Explorative Naturstoff-Genomik

Genombasierte Wege zur Identifikation bioaktiver bakterieller Naturstoffe

THAO N. PHAN[1,2], JULIA SPIES[1,2], MILENA BREITEBACH[1,2], ERIC J. N. HELFRICH[1,2]
[1] INSTITUT FÜR MOLEKULARE BIOWISSENSCHAFTEN, UNIVERSITÄT FRANKFURT A. M.
[2] LOEWE-ZENTRUM FÜR TRANSLATIONALE BIODIVERSITÄTSGENOMIK, FRANKFURT A. M.

Bacteria produce natural products to interact with their environment. These structurally diverse small molecules exhibit various bioactivities and have been exploited for the treatment of many diseases. The discovery pace of truly novel natural products using traditional methods has markedly declined. As an alternative, sophisticated bioinformatic tools have been developed that leverage genome sequence information for the targeted discovery of bioactive compounds. To chart microbial biosynthetic dark matter and identify non-canonical natural product biosynthetic pathways that escape unrecognised by current bioinformatic tools, artificial intelligence has recently been introduced into the genome mining process, holding promise to expand natural product chemical and biosynthetic space.

DOI: 10.1007/s12268-024-2192-z

■ Bakterien produzieren eine Vielzahl von Naturstoffen, die auch als Sekundärmetaboliten oder spezialisierte Metaboliten bezeichnet werden. Im Gegensatz zu Primärmetaboliten sind sie nicht essenziell für Wachstum, Entwicklung, Reproduktion oder Überleben eines Organismus; sie spielen jedoch eine wichtige Rolle für die Kontinuität des Produzenten in der Natur [1]. Naturstoffe erfüllen diverse ökologische Funktionen, einschließlich der Kommunikation und chemischen Verteidigung. Sie verfügen über eine enorme strukturelle Diversität, die mit einer Vielzahl unterschiedlicher Bioaktivitäten assoziiert ist. Daher stellen sie eine reiche Quelle an Leitstrukturen für pharmazeutische Wirkstoffe dar. Beispiele hierfür sind das Breitbandantibiotikum Erythromycin (1), das Antimykotikum Amphotericin B, das Immunsuppressivum Rapamycin, oder das Zytostatikum Daunorubicin.

Biosynthetische Prinzipien von Naturstoffen

Die meisten Naturstoffe folgen dem zentralen Dogma der Naturstoff-Biosynthese. Naturstoffe werden durch eine Vielzahl biochemischer Transformationen, die von biosynthetischen Enzymen katalysiert werden, hergestellt. Bei Bakterien sind die biosynthetischen Enzyme, die für die Herstellung eines Naturstoffs verantwortlich sind, in Genen codiert, die in direkter Nachbarschaft zueinander im Genom liegen und biosynthetische Gencluster (BGCs) bilden.

Die Naturstoff-Biosynthese erfolgt nach zwei unterschiedlichen Prinzipien. In beiden Strategien wird zunächst das Naturstoff-Grundgerüst aufgebaut und anschließend von modifizierenden Enzymen funktionalisiert [2]. Zum einen können Naturstoffe durch multifunktionale, modular aufgebaute Mega-Synthasen in einem schrittweisen Prozess, der einem Fließband gleicht, hergestellt werden. Jedes Modul setzt sich aus einzelnen Domänen zusammen und ist für den Einbau und die Modifikation eines Bausteins in das wachsende Naturstoff-Grundgerüst verantwortlich. Ein Beispiel dafür ist die 6-Deoxyerythronolid-B-Synthase (DEBS), die das Polyketid-Grundgerüst von Erythromycin aufbaut (**Abb. 1A**). Bei den modularen Mega-Synthasen korreliert die Enzymarchitektur mit der Naturstoff-Struktur, was die Vorhersage der Naturstoff-Grundstrukturen basierend auf Genom-Sequenzinformationen ermöglicht.

Zum anderen können Naturstoffe durch diskrete, monofunktionale Enzyme aufgebaut werden, die das Naturstoff-Grundgerüst nacheinander erzeugen und modifizieren (**Abb. 1B**). Diesem Biosynthese-Prinzip folgen beispielsweise ribosomal synthetisierte und posttranslational modifizierte Peptide (RiPPs). Ein RiPP-BGC besteht aus Genen, die ein Vorläuferpeptid und modifizierende Enzyme codieren. Das Vorläuferpeptid verfügt über eine Leadersequenz und eine Kernsequenz. Die Leadersequenz rekrutiert die im Gencluster codierten Enzyme, die das Kernpeptid modifizieren. Ist das Kernpeptid vollständig modifiziert, wird die Leadersequenz abgespalten (**Abb. 1B**). RiPPs sind eine heterogene Naturstoffklasse, die sich aus mehr als 40 Familien zusammensetzt. Sie besitzen, anders als Mega-Synthasen, keine konservierten Gene, die in allen RiPP-Familien vorkommen. Dies erschwert die Identifizierung bisher übersehener RiPP-Familien.

Traditionelle Methoden in der Naturstoff-Forschung

Die 1950–1960er-Jahre erwiesen sich als „goldenes Zeitalter" für die Entdeckung neuer Antibiotika. Mehr als die Hälfte der heute auf dem Markt befindlichen Antibiotika wurden in dieser Zeit identifiziert. Traditionell suchen Wissenschaftler:innen bioaktive Metaboliten mithilfe Bioaktivitäts-geleiteter Fraktionierungsstudien. Dabei testen sie Extrakte von Bakterienkulturen auf verschiedene Bioaktivitäten, z. B. auf ihr Potenzial, das Wachstum anderer Bakterien zu hemmen, und reinigen die identifizierten Antibiotika anschließend chromatografisch aus den Extrakten auf. Da die leicht zugängigen Substanzen bereits erforscht sind, führt die Verbindung aus Bioaktivitäts-geleiteter Fraktionierung und der Fokus auf wenige talentierte Naturstoff-Produzenten zur häufigen Wiederentdeckung bereits bekannter Naturstoffe. Daher war es erforderlich, komplementäre Ansätze zur Ent-

deckung neuartiger Verbindungen zu entwickeln.

Genom-Mining

Eine alternative Herangehensweise ist die Analyse bakterieller Genome auf die Präsenz potenzieller Naturstoff-BGCs. Die genomischen Analysen vieler gut erforschter Sekundärmetabolit-Produzenten haben gezeigt, dass selbst in umfassend untersuchten Stämmen das Naturstoff-Biosynthese-Potenzial um ein Vielfaches höher ist als basierend auf Bioaktivitäts-geleiteten Studien angenommen.

Während die Entwicklung von neuartigen Sequenzierungsverfahren es ermöglicht, viele Genome kosteneffizient zu sequenzieren, gestatten Einsichten in die Biosynthese von Naturstoffen die Entwicklung anspruchsvoller bioinformatischer Werkzeuge zur Identifizierung von Naturstoff-BGCs. Das „Genom-Mining" ist eine computergestützte Strategie zur automatischen Identifizierung und Annotation von Naturstoff-Biosynthese-Genclustern in genomischen Daten. Die Grundlagen des Genom-Minings beruhen auf folgenden Beobachtungen:

1. Die biosynthetischen Gene, die für die Herstellung eines Naturstoffs verantwortlich sind, liegen in bakteriellen Genomen geclustert vor. Daher können Naturstoff-BGCs basierend auf einem oder wenigen Genen, die für Schlüsselenzyme codieren, identifiziert werden [3].
2. Trotz der enormen strukturellen Vielfalt sind die biosynthetischen Prinzipien innerhalb einer Naturstoffklasse hoch konserviert. Die Gene, die für Schlüsselenzyme innerhalb einer Naturstoffklasse codieren, können basierend auf Sequenzhomologien identifiziert werden [4].
3. Basierend auf der Etablierung scheinbar universeller biosynthetischer Prinzipien ist es in vielen Fällen möglich, die Grundstruktur von Naturstoffen basierend auf der Genomsequenz vorherzusagen.

Regelbasierte Genom-Mining-Algorithmen

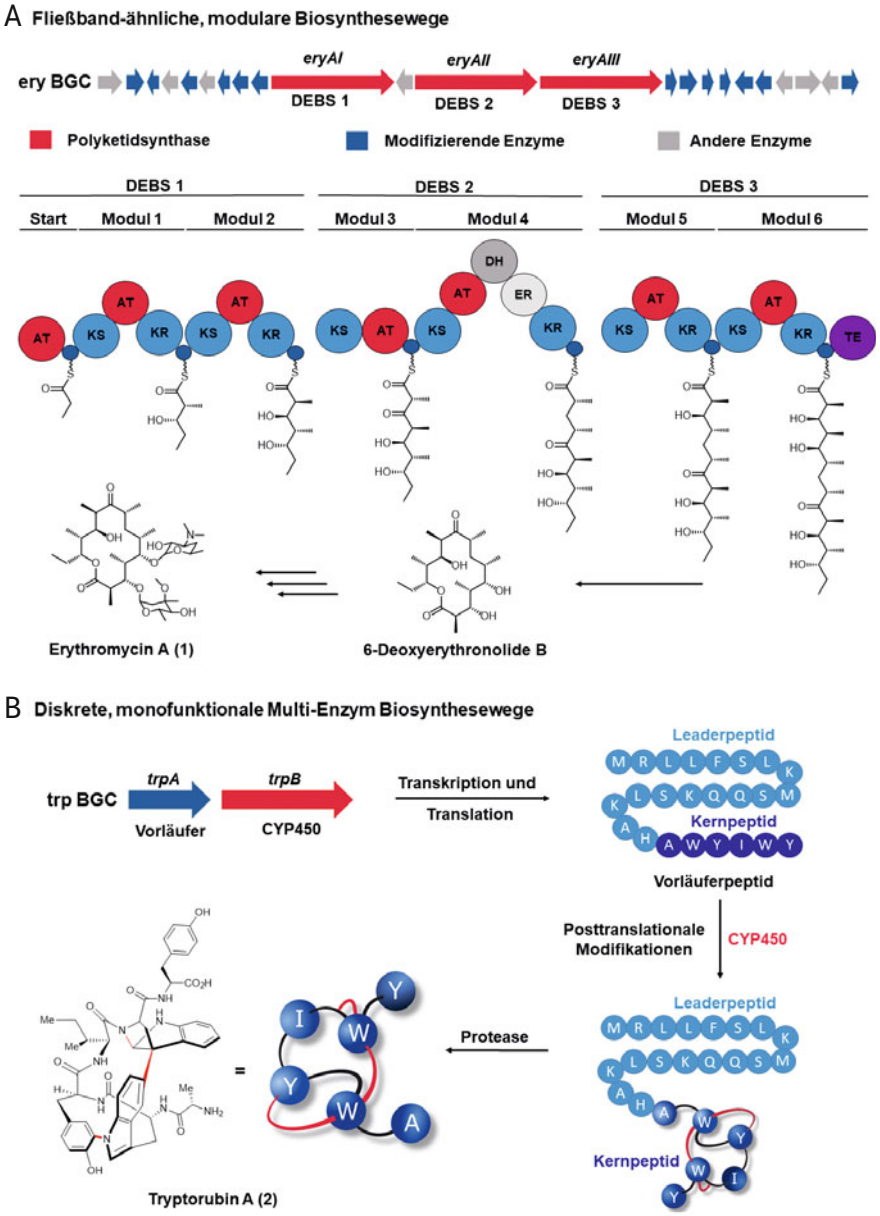
In der Regel verwenden klassische Genom-Mining-Algorithmen Profil-Hidden-Markov-Modelle (pHMMS), um Schlüsselenzym-codierende Gene zu identifizieren: Dazu werden zunächst bekannte Nukleotid- oder Aminosäuresequenzen von Genen, Enzymfamilien oder Domänen miteinander verglichen. Anschließend wird ein Profil erstellt, das die Wahrscheinlichkeit beschreibt, mit der bestimmte Basen oder Aminosäuren an jeder Position in den Sequenzen auftreten. Basierend auf diesem Profil wird ein pHHM erstellt, mithilfe dessen selbst Vertreter dieser Gene, Enzymfamilien oder Domänen mit geringer Sequenzidentität in Sequenz-Datenbanken identifiziert werden können. Auf diese Art werden zunächst Schlüsselenzym-codierende Gene identifiziert und anschließend die verbleibenden Gene des BGCs annotiert [5].

Vorteile und Nachteile des klassischen Genom-Minings

Regelbasierte Genom-Mining-Algorithmen können kanonische Naturstoff-Biosynthesewege, die scheinbar universellen biosynthetischen Prinzipien folgen, zuverlässig in großen Datensätzen identifizieren. Dies trifft besonders auf bakterielle Fließband-artige Mega-Synthasen zu. Zum einen lassen sich so bereits charakterisierte BGCs, deren assoziierte Metaboliten bekannt sind (Bekannte-Bekannte; **Abb. 2**), schnell identifizieren und aussortieren, um die Wiederentdeckung bekannter Metaboliten zu vermeiden. Für den weitaus größten Anteil der Naturstoff-BGCs, die von klassische Genom-Mining-Algorithmen identifiziert werden, sind die assoziierten Metaboliten noch nicht bekannt (Bekannte-Unbekannte; **Abb. 2**). In diesem Fall können neue Naturstoffe durch die Kultivierung des natürlichen Produzenten oder durch den Transfer und die Expression des BGCs in einen geeigneten heterologen Wirt charakterisiert werden.

Allerdings stößt das regelbasierte Genom-Mining bei der Erkennung neuartiger Biosynthesewege, die von den scheinbar universellen biosynthetischen Prinzipien abweichen, an seine Grenzen. Dies ist besonders bei bisher unerforschten Familien ribosomal synthetisierter und posttranslational modifizierter Peptide der Fall. Die Biosynthese von Tryptorubin A (2) (TrpA), das proliferative Eigenschaften aufweist, ist ein Beispiel für einen Naturstoff, dessen assoziiertes BGC von Genom-Mining-Algorithmen nicht erkannt wurde [6].

Das Hexapeptid TrpA wurde aus einem *Streptomyces*-Stamm isoliert und verfügt über eine ungewöhnlich komplexe dreidimensio-



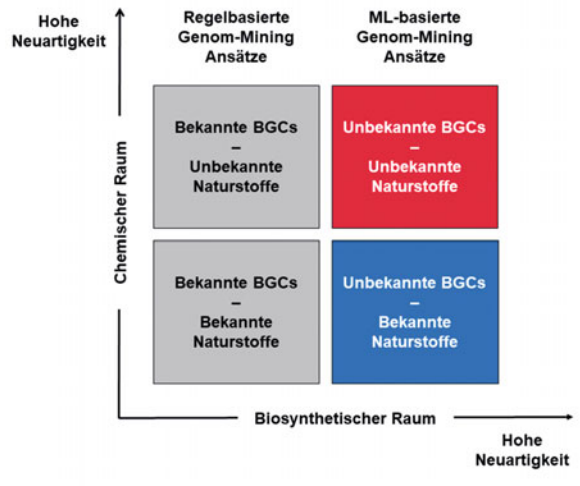
▲ **Abb. 1:** Beispiele für die beiden Naturstoff-Biosynthese-Prinzipien. **A**, Erythromycin-BGC und Modell für die Biosynthese von Erythromycin**.** Acyltransferasen (AT) wählen biosynthetische Bausteine aus und transferieren sie auf Acyl-Carrier-Proteine (kleine dunkelblaue Kreise). Die Ketosynthase (KS) knüpft die Bindung zwischen einem neuen Biosynthese-Baustein und dem wachsenden Polyketid. Dies resultiert in einer beta-Keto-Gruppe, die durch die Ketoreduktase (KR) zu einem Alkohol reduziert werden kann. Die Dehydratase (DH) erzeugt durch die Eliminierung von Wasser eine Doppelbindung, die durch eine Enoylreduktase (ER) komplett gesättigt werden kann. KR, DH und ER sind fakultative enzymatische Domänen, die nur in manchen Modulen vorkommen und dadurch für strukturelle Diversität sorgen. Die Thioesterase (TE) spaltet das vollständig assemblierte Polyketid von der Polyketidsynthase ab. Das resultierende Erythromycin Aglykon wird anschließend weiter modifiziert. **B**, Tryptorubin A-BGC und Modell für die Biosynthese von Tryptorubin A (TrpA). Das TrpA-Vorläuferpeptid setzt sich aus einem Leaderpeptid (hellblau) und einem Kernpeptid (dunkelblau) zusammen. Die Zytochrom-P450-Monooxygenase (CYP450) katalysiert die Knüpfung dreier Bindungen (rot markiert): eine C-C-Bindung zwischen den Seitenketten zweier Tryptophane (W-W), zwei C-N-Bindungen zwischen den Seitenketten von Tryptophan und Tyrosin (W-Y) sowie zwischen der Seitenkette von Tryptophan und dem Peptid-Grundgerüst. Diese Modifikationen resultieren in einer starren, hochkomplexen dreidimensionalen Struktur. Anschließend wird das modifizierte Kernpeptid von einer Protease vom Leaderpeptid abgespalten.

nale Struktur [7]. Da das *trp*-BGC von Genom-Mining-Algorithmen übersehen wird (Unbekannte-Bekannte), identifizierten wir das BGC durch manuelle retrobiosynthetische Analysen. Dabei zeigte sich, dass TrpA ein ribosomal synthetisiertes und posttranslational modifiziertes Peptid ist [6]. Das *trp*-BGC besteht lediglich aus zwei Genen, *trpA* und *trpB*, die ein Vorläuferpeptid und eine Zytochrom-P450-Monooxygenase (CYP450) codieren (**Abb. 1B**). Bemerkenswert ist, dass das CYP450 ausreichend ist, um die komplexe molekulare Struktur von Tryptorubin A zu erzeugen [6]. TrpA ist der erste Vertreter einer neuen RiPP-Familie, die wir Atropopeptide genannt haben. Mittlerweile umfassen die Atropopeptide weitere Familienmitglieder, die zum Teil durch einen auf maschinellem Lernen fußenden Genom-Mining-Algorithmus identifiziert wurden [6, 8].

Maschinelles Lernen zur Entdeckung neuer Naturstoffe

Maschinelles Lernen, ein Bereich der künstlichen Intelligenz, verwendet Algorithmen und statistische Modelle, um aus Daten zu lernen. Gegenüber dem regelbasierten Genom-Mining bietet dieser Ansatz den Vorteil, dass Naturstoffe, deren Biosynthese nicht den scheinbar universellen biosynthetischen Prinzipien folgt, identifiziert werden können. Beim überwachten Lernen werden auf maschinellem Lernen basierende Algorithmen mit gekennzeichneten Datensätzen zur Klassifikation und Vorhersage trainiert, um anschließend ungesehene Daten zu sortieren. Die ungewöhnliche, aber hochkomplexe Struktur von TrpA motivierte uns zur Entwicklung eines auf maschinellem Lernen basierenden Algorithmus, um den chemischen Raum der Atropopeptid-Familie zu erweitern und vollständig zu charakterisieren. Auf diese Weise können wir auch weit entferne Verwandte identifizieren, die die traditionellen Methoden wie BLAST nicht als solche identifizieren würden.

Um alle in Genom-Datenbanken verfügbaren Atropopetid-BGCs zu identifizieren, entwickelten wir einen Algorithmus, der Atropopeptid-modifizierende CYP450s anhand ihrer physisch-chemischen Eigenschaften (z. B. Substratspezifität) – nicht anhand von Sequenzhomologien wie traditionelle Verfahren – mit hoher Präzision bestimmt. Im zweiten Schritt werden die Vorläufergene und Gene, die weitere modifizierende Enzyme codieren, in der direkten Umgebung der CYP450-Gene annotiert. Die CYP450s, die charakteristisch für Atropopeptide sind, identifizierten wir also zunächst aus dem Pool aller CYP450s und setzten sie dann als Köder ein, um potenzielle Atropopeptid-BGCs zu identifizieren. Zur Validierung des Algorithmus wurden ausgewählte BGCs charakterisiert, was die Entdeckung von Atropopeptiden mit neuartigen Modifikationen, verschiedenen Peptidlängen und variabler struktureller Komplexität ermöglichte [8].



▲ **Abb. 2:** Die „Rumsfeld-Matrix" der vier Kategorien der Biosynthese-Gencluster/Naturstoff-Paare. Die Bekannten-Bekannten (bereits charakterisierte BGCs, deren assoziierte Metaboliten bekannt sind) und die Bekannten-Unbekannten (BGCs, deren assoziierten Metaboliten noch nicht bekannt sind) lassen sich durch klassische Genom-Mining-Algorithmen zuverlässig identifizieren. Bei den Unbekannten-Bekannten sind die Naturstoffe bekannt, aber die assoziierten BGCs werden von traditionellen Genom-Mining-Algorithmen übersehen. Die wahren, verborgenen biosynthetischen Schätze stellen die Unbekannten-Unbekannten dar, bei denen weder die BGCs noch die assoziierten Metaboliten bekannt sind. Diese biosynthetische dunkle Materie (Unbekannte-Unbekannte) kann man vermutlich mittels maschinellem Lernen identifizieren.

Das verwendete Prinzip kann in der Zukunft flexibel auf andere Naturstoffklassen angewendet werden, da die neue Genom-Mining-Strategie nicht auf die Substratspezifität von Atropopeptid-modifizierenden CYP450s beschränkt ist, sondern auf alle Enzyme, die Naturstoff-Grundgerüste modifizieren, angewendet werden kann. Durch die Verwendung dieses simplen Prinzips können in Zukunft ungewöhnlichen BGCs, die vermutlich für die Biosynthese neuartiger Naturstoffe verantwortlich sind (Unbekannte-Unbekannte; **Abb. 2**), zielgerichtet identifiziert werden.

Biosynthetische dunkle Materie erschließen

Der Einsatz bioinformatischer Werkzeuge hat die Naturstoff-Forschung revolutioniert. Da auf maschinellem Lernen basierende Algorithmen in der Lage sind, Naturstoff-BGCs zu identifizieren, die von hochmodernen bioinformatischen Algorithmen übersehen werden, birgt ihre Verwendung das Potenzial, den chemischen und biosynthetischen Raum von Naturstoffen zu erweitern und so nicht nur neue bioaktive Substanzen zu identifizieren, sondern auch ungewöhnliche Enzyme zu charakterisieren. Der hier vorgestellte, auf maschinellem Lernen beruhende Ansatz, stellt jedoch nur eine von vielen möglichen Lösungen dar, die biosynthetische dunkle Materie (Unbekannte-Unbekannte) zu erschließen. Das Potenzial der künstlichen Intelligenz (KI) ist in dieser Hinsicht noch bei weitem nicht ausgeschöpft. Fortschritte in der KI und die Etablierung geeigneter Datenbanken, die man für das Trainieren von KI-basierten Algorithmen verwenden kann, unterstützen die Entwicklung neuartiger Genom-Mining-Werkzeuge und tragen somit zur Entdeckung verborgener biosynthetischer Schätze bei.

Danksagung

Ein besonderer Dank geht an unsere Arbeitsgruppe für den konstruktiven Austausch von Fachwissen aus verschiedenen Disziplinen. Darüber hinaus möchten wir dem LOEWE-Zentrum für Translationale Biodiversitätsgenomik und dem Emmy-Noether-Programm der Deutschen Forschungsgemeinschaft für die Bereitstellung der finanziellen Mittel für unsere Forschung danken. ■

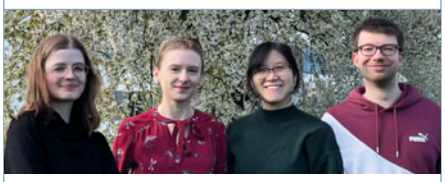
Literatur

[1] Medema MH, Rond T de, Moore BS (2021) Mining genomes to illuminate the specialized chemistry of life. Nat Rev Genet 22: 553–571
[2] Scott TA, Piel J (2019) The hidden enzymology of bacterial natural product biosynthesis. Nat Rev Chem 3: 404–425
[3] Weber T (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. Int J Med Microbiol 304: 230–235
[4] Ziemert N, Alanjary M, Weber T (2016) The evolution of genome mining in microbes - a review. Nat Prod Rep 33: 988–1005
[5] Biermann F, Helfrich EJN (2021) Hidden treasures: Microbial natural product biosynthesis off the beaten path. mSystems: e0084621
[6] Nanudorn P, Thiengmag S, Biermann F et al. (2022) Atropopeptides are a novel family of ribosomally synthesized and posttranslationally modified peptides with a complex molecular shape. Angew Chem Int Ed 61: e202208361
[7] Reisberg SH, Gao Y, Walker A et al. (2020) Total synthesis reveals atypical atropisomerism in a small-molecule natural product, tryptorubin A. Science 367: 458–463
[8] Biermann F, Tan B, Breitenbach M et al. (2023) Machine learning-based exploration, expansion and definition of the atropopeptide family of ribosomally synthesized and post-translationally modified peptides. bioRxiv: 2023.11.03.565440

Funding note: Open Access funding enabled and organized by Projekt DEAL.

Korrespondenzadresse:
Prof. Dr. Eric Helfrich
Institut für Molekulare Biowissenschaften
Goethe Universität Frankfurt
Max-von-Laue-Straße 9
D-60438 Frankfurt a. M.
eric.helfrich@bio.uni-frankfurt.de
www.helfrichlab.com

AUTORINNEN UND AUTOREN



Milena Breitenbach, Julia Spies, Thao Phan und Eric Helfrich (v. l. n. r.)

Thao Phan
2013–2019 Studium der Biowissenschaften und Molekularen Biotechnologie an der Universität Frankfurt a. M. **2018–2019** Forschungsaufenthalt an der Katholieke Universiteit Leuven, Belgien. **2020–2021** wissenschaftliche Mitarbeiterin am Leibniz-Institut für Neue Materialien, Saarbrücken und an der Universität Frankfurt a. M. Seit **2022** Doktorandin am Institut für Molekulare Biowissenschaften der Universität Frankfurt a. M.

Julia Spies
2017–2021 Bachelorstudium der Biowissenschaften und **2020–2022** Masterstudium der Molekulare Biowissenschaften an der Universität Frankfurt a. M. Seit **2023** Promotionsstudentin am Institut für Molekulare Biowissenschaften der Universität Frankfurt a. M.

Milena Breitenbach
2016–2020 Studium der Biowissenschaften an der Hochschule Fresenius. **2020–2021** Studium der Bioanalytik an der Hochschule Fresenius. **2021–2022** beschäftigt bei A&M Stabtest Mainz. Seit **2022** wissenschaftliche Mitarbeiterin an der Universität Frankfurt a. M.

Eric Helfrich
2007–2010 Bachelorstudium in Molekularer Biomedizin, Universität Bonn. **2010–2013** Masterstudium in Chemischer Biologie, Universität Jena. **2013–2017** Promotion an der Eidgenössischen Technischen Hochschule (ETH) in Zürich, Schweiz. **2018–2020** PostDoc an der Harvard Medical School, Boston, USA. Seit **2020** W1-tt-W2 Professor für Naturstoffgenomik an der Universität Frankfurt und dem LOEWE Zentrum für translationale Biodiversitätsgenomik in Frankfurt a. M.