Foto: Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)

Im Gespräch: Fabian Theis, München

# „Das ist der Zyklus der Systembiologie"

**Fabian Theis ist Leiter des Instituts für Computational Biology am Helmholtz Zentrum München. Im *Laborjournal*-Gespräch erklärt er, wie maschinelles Lernen und neuronale Netze dabei helfen, komplexe biologische Datensätze zu verstehen.**

Ein Computerprogramm hat kürzlich in einem Turnier über fünf Spiele einen der besten Spieler des asiatischen Brettspiels Go geschlagen. Wegen der Komplexität des Spiels hatte man es früher für quasi unmöglich gehalten, dass künstliche Intelligenz einem menschlichen Profi in diesem Strategiespiel einmal überlegen sein könnte. Die mächtigen lernenden Algorithmen, mit denen die Go-Software arbeitet, sind indes auch ein wichtiges Werkzeug der Systembiologen. Wie Fabian Theis unter anderem erzählt...

*Laborjournal: Herr Theis, Sie haben Physik und Mathematik in Regensburg studiert, forschen jetzt aber als „Computational Biologist". Woher kam denn Ihr Interesse an der Biologie?*

**Fabian Theis**: Das war bei mir ehrlich gesagt seit der Schulzeit zunächst eher schwach ausgeprägt. Aber ich habe mich immer sehr für Algorithmen und für das Programmieren interessiert. Meine Diplomarbeit in Physik habe ich dann aber in einer biophysikalischen Arbeitsgruppe gemacht, bei Elmar Lang an der Universität Regensburg. So kam ich zum Thema der neuronalen Netze.

*Neuronale Netze – da müssen wir jetzt kurz klären, was damit gemeint ist.*

**Theis**: Ein neuronales Netz, in seiner einfachsten Form, ist eine nichtlineare Funktion, die eine Reihe von Beobachtungen im Hinblick auf eine Antwort interpoliert. Ein Beispiel: Für einen Informatikwettbewerb meiner Schule hatten wir damals die Aufgabe, handschriftliche Ziffern zu klassifizieren. Jeder Mensch schreibt Ziffern ja ein wenig anders. Wie kann ein Algorithmus diese Zeichen möglichst korrekt unterscheiden?

*Was geht gerade noch als eine „Eins" durch, und was ist vielleicht eher eine schlampig hingeschmierte „Sieben" oder eine „Zwei"?*

**Theis**: Genau. Bei der Beantwortung solcher Klassifizierungs-Fragen haben wir mit künstlicher Intelligenz und Machine Learning in den letzten Jahrzehnten riesige Fortschritte gemacht.

*Mit der Methode des Machine Learning kann sich ein Algorithmus solche Unterscheidungsregeln selbst anhand von Beispielen erarbeiten?*

**Theis**: Es gibt da im Prinzip zwei Richtungen. Da ist zum einen das „überwachte" Machine Learning. Man gibt dabei Beispiele mit einer Zielfunktion an – also im Fall des Ziffern-Problems: Was sind Zweier, Dreier, Vierer und so weiter? Man könnte dem Computer nun Regeln vorgeben, die für möglichst viele Ziffern eindeutig sind. Wenn ich diese Regeln selber schreibe, verpasse ich aber viele Spezialfälle.

„Unüberwachte" Lernverfahren dagegen sind eher vergleichbar mit der Lernweise eines Babys. Bei dieser Art des Machine Learning lernt der Algorithmus, Muster in den Daten zu erkennen. Man zeigt dem Computer einfach viele Beispiele, wie verschiedene Menschen Ziffern schreiben, und der Algorithmus entwickelt selbst Regeln aus diesen Beispielen heraus. Computer sind mittlerweile sehr gut darin, solche Regeln zu erkennen.

*Und wie hängen nun die neuronalen Netze damit zusammen?*

**Theis**: Künstliche neuronale Netze beschreiben eine bestimmte Subklasse von Methoden des Machine Learning.

*Man sollte an dieser Stelle vielleicht zur Sicherheit erwähnen, dass neuronale Netze – entgegen des Namens – nicht viel mit biologischen Nervenzellen zu tun haben, das ist ja nur eine Metapher.*

**Theis**: Genau, mit Neurowissenschaft hat das erst mal nichts zu tun. Ein neuronales Netz ist eigentlich nichts anderes als eine Operationseinheit, die eine Reihe von Signalen aufnimmt, aufaddiert, verarbeitet, und dann einen Output generiert. Anfang der 2000er Jahre waren die neuronalen Netze quasi verschwunden, weil die nötige Computerleistung noch nicht da war. Aber dank der heutigen Rechenpower und dem Zuwachs der Big Data haben die neuronalen Netze diverse Forschungsfelder revolutioniert – beispielsweise das Gebiet der Computervision, also der Klassifizierung von Bildern durch Algorithmen. Früher war es hart, einem Computer beizubringen, Objekte in Bildern zu finden –, so dass der Algorithmus beispielsweise Möbelstücke unterscheiden kann: „Das ist ein Schrank, das ist ein Stuhl." Heute gelingt das schon ziemlich gut. Noch ein aktuelles Beispiel: Kürzlich hat ein Computer dank neuronaler Netze zum ersten Mal einen Profi-Spieler des Brettspiels Go geschlagen. Das hatte man lange Zeit gar nicht für möglich gehalten.

**„Der Algorithmus entwickelt selbst Regeln aus den Beispielen heraus. Computer sind mittlerweile sehr gut darin, solche Regeln zu erkennen."**

*Welche biologischen Fragen kann man mit solchen lernenden Algorithmen bearbeiten? Und welche neuen Erkenntnisse kann man durch Modelle generieren, die man durch reines Betrachten der Daten nicht gewinnen kann?*

**Theis**: In unserer Gruppe interessieren wir uns beispielsweise für zelluläre und molekularbiologische Datensätze. Die „Buchstaben" der Molekularbiologie verstehen wir ja schon recht gut, das ist ein günstiger Ausgangspunkt. Wir wissen mittlerweile ziemlich genau, wie Gene aufgebaut sind, wie der Weg vom Gen über mRNA zum Protein führt, inklusive Splicing und so weiter. Wir lernen auch verschiedene Zelltypen und ihre Entstehung immer besser zu verstehen, und wir wissen recht gut, welche Mechanismen dabei aktiv sind. Aber das Zusammenwirken der Prozesse ist sehr komplex und durch reines Analysieren einzelner Datenreihen oft nicht mehr zu begreifen. Die Frage für Systembiologen ist also: Wie kann man diese Prozesse strukturierter, in einem größeren Ansatz analysieren...

*.. und dadurch komplexe Interaktionen und Feedback-Mechanismen besser verstehen?*

**Theis**: Genau. Welche Daten wir dabei als Ausgangspunkt zur Verfügung haben, ist ziemlich Technologie-getrieben. Wir können heute mittels Hochdurchsatz-Technologie ganze Genome zu einem Bruchteil der Kosten des damaligen Humangenomprojekts sequenzieren. Die Daten werden komplexer und vielfältiger. Für das Machine Learning bedeutet das: Wir haben heute nicht nur mehr Rechenpower, sondern auch viele Beispiele, an denen unsere Algorithmen lernen können.

*Hängt die Theorie dem Strom der Big Data da eigentlich hinterher? Johannes Jäger, theoretischer Evolutionsbiologe und Leiter des Konrad-Lorenz-Instituts, hat sich an dieser Stelle vor kurzem beklagt, dass die Theorie in der Biologie weniger Wertschätzung erfahre als die Generierung immer neuer Datenberge (siehe* Laborjournal *11/2015: 20-22).*

**Theis**: Es gibt einen klaren Trend zur Computational Biology. Die Physiker setzen die Mathematik seit langem erfolgreich als universelle Sprache zur Beschreibung der Welt ein und finden damit prädiktive Regeln. So etwas möchten wir in der Biologie auch erreichen, das ist unsere Vision Zum Beispiel wollen wir irgendwann mal *In Silico*-Drug Design machen.

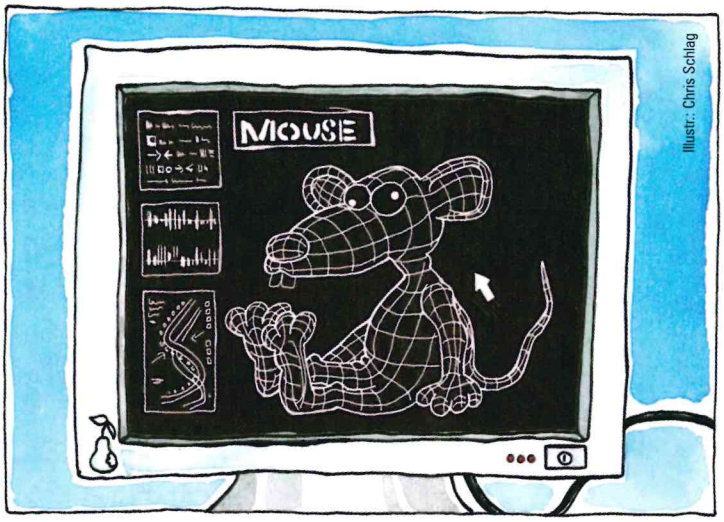
*Man möchte also vorhersagen, welche Wirkungen eintreten, wenn ein Patient einen neuen Wirkstoff einnimmt – ohne ein Experiment zu machen?*

**Theis**: Genau. Dose-Response, Efficiency,... – all diese Parameter. Ein anderes Beispiel, das uns auch in unserer Arbeitsgruppe interessiert: Wie kann man das Entwicklungsprogramm eines Organismus' beschreiben? Wie entstehen die verschiedenen Zelltypen, und wie kann man diese Zelltypen definieren? Auch dabei helfen Modelle und Algorithmen.

*Biologisch macht man das zum Beispiel mit Oberflächenmarkern.*

**Theis**: Ja, aber wenn man dann genauer hinschaut, entdeckt man, dass scheinbar homogene Populationen, die einen Zelltyp beschreiben sollen, viel heterogener sind als gedacht – und als es uns der Marker zeigt. Diese Heterogenität kann man mit Methoden der Computational Biology systematisch analysieren.

Bei uns läuft das meistens so ab: Am Anfang steht häufig eine statistische Analyse des Datensatzes – wir beschreiben also die Struktur in den Daten. Im zweiten Schritt bauen wir dann das Modell, das diese Daten einbezieht. Das ist ein wichtiger Unterschied zur klassischen theoretischen Biologie oder Biomathematik. Auch früher hatte man ja spannende Modelle aufgestellt, um Biologie mit mathematischen Mitteln zu beschreiben. Aber die Modelle waren damals oft wenig quantitativ, und man konnte kaum Vorhersagen daraus ableiten – zum Beispiel, weil die Modelle Parameter hatten, die man experimentell gar nicht zu fassen bekam. Ich glaube aber, es ist entscheidend, dass wir heute die




„Es ist nicht sinnvoll, von dem ‚wahren Modell' zu sprechen. Das Modell ist *per definitionem* eine Abstraktion der biologischen Wirklichkeit."

klassische Biomathematik, mit ihren mechanistischen Modellen, erweitern um die Statistik. Denn die Statistik brauchen wir, um die Modelle mit den experimentellen Daten zu verbinden.

*Dann ist es gar nicht mal so schlecht, wenn ein Modell dabei mal „scheitert", also durch ein Experiment widerlegt wird?*

**Theis**: Das ist ein wichtiger Punkt. Es ist eigentlich nie sinnvoll, von dem „wahren Modell" zu sprechen. Das Modell ist *per definitionem* eine Abstraktion der biologischen Wirklichkeit. Das heißt, da wird immer ein Fehler drin sein. Wir wollen auch gar nicht 1:1 die Daten wiedergeben, die biologische Wirklichkeit also nur abbilden. Denn dann verstehen wir mechanistisch nicht viel Neues. Wir wollen, wir *müssen* also approximieren. Ich kann die Richtigkeit eines Modell letztlich nie beweisen. Ich kann zwar versuchen zu zeigen, dass meine Vorhersage eintritt. Aber das zeigt mir noch nicht, dass mein Modell wirklich richtig ist – vielleicht habe ich einfach den falschen Versuch gemacht.

Aber wenn meine Vorhersage *nicht* eintritt, weiß ich sicher: Mein Modell stimmt so nicht. Ich vermeide jedenfalls zu sagen: „Das ist jetzt das richtige Modell." Ich kann aber schon sagen: „Das ist das wahrscheinlichste Modell *für meine Daten*."

*Wie problematisch ist für die Modellierung das Rauschen in typisch biologischen Big Data, wie etwa bei Hochdurchsatz-Transkriptionsdaten oder ähnlichen Methoden?*

**Theis**: Es ist mittlerweile viel besser geworden, vor allem durch neue Technologien wie Deep Sequencing oder High-Throughput Imaging. Aber wir haben in der Biologie diese krassen Effekte durch die Einzelfall-Natur der Experimente. Es gibt viele Parameter, die ganz spezifisch für den jeweiligen Versuch sind. Das kann das Handling im Labor sein oder die Einstellung der Geräte. Im schlimmsten Fall hängt das Resultat auch noch von der Person des jeweiligen Experimentators ab. Diese *Confounding Factors,* also Störeinflüsse, können es uns schwer machen, Schlüsse zu ziehen, die tatsächlich über das einzelne Experiment hinausgehen. Denn unser Ziel ist ja eigentlich, in einem Modell viele Datensätze aus vielen Experimenten zu integrieren.

Dazu ein Beispiel: Wir hatten für ein Projekt zusammen mit Epidemiologen Blutproben analysiert. Werden die Proben nicht alle im selben Labor analysiert, haben wir schon ein Problem. Das Alter und das Geschlecht der Probanden spielen auch eine Rolle. Und wenn man Stoffwechselprodukte untersucht, sieht man in den Daten, ob der Proband kurz vorher gegessen hat oder nicht. Wenn man das vermeidet, indem alle Probanden vor der Blutabnahme fasten, dann spielt vielleicht das Gewicht eine Rolle.

Wir arbeiten auch viel mit Einzelzellversuchen, und da gibt es solche Effekte natürlich ebenfalls. Der Zellzykluszustand beispielsweise ist so ein Faktor, den man

berücksichtigen muss, denn das Zellwachstum kann man selten hundertprozentig synchronisieren.

All diese Störfaktoren muss man herausrechnen. Und für diese Korrekturen brauchen wir nicht nur die klassische Modellierung, sondern eben auch die Biostatistik – und diese beiden Methoden bringen wir zusammen. Für mich gehen Statistik und Modellierung Hand in Hand.

Ein systembiologisches Modell sollte ja prädiktiv sein, das heißt, in der Praxis muss mein Modell in der Lage sein, etwa die Konzentration eines Stoffs zu einem bestimmten Zeitpunkt vorherzusagen. Diese Vorhersage ist dann wiederum durch neue Daten validierbar.

Das ist, kurz gesagt, der „Zyklus der Systembiologie", wir nähern uns so immer mehr dem Punkt, wo wir tatsächlich das biologisch relevante Wissen vorfinden.

*Sie haben ja schon Beispiele aus Ihrer Arbeit erwähnt. Welche biologischen Fragen interessieren Sie noch?*

**Theis**: Ein Thema, das wir gerade mit Timm Schröder von der ETH Zürich bearbeiten, dreht sich um Hämatopoese, die Bildung der verschiedenen Blutzelltypen aus Blutstammzellen. Wann wird diese Entscheidung über die Zugehörigkeit zu einer bestimmten Zelllinie getroffen? Um das zu beantworten, können wir beispielsweise Proteinexpression verfolgen, oder – und das ist das Neue an unserer Studie – wir können auch die Änderungen in der Morphologie der Zellen verfolgen. Die Zellformen können wir heranziehen, um vorherzusagen, welcher Blutzelltyp aus einer Vorläuferzelle entstehen wird. Auf dieser Idee aufbauend haben wir ein tiefes neuronales Netz trainiert, mit dem wir den zukünftigen Zelltyp einer Linie sehr früh vorhersagen können, einfach aufgrund der Morphologie und der Wachstumsgeschwindigkeit – und schon bis zu drei Generationen, bevor der jeweilige Zelltyp-Marker sichtbar ist.

**„Für das Machine Learning bedeutet das: Wir haben heute nicht nur mehr Rechenpower, sondern auch viele Beispiele, an denen unsere Algorithmen lernen können."**

Wir arbeiten aber auch mit menschlichen Kohorten-Daten, beispielsweise zusammen mit Anette Ziegler hier am Institut für Diabetesforschung des Helmholtz Zentrums. Bei diesem Projekt erstellen wir einen genetischen Risk-Score, basierend auf Daten von mehreren Loci, um möglichst präzise den Beginn von Typ-1-Diabetes bei Kindern vorherzusagen. Spannend ist hier die Integration komplexer Daten für immer bessere Vorhersagen, auch im klinischen Alltag.

*Das ist wohl auch ein Charakteristikum von Theoretikern in der Biologie: Dass sie mit ganz unterschiedlichen Arbeitsgruppen und Themen in Kontakt kommen?*

**Theis**: Stimmt, da muss man auch ein wenig darauf achten, wo man sich seine Expertise aufbaut. Ich finde es aber immer sehr spannend, mit verschiedenen Partnern aus der Biologie zusammenzuarbeiten. Und ich freue mich immer auf die kreativen Fragen, die dabei auf einen zukommen.

*INTERVIEW: HANS ZAUNER*