

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Zuwendungsempfänger

Hasso-Plattner-Institut für Digital Engineering gGmbH

Thema der Förderung

DeepPath - Deep learning based identification of pathogens from next generation sequencing data im Computational Life Science Call des BMBF

Verantwortliche

Prof. Dr. Bernhard Renard

Förderkennzeichen

031L0248 CompLS2

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Schlussbericht zu Nr. 3.2
Für das Projekt am HPI zum Antrag
DeepPath - Deep learning based identification of pathogens from next generation sequencing data
im Computational Life Science Call des BMBF
Förderkennzeichen 031L0248 CompLS2

Prof. Dr. Bernhard Renard

I. Kurze Darstellung zu

1. Aufgabenstellung,

Im Rahmen dieses Projekts sollte das Potenzial von Deep-Learning-Ansätzen nutzen, um das pathogene Potenzial von Nukleotidsequenzen vorherzusagen, um neue mikrobielle Krankheitserreger oder Risiken aus synthetischen biologischen Experimenten zu erkennen

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde,

In einer global vernetzten und dicht besiedelten Welt können sich Krankheitserreger leichter ausbreiten und entwickeln als sie es je zuvor getan haben. Wie die jüngsten Ausbrüche der Ebola- und Zika-Viren gezeigt haben, sind die Risiken selbst für diese bisher bekannten Agenten unberechenbar und ihre Ausbreitung schwer zu kontrollieren. Außerdem ist es fast sicher, dass noch unbekanntere Erregerarten und -stämme entdeckt werden müssen, aufgrund ihrer ständigen, extrem schnellen Entwicklung und unerforschten Biodiversität sowie der zunehmenden Verbreitung menschlicher Ressourcen. Einige dieser neuartigen Krankheitserreger können Epidemien verursachen (ähnlich dem Ausbruch des Jahres 2011 von Shiga-toxischer Escherichia coli O104:H4-Stamm in Deutschland oder die Ausbrüche des SARS- und MERS-Coronavirus. 2002 und 2012) oder sogar Pandemien (z.B. der "Schweinegrippe" H1N1/09 Influenza-A-Stamm). Viele Erreger können mehr als einen Wirt haben, was die Bewertung und Vorhersage der Risiken noch schwieriger macht. Da der hochmoderne Ansatz für den Open-View-Nachweis von Krankheitserregern die DNA-Sequenzierung ist, ist er entscheidend für die Entwicklung automatisierter Pipelines zur Charakterisierung des pathogenen Potenzials von derzeit nicht identifizierbaren Sequenzen.

Das Screening gegen potenziell gefährliche Sequenzen kann auch als Mittel zur Sicherstellung der verantwortungsvollen Forschung in der Synthetischen Biologie gesehen werden. Während die Absichten und Ambitionen der synthetischen Biologie im Allgemeinen wohlwollend sind, führt das Verändern von lebenden Organismen und Viren zu vielen Kontroversen. Zwei kontroverse Studien modifizierten den Influenza A/H5N1 ("Vogelgrippe") Virus, das bei Säugetieren in der Luft übertragbar sein soll. Darüber hinaus bieten sich Möglichkeiten das Host-Spektrum von Viren (z.B. Baculoviren) zu erweitern. Eine erste Studie zur erfolgreichen Synthese eines Baculovirus zitierte die Konstruktion von Pockenviren als mögliche Anwendung. Genetische Veränderungen von Filamentpilzen werden ebenfalls immer beliebter. Wie einige Pilzarten sind gefährliche Krankheitserreger, was zur Entwicklung neuer Dual-Use-Anwendungen führen kann und potenzielle bioterroristische Risiken trägt. Diese Fragen müssen berücksichtigt werden und werden von der wissenschaftlichen Community selbst ausdrücklich angesprochen, aber gesetzliche Regelungen, Richtlinien und bioethische Diskussionen sind weiterhin notwendig. Wissenschaftlich, objektiv und genaue Methoden zur Risikokontrolle sind ein Muss, aber es gibt kein festgelegtes Protokoll zur Bewertung von synthetischen Produkten. Da ein Krankheitserreger sowohl zufällig als auch absichtlich entstehen kann (entweder für die Forschung oder den Bioterrorismus), könnte man argumentieren, dass alle neuartigen Sequenzen, einschließlich derjenigen, die von ertrauenswürdigen Experten entwickelt wurden auf Anzeichen von Pathogenität untersucht werden sollten. Eine umfassende Vorstellung davon, wie eine solche Aufsicht aussieht, wurde vom US National Research Council vorgeschlagen. Da sich die Nukleotidsequenzierung zum Stand der Technik im Bereich des Open-View-Erregernachweises entwickelt hat, werden neue Pipelines aufgebaut. und Algorithmen werden benötigt, um die bei jedem Durchlauf anfallende Datenfülle effizient und präzise zu verarbeiten. Die Bewertung und Minderung von Risiken, die allein auf der

DNA-Sequenz basieren, sollte mit Hilfe von Berechnungsmethoden erfolgen, um relevante Patterns zu erkennen und Vorhersagen für neue Sequenzen zu treffen. Daher sind auf das maschinelle Lernen basierte Ansätze eine vielversprechende Alternative zu den traditionellen Werkzeugen der Sequenzanalyse .

3. Planung und Ablauf des Vorhabens,

Das Vorhaben wurde am HPI durchgeführt. Es gliedert sich in insgesamt sieben Arbeitspakete, die bearbeitet wurden. Ursprünglich war geplant, das Projekt am Robert Koch-Institut anzusiedeln, wurde aber vor Projektbeginn durch den Wechsel des Projektleiters am Hasso-Plattner-Institut angesiedelt. In der Durchführung kam es zu Verzögerungen durch die Personalgewinnung und den temporären Ausfall zweier Projektmitarbeiter aus persönlichen Gründen. Ferner kam es insbesondere durch die Corona-Krise zu einigen Änderungen im Projektablauf, gleichzeitig haben wir einige Ansätze direkt auf Corona-Daten angewendet.

4. wissenschaftlichem und technischem Stand, an den angeknüpft wurde,

Es wurde an Vorarbeiten und unserer Deepac Software angeknüpft. Ferner wurde eine umfangreiche Literaturrecherche über pubmed und google-scholar durchgeführt, deren Ergebnisse im Projektantragsdokumentiert sind. In der Zwischenzeit haben keine wesentlichen Änderungen stattgefunden.

5. Zusammenarbeit mit anderen Stellen.

Es gab eine intensive Zusammenarbeit mit der Gruppe von Regina Barzilay am MIT. Hier wurde erfolgreich ein gemeinsamer Projektantrag gestellt, um Ideen des Projekts gemeinsam fortzuführen. Ferner haben wir im Projekt sehr eng mit dem Robert Koch-Institut zusammengearbeitet.

II. Eingehende Darstellung

1. der Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele,

Mit Hilfe der verwendeten Zuwendung wurden Projektmitarbeiter beschäftigt (s. Mittelnachweis) sowie Reisen durchgeführt, wir konnten auf Rechneranschaffungen verzichten, weil wir vorhandene Infrastruktur nutzen konnten, mussten hierfür aber mehr Personalaufwand zur Adaption einsetzen. Die Förderung hat es uns erlaubt, die Technologieentwicklung voranzutreiben. Im Detail wurden die folgenden Schritte durchgeführt (in der Gliederung des ursprünglichen Projektentwurfs mit seinen Zielen).

WP 1: Viral host range prediction with deep learning

Dieses Arbeitspaket wurde erfolgreich abgeschlossen. Wir haben eine Software entwickelt, (deepac-vir), die virale host-range vorhersagt und deutlich den bisherigen State-of-the-art (durch einfache Sequenzvergleiche) übertrifft. Hierbei haben wir sowohl CNN als auch LSTM-Architekturen entworfen, in Pipelines implementiert und getestet, die leichte Unterschiede bzgl. Sensitivität und Spezifität aufweisen (und sich daher für unterschiedliche Einsatzbedingungen eignen), wie in Tabelle 1 dargestellt.

	Bacc.	AUPR	Rec.	Spec.
CNN _{All} (ours)	91.7	91.2	89.3	94.2
LSTM _{All} (ours)	86.3	85.8	96.2	76.4
BLAST (reads)	90.3	n/a	85.5	95.1
k-NN (genome)	82.8	n/a	93.9	71.6
BLAST (genome)	90.5	n/a	86.3	94.6
k-NN (contigs)	83.0	n/a	94.3	71.6
BLAST (contigs)	88.4	n/a	87.1	89.7

Tabelle 1: Balanced Accuracy (Bacc.), Area under Precision Recall Curve (AUPR), Recall (Rec.) und Spezifität (Spec) für unsere CNN and LSTM-Ansätze sowie Vergleich zu bisherigen Ansätzen.

Die Ergebnisse wurden in wissenschaftlichen Journalen publiziert [1], auf Fachtagungen vorgestellt [2,3], die Software wurde unter einer open source Lizenz veröffentlicht [4] und die Öffentlichkeit wurde informiert, bspw. in gesundhyte [5] und dem Tagesspiegel [6]

WP2.Transfer und Multi-Task-Lernen für nicht-virale Sequenzen

Uns ist es gelungen, den Ansatz von Deepac und den Ergebnissen aus WP1 auch für mehrere Erregerklassen auf einmal zu übertragen. Hierfür konnten wir sogar pathogene Pilze in die Erregerbetrachtung einbeziehen und besser als bisher möglich in einem einheitlichen System klassifizieren (vgl. Tabelle 2). Hierfür haben wir die in WP1 beschriebenen Deep-Learning Systeme entsprechend angepasst.

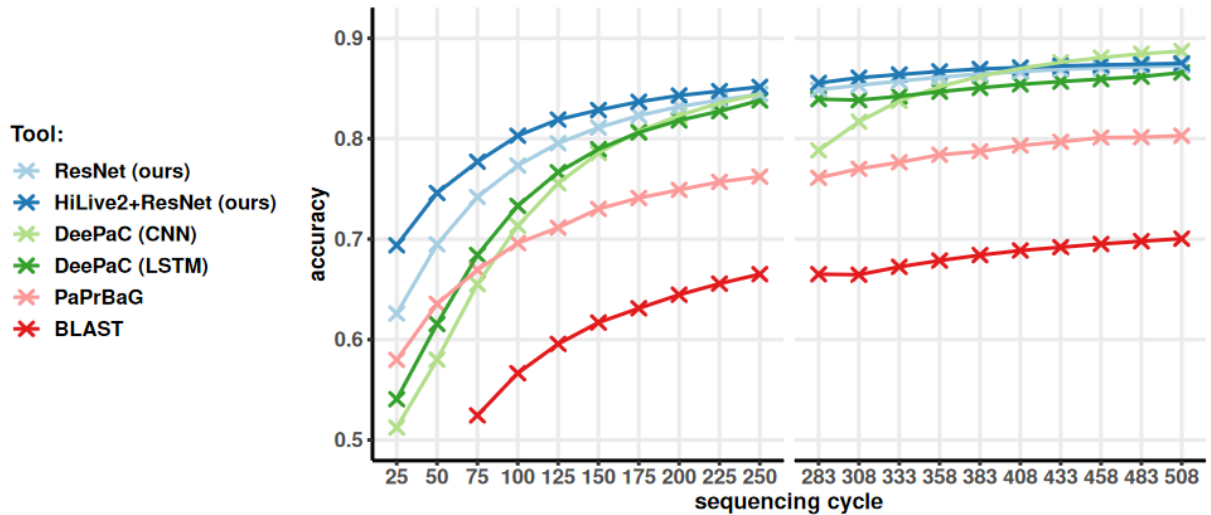
		Acc.	F1	Prec.	Rec.	AUPR
All classes	4-class ens. (ours)	87.6	87.7	87.7	87.6	93.4
	BLAST	78.3	84.0	90.6	78.3	-
Non-pathogens	4-class ens. (ours)	77.4	78.7	80.1	77.4	86.7
	BLAST	66.5	71.6	77.5	66.5	-
Path. bacteria	4-class ens. (ours)	87.2	85.1	83.2	87.2	90.4
	BLAST	83.8	87.5	91.6	83.8	-
Human viruses	4-class ens. (ours)	90.9	93.7	96.7	90.9	98.4
	BLAST	78.9	87.9	99.2	78.9	-
Fungi	4-class ens. (ours)	95.0	92.9	90.9	95.0	97.9
	BLAST	84.1	88.9	94.2	84.1	-

Tabelle 2: Accuracy (Acc.), F1- Score (F1), Precision (Prec.), Recall (Rec.) und Area under Precision Recall Curve (AUPR) für verschiedene Erregerklassen und Vergleich zu Blast (vgl. Tabelle 1) als bisherigen State-of-the-Art.

Die Ergebnisse wurden publiziert [7], genauso wie der entsprechende Open Source Quellcode [8].

WP 3: Neural networks for real-time inference

Wir konnten zeigen, dass sich bereits während der Laufzeit von Sequenzierern eine Einbindung von unseren Verfahren ermöglichen lässt (Figure 1). Hierfür wurden entsprechende Anpassungen an Modellen und Code vorgenommen.



WP 4: Evaluation and practical applications

Unter Verwendung der in WP1.1 und WP2 etablierten Verfahren für Trainingsdaten wurden simulierte Hold-Out-Testdatensätze erzeugt und Realdaten identifiziert. Sie enthalten Viren, Fungi und Bakterien. Eine weitere Fortentwicklung hängt von Fortschritten in WP5 bzgl. der Interpretierbarkeit ab.

WP 5: Interpretability and visualization

WP 5.1: Interpretation of Neural Nets

Aufbauend auf unseren ersten Arbeiten insbesondere mit dem DeepLIFT- und DeepSHAP-Frameworks haben wir uns insbesondere auf die Detektion von genomischen Interaktionsaspekten befasst und konnten hier verschiedene Interaktions-Effekte aus neuronalen Netzen herausziehen und visualisieren, wie auch in Abbildung 1 visualisiert.

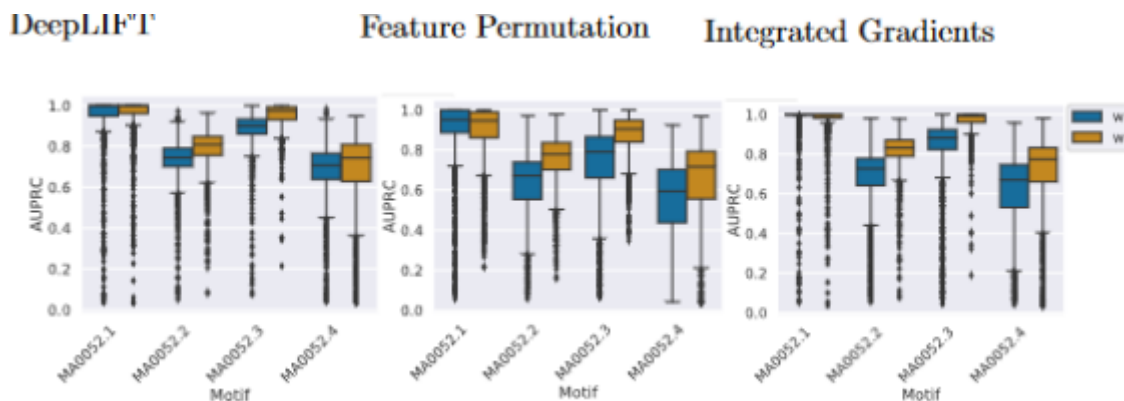


Abbildung 1: Beispiel für die Detektionsgüte verschiedener Interaktionseffekte basierend auf unterschiedlichen Frameworks.

WP 5.2: Analysis of pathogen genomes

Basierend auf den Ergebnissen der verschiedenen Pathogenklassen wurde auch die Interpretierbarkeit auf alle betrachteten Klassen entsprechend angepasst und erweitert. Die Güte ist dabei stark von der Verfügbarkeit der Trainingsdaten abhängig.

WP 6: Software engineering and deployment

Die entwickelten Software-Tools wurden vollständig kompatibel mit Echtzeit-Mapping-Ansätzen (HiLive2) gemacht, und es wurde ein dediziertes Plugin entlang automatisierter Testverfahren implementiert, um eine effiziente Kommunikation zwischen den Mapping- und Deep-Learning-

Modulen neuartiger Pathogenerkennungs-Workflows zu ermöglichen. Eine Anbindung an Nanopore Sequenzierungen wurden eingebunden. Ferner müssen noch Ansätze aus Arbeitspaket 5 nach Abschluss integriert und getestet und dokumentiert werden.

WP7. Management und Verbreitung

Publikationen und Präsentationen auf Fachtagungen wurden erreicht [1-10].

2. der wichtigsten Positionen des zahlenmäßigen Nachweises,

Der mit Abstand größte und für den Projekterfolg wichtigste Punkt des zahlenmäßigen Nachweises waren Projektausgaben in Höhe von ca. 244.000 € für Wissenschaftliche Mitarbeitende, die sowohl Modell- als auch Softwareentwicklung sowie Evaluierung vorangetrieben haben. Dies wurde unterstützt, insbesondere in der Softwareentwicklung durch studentische Hilfskräfte. Anders als geplant bedingt durch die Corona-Pandemie hatten wir keine Kosten für Dienstreisen und durch den Institutionenwechsel keine Infrastrukturkosten.

3. der Notwendigkeit und Angemessenheit der geleisteten Arbeit,

Die Corona-Krise hat uns erneut darin bestärkt, dass die frühzeitige Diagnostik von zoonotischen Infektionskrankheiten zentrale Wichtigkeit besitzt und das von uns entwickelte Framework hierfür eine hohe Relevanz haben kann. Selbst in der Frühphase der Pandemie konnten wir hier vielbeachtete Beiträge mit unserem Framework leisten. Ohne die finanzielle Unterstützung durch das BMBF wäre das Projekt nicht durchführbar gewesen, da keine entsprechenden institutionellen Mittel zur Verfügung standen. Die geleistete Arbeit in den oben beschriebenen WPs war hierfür notwendig und in einem angesichts des potentiellen Nutzen angemessenen Rahmens.

4. des voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans,

Während des ersten Jahrs der Corona-Krise konnten wir bereits den Nutzen unseres Ansatzes zur erklärbaren Charakterisierung von pathogenen Genomregionen aufzeigen und mit Kollegen am RKI zusammen anwenden. Dort ist unsere Software inzwischen auch im Standardsoftwarestack integriert, genauso wie wir sie bereits dem Emerging Infections Team der WHO und der Bill & Melissa Gates Foundation vorstellen konnten.

Wissenschaftlich haben wir basierend auf den Ergebnissen eine Kooperation mit der Arbeitsgruppe von Regina Barzilay am MIT begonnen, wo wir Erkenntnisse aus dem Projekt nutzen, um Therapien gegen Antibiotika-resistente Bakterien zu designen. Hierfür konnten wir erfolgreich Mittel einwerben und die Projektmitarbeitende aus diesem Projekt überführen. Wir sehen hier wissenschaftlich, aber auch vom Impact, ein großes Potential.

Wir planen, wie bereits oben erwähnt, noch Patente für Ideen aus diesem Projekt zu beantragen und haben diesbezüglich bereits eine Freigabe seitens des Hasso-Plattner-Instituts erhalten und mit dem Ausgründungszentrum des HPIs Business Pläne diskutiert. Wir haben von unserem Fachgebiet eine Ausgründung (seqstant GmbH) zur Erreger-Schnelldiagnostik durchgeführt und hierfür im Vorfeld Mittel zur Geschäftsentwicklung beantragt und bewilligt bekommen (Exist Forschungstransfer). Es ist angedacht, dass diese Ausgründung auch Innovationen aus diesem Projekt nutzen wird (bspw. im Rahmen des open source codes).

5. des während der Durchführung des Vorhabens dem ZE bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen,

Wir beobachten mit Interesse die Weiterentwicklung von Lage Language Modells. Aktuell sehen wir aber hier keinerlei Beeinträchtigung der Ziele unseres Vorhabens.

6. der erfolgten oder geplanten Veröffentlichungen des Ergebnisses nach Nr. 6.

Ergebnisse der Arbeit wurden wie folgt unter Erwähnung des Mittelgebers publiziert:

[1] Bartoszewicz, Jakub M., Anja Seidel, and Bernhard Y. Renard. "Interpretable detection of novel human viruses from genome sequencing data." *NAR genomics and bioinformatics* 3.1 (2021): lqab004.

[2] Bartoszewicz, Jakub M., Anja Seidel, and Bernhard Y. Renard. "Interpretable detection of novel human viruses from genome sequencing data AI for Public Health Workshop at ICLR'21

[3] Bartoszewicz, Jakub M, Inaugural Lecture of Lecture Series of the European Virus Bioinformatics Center

[4] <https://gitlab.com/dacs-hpi/DeePaC-vir>.

[5] Vorhersage des Infektionspotenzials neuartiger Viren, *Gesundhyte* 2021:14, 38-39

[6] Tagesspiegel Background Digitalisierung & KI, 12.03.2020

[7] Bartoszewicz, J. M., Nasri, F., Nowicka, M., & Renard, B. Y. (2022). Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection. *Bioinformatics*, 38(Supplement_2), ii168-ii174.

[8] <https://gitlab.com/dacs-hpi/deepac> bzw. <https://gitlab.com/dacs-hpi/deepac-live>

[9] Bartoszewicz, Jakub M., Ulrich Genske, and Bernhard Y. Renard. "Deep learning-based real-time detection of novel pathogens during sequencing." *Briefings in Bioinformatics* 22.6 (2021): bbab269.

Anlagen:

Erfolgskontrollbericht
Berichtsblatt

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Zuwendungsempfänger

Hasso-Plattner-Institut für Digital Engineering gGmbH

Thema der Förderung

DeepPath - Deep learning based identification of pathogens
from next generation sequencing data
im Computational Life Science Call des BMBF

Verantwortliche

Prof. Dr. Bernhard Renard

Förderkennzeichen

031L0248 CompLS2

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Berichtsblatt

1. ISBN oder ISSN geplant	2. Berichtsart (Schlussbericht oder Veröffentlichung) Veröffentlichung
3. Titel Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection	
4. Autor(en) [Name(n), Vorname(n)] Bartoszewicz, Jakub Nasri, Ferdous Nowicka, Melania Renard, Bernhard	5. Abschlussdatum des Vorhabens 31.07.2023
	6. Veröffentlichungsdatum 18.09.2022
	7. Form der Publikation Zeitschriftenartikel
8. Durchführende Institution(en) (Name, Adresse) Hasso Plattner Insitut	9. Ber. Nr. Durchführende Institution --
	10. Förderkennzeichen 031L0248 CompLS2
	11. Seitenzahl 7 +30 Seiten Appendix
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 83
	14. Tabellen 2 (+7 im Appendix)
	15. Abbildungen 1 (+10 im Appendix)
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung <p>Neu auftretende Krankheitserreger stellen eine wachsende Bedrohung dar, aber große Datensammlungen und Ansätze zur Vorhersage des mit neuartigen Erregern verbundenen Risikos sind auf Bakterien und Viren beschränkt. Pathogene Pilze, die ebenfalls eine ständige Bedrohung für die öffentliche Gesundheit darstellen, sind noch zu wenig erforscht. Einschlägige Daten sind nach wie vor vergleichsweise spärlich und über viele verschiedene Quellen verstreut, was die Entwicklung sequenzbasierter Nachweismethoden für neuartige Pilzerreger behindert. Es gibt keine Vorhersagemethode, die für Erreger aller drei Gruppen funktioniert, obwohl die Ursache einer Infektion oft nur schwer anhand der Symptome zu erkennen ist.</p> <p>Wir stellen eine kuratierte Sammlung von Daten zum Wirtsspektrum von Pilzen vor, die Datensätze zu menschlichen, tierischen und pflanzlichen Krankheitserregern sowie zu anderen pflanzenassoziierten Pilzen umfasst und mit öffentlich verfügbaren Genomen verknüpft ist. Wir zeigen, dass sich damit das pathogene Potenzial neuer Pilzarten direkt aus DNA-Sequenzen vorhersagen lässt, entweder mit Sequenzhomologie oder Deep Learning. Wir entwickeln erlernte, numerische Darstellungen der gesammelten Genome und visualisieren die Landschaft der Pilzpathogenität. Schließlich trainieren wir Multi-Klassen-Modelle, die vorhersagen, ob Next-Generation-Sequencing-Reads von neuartigen Pilz-, Bakterien- oder Virusbedrohungen stammen. Die mit unserer Datensammlung trainierten neuronalen Netze ermöglichen eine genaue Erkennung neuartiger Pilzpathogene. Ein kuratierter Satz von über 1400 Genomen mit Wirts- und Pathogenitäts-Metadaten unterstützt das Training von Machine-Learning-Modellen und Sequenzvergleichen, die nicht auf die Erkennung von Pathogenen beschränkt sind.</p>	
19. Schlagwörter Pathogendetektion, Echtzeitdetektion, Deep Learning, Explainable Artificial Intelligence, Bioinformatik	
20. Verlag Oxford University Press	21. Preis (open access)