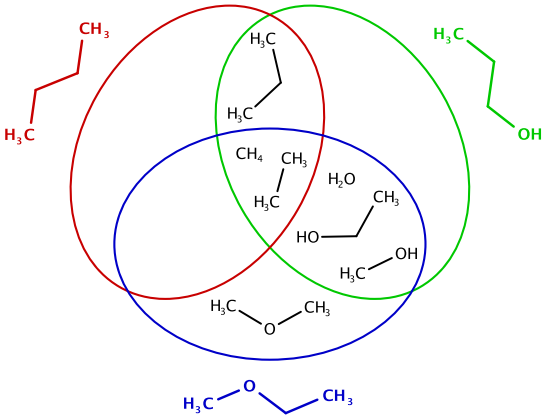

Abbildung 2: *Substrukturen (schwarz) für drei unterschiedliche Moleküle (rot, grün, blau). Die farbigen Ellipsen zeigen an welche Fragmente in den jeweils gleichfarbigen Molekülen vorkommen. Die meisten Substrukturen kommen in zwei oder sogar allen Molekülen vor.*

ne Wasserstoffatome zu zählen) abzudecken. Ein künstliches neuronales Netzwerk, welches mit einem aus diesen Molekülen aufgebauten Datensatz trainiert wurde, ist in der Lage den Faltungsprozess von Deka-Alanin – ein Protein, welches oft als Modellsystem verwendet wird und aus über einhundert Atomen besteht – vorherzusagen. Dabei entdeckte das Netzwerk sogar eine neuartige „Kranz"-Struktur, welche eine ähnliche Stabilität wie die bereits bekannte Helix-Struktur aufweist. Eine universelle Potentialenergiefläche, die das Ziel des Projekts ist, würde ähnliche Untersuchungen für alle Bereiche der Chemie ermöglichen.

WWW

https://www.ml.tu-berlin.de/menue/members/klaus_robert_mueller

Weitere Informationen

[1] K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller, A. Tkatchenko, *Nat. Commun.* **8**, 13890 (2017). doi:10.1038/ncomms13890

[2] B. Huang, O. A. von Lilienfeld, *arXiv preprint arXiv:1707.04146* (2017).

[3] O. T. Unke, M. Meuwly, *J. Chem. Theory Comput.* **15**(6), 3678–3693 (2019). doi:10.1021/acs.jctc.9b00181

Förderung

KRM wird von Projekten des BMBF und der DFG gefördert. OTU wird vom Schweizerischen Nationalfonds unterstützt.