

Sachbericht Teil I und II

Verbundprojekt: Sprechen Sie Toxin? Die Botschaften der Bakteriophagen – SSTDBB

Teilvorhaben: Grammatik

Förderkennzeichen: 16DKWN136A

Laufzeit: 01.09.2022 bis 31.08.2025

Prof. Dr. Burkhard Rost

Dr. Ivan Koludarov

März 2026

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter den Förderkennzeichen 16DKWN136A und 16DKWN136B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.

Finanziert durch die Europäische Union – NextGenerationEU. Die geäußerten Ansichten und Meinungen sind ausschließlich die des Autors/der Autoren und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Europäischen Kommission wider. Weder die Europäische Union noch die Europäische Kommission können für sie verantwortlich gemacht werden.

Teil I: Kurzbericht

Ursprüngliche Aufgabenstellung und Stand der Forschung

Das Projekt „Sprechen Sie Toxin? Die Botschaften der Bakteriophagen“ (SSTDBB) wurde im September 2022 gestartet mit dem Ziel, Techniken des maschinellen Lernens in den Bereich der Toxinologie zu bringen. Im Teilvorhaben „Grammatik“ lag der Fokus darauf zu verstehen, was Protein-Sprachmodelle (pLMs) in den Primärsequenzen von Proteinen entdecken, und diese Erkenntnisse für die Identifikation und Klassifikation von Toxinen nutzbar zu machen. Das Projekt wurde als interdisziplinäre Zusammenarbeit zwischen dem Lehrstuhl für Bioinformatik (Prof. Dr. Burkhard Rost, Dr. Ivan Koludarov) und dem Labor von PD Dr. Luisa F. Jimenez Soto konzipiert. Zum Zeitpunkt des Projektbeginns fehlten im Bereich der Toxinologie systematische Ansätze zur Nutzung moderner Deep-Learning-Methoden, insbesondere von Protein-Sprachmodellen, die in der allgemeinen Bioinformatik bereits große Erfolge erzielt hatten.

Ablauf des Vorhabens

Das Projekt gliederte sich in drei Phasen: In der Grundlegungsphase (2022–2023) wurden die Evolution der Drei-Finger-Toxin-Proteinfamilie (3FTX) aufgeklärt (publiziert in *Nature Communications*), umfassende ML-Workshops für das Partnerlabor durchgeführt und die Winter School für Toxinologen 2023 organisiert. Die Entwicklungsphase (2024) umfasste die Einstellung studentischer Hilfskräfte, die Kuration eines umfassenden Toxin-Datensatzes sowie die Entwicklung eines Conotoxin-Klassifikators mittels Protein-Einbettungen. Nach der Absage der Winter School 2024 wurde die Strategie auf digitale Bildungsinhalte umgestellt. In der Abschlussphase (2025) wurden mehrere Manuskripte fertiggestellt, internationale Präsentationen durchgeführt (CSIRO Canberra, Workshop in Marokko) und die Entwicklung von Online-Bildungsmaterialien eingeleitet.

Wesentliche Ergebnisse

Die Hauptergebnisse des Projekts umfassen:

- Aufklärung der Evolution der Drei-Finger-Toxine (*Nature Communications*)
- Entwicklung von Toxin/Nicht-Toxin-Prädiktoren unter Verwendung modernster Protein-Einbettungen (Manuskript eingereicht)
- Erstellung und Kuration hochwertiger Toxin-Datensätze (tierisch und bakteriell), öffentlich verfügbar unter DOI: 10.5282/ubm/data.423
- Tiefenanalyse der ToxProt-Datenbank hinsichtlich Vollständigkeit, Datenqualität und KI-Forschungsbereitschaft (Manuskript eingereicht)
- Manuskript zur Evolution von Peptidtoxinen bei Ameisen (eingereicht bei hochrangiger Fachzeitschrift)
- Erstellung eines spezialisierten ML-Modells zur Vorhersage funktioneller Untergruppen von Conotoxinen (Manuskript in Vorbereitung)
- Erfolgreicher Wissenstransfer durch Workshops, Winter School 2023 und internationale Präsentationen (Australien, Marokko)

- Etablierung internationaler Kooperationen (Prof. Holford NYU, European Venom Network, CSIRO)

Hinweis: Die Entwicklung der Online-Bildungsmaterialien (Videoserie zu Protein-Sprachmodellen in der Toxinologie) ist noch nicht abgeschlossen. Die Materialien befinden sich derzeit in der Produktion und werden voraussichtlich im Frühjahr 2026 fertiggestellt und veröffentlicht.

Die Zusammenarbeit mit dem Partnerlabor (PD Dr. Jimenez Soto) war während der gesamten Projektlaufzeit eng und produktiv und umfasste wöchentliche Treffen beider Projektgruppen.

Teil II: Eingehende Darstellung

SSTDBB

Förderkennzeichen: 16DKWN136A

Laufzeit: 01.09.2022 bis 31.08.2025

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter den Förderkennzeichen 16DKWN136A und 16DKWN136B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.

1. Aufgabenstellung

Das Verbundprojekt SSTDBB hatte das übergeordnete Ziel, maschinelles Lernen (ML) und insbesondere Protein-Sprachmodelle (pLMs) systematisch in die Toxinologieforschung einzuführen. Im Teilvorhaben „Grammatik“ lag der Schwerpunkt auf dem Verständnis dessen, was pLMs über Proteinsequenzen – insbesondere über Toxine – lernen, und auf der Nutzung dieser Erkenntnisse für die Entwicklung von Vorhersagewerkzeugen.

Die konkreten Ziele des Teilvorhabens umfassten:

- Untersuchung des Proteinraums toxischer und nicht-toxischer Proteine mittels pLMs
- Erstellung und Kuration hochwertiger Toxin-Datensätze
- Entwicklung von ML-Modellen zur Toxin-Identifikation und -Klassifikation
- Wissenstransfer an das Partnerlabor durch strukturierte Workshop-Serien
- Organisation von Bildungsveranstaltungen (Summer/Winter School) für die Toxinologie-Gemeinschaft
- Verbreitung der Ergebnisse durch Publikationen und internationale Kooperationen

2. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Zum Zeitpunkt des Projektbeginns (September 2022) befand sich das Feld der computergestützten Toxinologie in einem raschen Umbruch. Protein-Sprachmodelle wie ProtT5 und ESM-2, die auf der Transformer-Architektur basieren, hatten bereits beeindruckende Ergebnisse bei allgemeinen Proteinaufgaben wie Strukturvorhersage, Funktionsannotation und Homologieerkennung erzielt. Ihre systematische Anwendung auf Toxine war jedoch kaum erforscht.

Existierende Toxin-Prädiktoren wie ToxinPred und ClanTox basierten überwiegend auf handverlesenen Merkmalen (Aminosäurezusammensetzung, physikochemische Eigenschaften) und unterschieden nicht systematisch zwischen tierischen und bakteriellen Toxinen – ein grundlegender Nachteil, wie unsere Arbeit zeigen sollte.

Die Veröffentlichung von AlphaFold2 (Jumper et al., 2021) hatte die Strukturvorhersage revolutioniert, und Tools wie Foldseek (van Kempen et al., 2023) ermöglichten erstmals schnelle strukturbasierte Suchen im Proteom. Methoden wie LoRA (Hu et al., 2021) eröffneten effiziente

Wege zur Feinabstimmung großer Modelle ohne massiven Rechenaufwand. Diese Technologien integrierten wir sukzessive in unsere Methodik.

3. Planung und Ablauf des Vorhabens

Das Vorhaben gliederte sich in drei Phasen, die im Wesentlichen dem ursprünglichen Zeitplan entsprachen:

Phase I: Grundlegung (2022–2023)

In dieser Phase wurden die Grundlagen für alle nachfolgenden Arbeiten gelegt:

- Durchführung einer umfassenden Workshop-Serie zu maschinellem Lernen und Protein-Sprachmodellen für das Partnerlabor (PD Dr. Jimenez Soto). Behandelt wurden: Random Forest, SVM, XGBoost, Feedforward- und Convolutional Neural Networks, NLP, Einbettungen, pLM-basierte Modellentwicklung.
- Aufklärung der Evolution der Drei-Finger-Toxine (3FTX) durch Kombination phylogenetischer Ansätze, Syntenie-Analyse und ML (AlphaFold2, ProtT5). Identifikation des unmittelbaren nicht-toxischen Vorfahren als nicht-sekretorisches LY6. Publikation in Nature Communications.
- Erstellung eines nicht-redundanten Datensatzes von 50.000 Proteinen, der alle wichtigen toxischen und nicht-toxischen Proteinfamilien abdeckt.
- Organisation und Leitung der Winter School für Toxinologen 2023 – erfolgreiche Einführung in Deep-Learning-Methoden für Teilnehmer verschiedener Karrierestufen.
- Präsentation auf der Konferenz Venom2Drugs 2023 in Australien.
- Erstellung und Kuration des Exotoxin-Datensatzes in Zusammenarbeit mit dem Partnerlabor. Manuskript eingereicht bei Toxins (ID: toxins-2799577). Daten öffentlich verfügbar unter DOI: 10.5282/ubm/data.423.

Phase II: Entwicklung und Erweiterung (2024)

- Einstellung von vier studentischen Hilfskräften (HiWis) zur Unterstützung der Datensatzkuration. Die aufwändige Kuration eines umfassenden tierischen Toxin-Datensatzes aus Publikationen und Datenbanken konnte damit erheblich beschleunigt werden.
- Entwicklung eines Klassifikators zur Unterscheidung von Toxinen und Nicht-Toxinen mittels Protein-Einbettungen.
- Entwicklung eines spezialisierten ML-Modells zur Vorhersage funktioneller Untergruppen von Conotoxinen. Durch die Nutzung von Einbettungen konnten strukturelle Muster identifiziert werden, die klassischen Methoden entgehen.
- Absage der Winter School 2024 aufgrund unzureichender Anmeldungen. Daraufhin grundlegende Überarbeitung der Kommunikationsstrategie und strategische Neuausrichtung auf digitale Bildungsinhalte und Online-Reichweite.
- Initiierung neuer Kooperationen mit Prof. Mandä Holford (NYU) und dem European Venom Network.

Phase III: Abschluss und Verbreitung (2025)

- Einreichung einer Publikation zur Tiefenanalyse von ToxProt (toxinologiefokussierter Teil von SwissProt): Untersuchung der Vollständigkeit, Datenqualität und KI-Forschungsbereitschaft (Leitung: IK, mit HiWi-Unterstützung).
- Fertigstellung des Manuskripts zum Toxin-vs-Nicht-Toxin-Prädiktor, der seit Projektbeginn entwickelt wurde (Leitung: IK, mit HiWi-Unterstützung).
- Vorbereitung und Einreichung eines Manuskripts zur Evolution von Peptidtoxinen bei Ameisen bei einer hochrangigen Fachzeitschrift (Leitung: IK).
- Internationaler Wissenstransfer: Erweiterter Workshop zu Protein-Sprachmodellen bei CSIRO, Canberra (27.–28. Oktober 2025); 2-stündige Sitzung zu „Protein-Sprachmodelle in der Toxinidentifikation“ beim 2. Workshop zur Translationalen Giftmedizin in Marokko (beide Präsentationen von Ivan Koludarov).

Hinweis zur Digitalen Bildungsinitiative: Nach Rücksprache mit der Projektträgerschaft wurde das Konzept der Präsenz-Toxinologie-Schule in eine Serie von Online-Bildungsvideos transformiert. Die Entwicklung dieser Online-Lernmaterialien ist jedoch noch nicht abgeschlossen. Die Materialien befinden sich derzeit in der Produktion und werden voraussichtlich im Frühjahr 2026 fertiggestellt und der Forschungsgemeinschaft frei zugänglich gemacht. Wir sind zuversichtlich, dass diese Materialien durch ihre breitere Reichweite über geografische Grenzen hinweg und ihre dauerhafte Verfügbarkeit einen nachhaltigen Beitrag zur Datenkompetenzsteigerung in der Toxinologie leisten werden.

4. Inhaltliche Ergebnisse

Darstellung der inhaltlichen Ergebnisse anhand der Arbeitspakete. Besonderes Augenmerk liegt darauf, welche neuen Forschungserkenntnisse und -methoden durch die Zusammenarbeit der fachlichen und datenwissenschaftlichen Mitarbeiter:innen gewonnen wurden und wie die Projektergebnisse zur Datenkompetenzsteigerung in der Toxinologie beitragen.

AP 1: Untersuchung des Proteinraums

Die systematische Untersuchung des Proteinraums toxischer und nicht-toxischer Proteine ergab mehrere grundlegende Erkenntnisse:

Drei-Finger-Toxine (3FTX): Durch die Kombination traditioneller phylogenetischer Methoden mit modernen ML-Ansätzen (AlphaFold2 für Strukturvorhersagen, ProtT5 für Sequenz-Einbettungen) gelang es, eine detaillierte Evolutionsgeschichte der 3FTX zu rekonstruieren. Wir identifizierten ein nicht-sekretorisches LY6-Protein, einzigartig für Schuppenkriechtiere, als unmittelbaren Vorfahren. Der Verlust einer Membrananker-Domäne und Veränderungen in der Genexpression ebneten den Weg für die Evolution dieser wichtigen Giftfamilie. Diese Ergebnisse wurden in Nature Communications veröffentlicht und demonstrieren exemplarisch die Stärke des interdisziplinären Ansatzes.

Exotoxin-Analyse: Durch die Erstellung eines umfassenden Datensatzes tierischer und bakterieller Toxine konnten wir zeigen, dass sich prokaryotische und metazoische Toxine in Aminosäurezusammensetzung und Länge so grundlegend unterscheiden, dass ihre

gemeinsame Verwendung in einem einzigen Toxin-Prädiktor zu systematischen Verzerrungen führen kann. Diese fundamentale Erkenntnis hat direkte Auswirkungen auf die Gestaltung zukünftiger Vorhersagemodelle.

ToxProt-Analyse: Eine umfassende Tiefenanalyse der ToxProt-Datenbank (dem toxinologiefokussierten Teil von UniProtKB/Swiss-Prot) ergab wichtige Einblicke in die Vollständigkeit, Datenqualität und KI-Forschungsbereitschaft dieser zentralen Ressource. Die Ergebnisse sind als Publikation eingereicht.

AP 2: Modellentwicklung

Toxin/Nicht-Toxin-Prädiktor: Auf Basis der kuratierten Datensätze und modernster Protein-Einbettungen wurde ein Klassifikator entwickelt, der Toxine von Nicht-Toxinen unterscheidet. Das Manuskript befindet sich in der Einreichung.

Conotoxin-Klassifikator: Ein spezialisiertes ML-Modell zur Vorhersage funktioneller Untergruppen von Conotoxinen wurde entwickelt. Durch die Nutzung von pLM-Einbettungen konnten strukturelle und funktionelle Muster identifiziert werden, die klassischen sequenzbasierten Methoden entgehen. Das Manuskript ist in Vorbereitung.

AP 3: Wissenstransfer und Kapazitätsaufbau

Workshops: Die gesamte geplante Workshop-Serie wurde erfolgreich durchgeführt. Das Partnerlabor von PD Dr. Jimenez Soto verfügt nun über die Kompetenz, ML-Methoden eigenständig in der Toxinologieforschung einzusetzen.

Winter School 2023: Erfolgreich durchgeführt; Toxinologen verschiedener Karrierestufen wurden in Deep-Learning-Methoden eingeführt.

Internationale Präsentationen: Venom2Drugs 2023 (Australien), CSIRO Canberra (Oktober 2025), 2. Workshop zur Translationalen Giftmedizin (Marokko, 2025).

Online-Bildungsmaterialien: Die geplante Transformation der Präsenz-Toxinologie-Schule in eine Serie von Online-Bildungsvideos wurde eingeleitet, ist jedoch noch nicht abgeschlossen. Die Fertigstellung und Veröffentlichung ist für das Frühjahr 2026 geplant.

AP 4: Datensätze

Alle erstellten Datensätze folgen den Prinzipien von Open Science und FAIR-Daten:

- Exotoxin-Datensatz: Kuratierte Sammlung tierischer und bakterieller Toxine, öffentlich verfügbar über das LMU-Repository (DOI: 10.5282/ubm/data.423)
- Nicht-redundanter Datensatz von 50.000 Proteinen für pLM-Untersuchungen
- Conotoxin-Datensatz mit funktionellen Annotationen
- Kuratierter Datensatz für den Toxin/Nicht-Toxin-Prädiktor

5. Wichtigste Positionen des zahlenmäßigen Nachweises

Die Mittel wurden überwiegend für Personalkosten eingesetzt. In den Jahren 2022–2023 gestaltete sich die Suche nach qualifizierten studentischen Hilfskräften schwierig, da das erforderliche Kompetenzprofil (Bioinformatik, ML, Programmierung) stark nachgefragt ist. Die entsprechenden Aufgaben wurden daher teilweise durch zusätzliche Doktorandenstunden abgedeckt. Ab 2024 konnten vier HiWis eingestellt werden, die maßgeblich zur Datensatzkuration und Modellentwicklung beitrugen. Reisekosten fielen für die internationalen Konferenz- und Workshopeteilnahmen an (Australien, Marokko). Sachmittel wurden im Rahmen des bewilligten Budgets eingesetzt.

6. Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Die durchgeführten Arbeiten waren notwendig und angemessen, um das Projektziel zu erreichen. Die aufwändige Datensatzkuration war unerlässlich, da die Qualität von ML-Modellen direkt von der Qualität der Trainingsdaten abhängt. Die Workshop-Serien ermöglichten es dem Partnerlabor, eigenständig ML-Methoden anzuwenden – eine nachhaltige Investition in die Forschungskapazität. Die internationalen Präsentationen waren entscheidend für die Verbreitung der Ergebnisse und den Aufbau von Kooperationen.

7. Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen

Alle Projektergebnisse folgen den Prinzipien von Open Science und FAIR-Daten:

- Datensätze: Öffentlich verfügbar über institutionelle Repositories mit permanenten DOIs. Die Datensätze werden der Forschungsgemeinschaft jahrelang als Referenz dienen.
- Code und Analyse-Pipelines: Verfügbar auf GitHub für die freie Nachnutzung.
- Bildungsmaterialien: Die Online-Videomaterialien werden nach Fertigstellung (Frühjahr 2026) frei online zugänglich sein und als dauerhafte Bildungsressource dienen.
- Publikationen: In Open-Access-Zeitschriften, wo möglich.

Das Projekt hat ein neues Paradigma für die computergestützte Toxinologie etabliert. Die Wirkung wird sich fortsetzen durch: ausgebildete Forscher, die ML-Methoden selbstständig anwenden; internationale Kooperationen (NYU, European Venom Network, CSIRO), die eine kontinuierliche Weiterentwicklung sicherstellen; sowie eine Grundlage für KI-gestützte Wirkstoffforschung aus Toxinen.

8. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Während des Projektzeitraums haben sich im Bereich Protein-ML wesentliche Fortschritte ergeben, die wir sukzessive in unsere Methoden integriert haben:

- LoRA (Low-Rank Adaptation, Hu et al., 2021): Ermöglicht die effiziente Feinabstimmung vortrainierter Modelle ohne massiven Rechenaufwand.

- AlphaFold2 (Jumper et al., 2021) und Foldseek (van Kempen et al., 2023): Ermöglichten strukturbasiertes Protein-Clustering und verbesserte Toxinkandidaten-Auswahl.
- SHAFF-Algorithmus (Bénard et al., 2022): Angewendet für die Interpretierbarkeit unserer Vorhersagemodelle.
- Neueste Protein-Sprachmodelle und Einbettungstechniken: Kontinuierliche Integration neuer pLM-Varianten zur Verbesserung der Vorhersagequalität.

9. Erfolgte oder geplante Veröffentlichungen der Projektergebnisse

Veröffentlichte Publikationen:

- Koludarov I. et al. (2023): Evolution der Drei-Finger-Toxine. Nature Communications. (Published)
- Exotoxin-Datensatz-Analyse: Eingereicht bei Toxins (Manuskript-ID: toxins-2799577). Daten verfügbar unter DOI: 10.5282/ubm/data.423.

Eingereichte/In Vorbereitung befindliche Publikationen:

- Tiefenanalyse von ToxProt: Untersuchung der Vollständigkeit und KI-Bereitschaft (eingereicht)
- Toxin-vs-Nicht-Toxin-Prädiktor mittels Protein-Einbettungen (eingereicht)
- Evolution von Peptidtoxinen bei Ameisen (eingereicht bei hochrangiger Fachzeitschrift)
- Conotoxin-Klassifikator mittels ML (Manuskript in Vorbereitung)

Konferenzbeiträge und Workshops:

- Venom2Drugs 2023, Australien: Vortrag zu Protein-Sprachmodellen in der Toxinologie
- Winter School für Toxinologen 2023: Organisation und Leitung
- CSIRO, Canberra (27.–28. Oktober 2025): Erweiterter Workshop zu pLMs in der Toxinologie
- 2. Workshop zur Translationalen Giftmedizin, Marokko (2025): 2-stündige Sitzung zu Protein-Sprachmodellen in der Toxinidentifikation

10. Literaturverzeichnis

- [1] Hu, E.J. et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." arXiv: 2106.09685.
- [2] Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature 596, 583–589.
- [3] van Kempen, M. et al. (2023). "Fast and accurate protein structure search with Foldseek." Nature Biotechnology.
- [4] Bénard, C. et al. (2022). "SHAFF: Fast and consistent SHAPley eFfect estimates via random Forests." AISTATS 151, 5563–5582.