

Beam-Search zum automatisierten Entwurf und Scoring neuer ROR-Liganden mithilfe maschineller Intelligenz

Journal Article

Author(s):

Moret, Michael; Helmstädter, Moritz; [Grisoni, Francesca](#) ; [Schneider, Gisbert](#) ; Merk, Daniel

Publication date:

2021-08-23

Permanent link:

<https://doi.org/10.3929/ethz-b-000501729>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Angewandte Chemie. International Edition 133(35), <https://doi.org/10.1002/ange.202104405>

Funding acknowledgement:

182176 - De novo molecular design by deep learning (SNF)

De-novo-Design

Beam-Search zum automatisierten Entwurf und Scoring neuer ROR-Liganden mithilfe maschineller Intelligenz**

Michael Moret[†], Moritz Helmstädter[†], Francesca Grisoni, Gisbert Schneider* und Daniel Merk*

Abstract: Chemische Sprachmodelle ermöglichen ein De-novo-Wirkstoff-Design ohne explizite chemische Konstruktionsregeln. Während solche Modelle angewendet wurden, um neuartige Verbindungen mit angestrebter biologischer Aktivität zu generieren, bleibt die tatsächliche Priorisierung und Auswahl der vielversprechendsten Moleküldwürfe („Designs“) eine Herausforderung. Wir haben hier die von chemischen Sprachmodellen gelernten Wahrscheinlichkeiten mithilfe des Beam-Search-Algorithmus als Modell-intrinsische Technik für das Moleküldesign und die Bewertung der Designs („Scoring“) genutzt. Die prospektive Anwendung dieser Methode führte zu neuartigen inversen Agonisten der Retinoid-related-Orphan-Rezeptoren (RORs). Jedes Design war in drei Reaktionsschritten synthetisierbar und zeigte eine niedrig-mikromolare bis nanomolare Potenz gegenüber ROR γ . Als Modell-intrinsische Technik eliminiert das Beam-Search-Sampling die strikte Notwendigkeit externer Molekül-Scoring-Funktionen und erweitert damit die Anwendbarkeit generativer künstlicher Intelligenz in der datengetriebenen Wirkstoffforschung.

Einleitung

Generatives Deep Learning,^[1,2] eine Klasse maschineller Lernmodelle, die in der Lage sind, neue Daten zu generieren, kann für das computergestützte De-novo-Design pharmakologisch aktiver Verbindungen eingesetzt werden.^[3–5] Deep-Learning-basierte Algorithmen für das Moleküldesign können spezifische chemische Merkmale aus „rohen“ Moleküldarstellungen, wie z. B. molekularen Graphen und dem Simplified Molecular Input Line Entry System (SMILES, Abbildung 1 a)^[11] extrahieren,^[6–10] was ihnen potenziell den Zugang zu unerforschten Regionen des chemischen Raums er-

Zitierweise: *Angew. Chem. Int. Ed.* **2021**, *60*, 19477–19482
Internationale Ausgabe: doi.org/10.1002/anie.202104405
Deutsche Ausgabe: doi.org/10.1002/ange.202104405

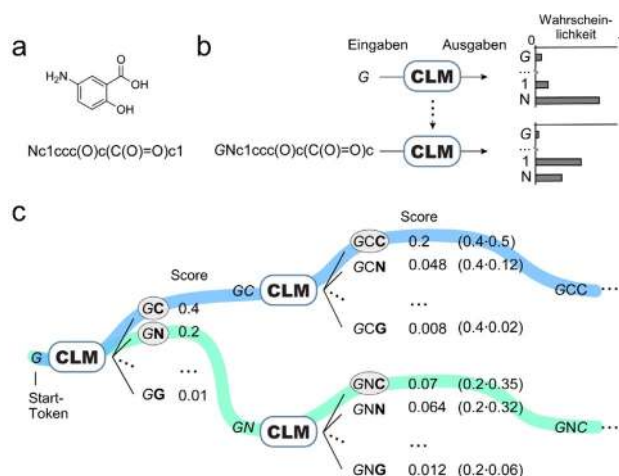


Abbildung 1. Moleküldesign durch chemische Sprachmodelle (CLM) und Beam-Search-Sampling. a) Kekulé-Struktur eines Beispielmoleküls mit seinem korrespondierenden SMILES-String. b) CLM-Training. Das CLM lernt die Wahrscheinlichkeit jedes SMILES-Zeichens („Token“) basierend auf den vorherigen Token im SMILES-String vorherzusagen. c) Beam-Search mit $k=2$: Der Algorithmus behält die zwei wahrscheinlichsten SMILES-Strings im Blick (farblich hervorgehoben). In diesem Beispiel erfolgt die Generierung des SMILES-Strings von links nach rechts.

möglich.^[12] Frühere Studien haben gezeigt, dass chemische Sprachmodelle (Chemical Language Models, CLMs),^[13,14] insbesondere auf SMILES-Strings trainierte generative Deep-Learning-Modelle, neuartige Moleküle mit experimentell validierter biologischer Aktivität generieren können.^[9,15,16] CLMs haben dabei die Fähigkeit bewiesen, fo-

[*] M. Moret,^[†] Prof. Dr. F. Grisoni, Prof. Dr. G. Schneider
ETH Zurich, Department of Chemistry and Applied Biosciences
Vladimir-Prelog-Weg 4, 8093 Zurich (Schweiz)
E-Mail: gisbert@ethz.ch

M. Helmstädter,^[†] Prof. Dr. D. Merk
Goethe University Frankfurt
Institute of Pharmaceutical Chemistry
Max-von-Laue-Straße 9, 60438 Frankfurt (Deutschland)
E-Mail: merk@pharmchem.uni-frankfurt.de

Prof. Dr. F. Grisoni
Eindhoven University of Technology
Institute for Complex Molecular Systems
Department of Biomedical Engineering
Groene Loper 7, 5612AZ Eindhoven (Niederlande)

Prof. Dr. G. Schneider
ETH Singapore SEC Ltd
1 CREATE Way, #06-01 CREATE Tower
Singapore 138602 (Singapur)

Prof. Dr. D. Merk
LMU München, Department of Pharmacy
Butenandtstraße 7, 81377 München (Deutschland)

[†] Diese Autoren haben zu gleichen Teilen zu der Arbeit beigetragen.

[**] Eine frühere Version dieses Manuskripts ist auf einem Preprint-Server hinterlegt worden (<http://doi.org/10.26434/chemrxiv.14153408.v1>).

Hintergrundinformationen und Identifikationsnummern (ORCIDs) der Autoren sind unter:
<https://doi.org/10.1002/ange.202104405> zu finden.

© 2021 Die Autoren. Angewandte Chemie veröffentlicht von Wiley-VCH GmbH. Dieser Open Access Beitrag steht unter den Bedingungen der Creative Commons Attribution License, die jede Nutzung des Beitrages in allen Medien gestattet, sofern der ursprüngliche Beitrag ordnungsgemäß zitiert wird.

kussierte chemische Merkmale aus kleinen Sammlungen von Template-Molekülen mittels Transferlernen zu erlernen.^[15,17,18] Die Methode des Transferlernens ermöglicht die Wiederverwendung von zuvor gelerntem Wissen in einer neuen Aufgabe, für die nur wenige Daten verfügbar sind, und wird in zwei Schritten durchgeführt. Im ersten Schritt wird ein Modell mit einer großen Menge von Daten trainiert, die sich auf die zu erfüllende Aufgabe beziehen („Pre-Training“). Im Falle von CLMs wird dies üblicherweise mit großen Molekülsammlungen in der Größenordnung von 200 000 bis 1 000 000 Molekülen erreicht.^[9,16,17] Das Pre-Training ermöglicht es dem generativen Modell, a) die SMILES-Syntax (d. h. wie alphanumerische Zeichen zusammengesetzt werden sollten, um Strings zu generieren, die validen Molekülen entsprechen, Abbildung 1) und b) die Eigenschaften des Pre-Trainingsdatensatzes, wie z. B. die physikochemischen Eigenschaften und die chemische Synthetisierbarkeit der Moleküle im Datensatz zu erfassen. Im zweiten Schritt wird das vortrainierte CLM mit einer kleineren Anzahl aufgabenspezifischer Moleküle weiter trainiert („Fine-Tuning“).^[13,19,20] Während dieses Transferlernprozesses wird das CLM auf den chemischen Raum von Interesse ausgerichtet, also auf Moleküle mit den angestrebten biologischen und physikochemischen Eigenschaften. Diese Fähigkeit, von wenigen Daten zu lernen („few-shot learning“^[21,22]), macht CLMs besonders für die Anwendung auf biologische Zielstrukturen wertvoll, für die nur wenige Liganden bekannt sind. Das vollständig trainierte CLM kann dann verwendet werden, um neue Moleküle in Form von SMILES-Strings zu entwerfen. Diese Datengenerierung erfolgt durch die schrittweise Vorhersage von jeweils einem Zeichen („Token“) eines SMILES-Strings basierend auf allen vorherigen Token. Erwähnenswert ist dabei, dass dieser Prozess keine vordefinierten Regeln für das Moleküldesign erfordert, da CLMs nur von den SMILES-Strings lernen, die für das Training verwendet werden.

Bisherige Anwendungen von CLMs im De-novo-Design haben das sog. Temperatur-Sampling genutzt, um große virtuelle Molekülbibliotheken zu generieren.^[9,13,15] Temperatur-Sampling erlaubt es, neue SMILES-Strings zu erstellen, indem Token zum (wachsenden) String entsprechend den vom CLM gelernten Wahrscheinlichkeiten hinzugefügt werden, wobei wahrscheinlichere Token an einer bestimmten Position häufiger gewählt werden (Abbildung 1 b). Allerdings sind die so generierten SMILES-Strings ggf. nicht immer chemisch sinnvoll (invalide Strings), oder sie stimmen aufgrund der Zufallskomponente des Temperatur-Samplings nicht mit der Merkmalsverteilung der Trainingsdaten überein. Daher bedarf es üblicherweise zusätzlicher Methoden, um die vielversprechendsten Designs aus den virtuellen Molekülbibliotheken auszuwählen. Dies geschieht z. B. basierend auf der Ähnlichkeit zu bekannten biologisch aktiven Molekülen, mittels externer Aktivitätsvorhersage oder durch sog. Belohnungsfunktionen.^[9,13,15,23]

Als Modell-intrinsische Alternative zum Temperatur-Sampling nutzen wir hier den Beam-Search-Algorithmus, der dem CLM die gleichzeitige Erstellung und Priorisierung der molekularen Designs in automatisierter Weise ohne zusätzliche Selektionsmethoden ermöglicht.^[24,25] In einer prospektiven Anwendung zur Entwicklung neuer Retinoid-related-

Orphan-Rezeptor (ROR)-Liganden^[26] wurde das Beam-Search-Sampling und -Scoring erfolgreich validiert.

Die RORs wurden als molekulare Zielstrukturen ausgewählt, da sie attraktive, aber nicht umfassend untersuchte potenzielle Wirkstofftargets darstellen. Sie bilden eine Familie Ligand-aktivierter Transkriptionsfaktoren, die hauptsächlich als Monomere agieren und unter anderem an der zirkadianen Kontrolle der Energiehomöostase^[27,28] und der Regulation des Immunsystems^[29,30] beteiligt sind. RORs besitzen vielversprechendes pharmakologisches Potenzial in verschiedenen Indikationen, insbesondere bei Autoimmunerkrankungen.^[29,30] Bis heute hat jedoch noch kein ROR-Ligand eine Arzneistoffzulassung erreicht, was zum Teil auf substanzspezifische Schwachpunkte wie schlechte Wasserlöslichkeit, mangelnde Selektivität und klinische Sicherheitsbedenken zurückzuführen ist.^[29,31,32]

Ergebnisse und Diskussion

Chemisches Sprachmodell und Beam-Search-Sampling zum De-novo-Design

Als mögliche Alternative zum Temperatur-Sampling in Kombination mit einer externen Priorisierungsmethode haben wir den Beam-Search-Algorithmus^[33] zur Erzeugung von Molekülen aus einem CLM untersucht. Basierend auf den Wahrscheinlichkeiten, die ein CLM erlernt, kann theoretisch eine große Anzahl von SMILES-Strings erzeugt werden, es ist rechnerisch aber nicht machbar, alle möglichen Strings zu generieren. Es kann jedoch die Hypothese aufgestellt werden, dass die Wahrscheinlichkeit für die Erzeugung eines bestimmten SMILES-Strings mit der Qualität des korrespondierenden Moleküls im Hinblick auf das Design-Ziel korreliert, wie es im Fine-Tuning-Set repräsentiert ist (z. B. angestrebte biologische Aktivität, physikochemische Eigenschaften). Mithilfe einer heuristischen Methode wie der Beam-Search können die wahrscheinlichsten Strings, die ein CLM generieren kann, gefunden werden.

Während der Molekülgenerierung durch den Beam-Search-Algorithmus („Beam-Search-Sampling“) fügt der Algorithmus schrittweise Token zu einem SMILES-String hinzu, während er die k wahrscheinlichsten SMILES-Strings behält. Um ein neues Token hinzuzufügen, berechnet der Algorithmus die bedingte Wahrscheinlichkeit jedes möglichen Token basierend auf den Token im bestehenden String und definiert die k wahrscheinlichsten Token, um den String zu erweitern (Abbildung 1 c). Die Menge der k wahrscheinlichsten gewählten Optionen basiert auf einer Bewertungsfunktion („Beam-Search-Score“), die als Produkt der Wahrscheinlichkeiten der einzelnen Token berechnet wird (Abbildung 1 c). Dieser Prozess wird so lange wiederholt, bis der SMILES-String vollständig ist (d. h. das „Ende-des-Strings“-Token hinzugefügt wird) oder eine vordefinierte maximale String-Länge erreicht ist. Auf diese Weise kann der Beam-Search-Algorithmus verwendet werden, um basierend auf 1) dem zugrundeliegenden Modell und 2) dem Beam-Search-Score hochwahrscheinliche SMILES-Strings zu erzeugen. Der Beam-Search-Score erlaubt es dabei, diese De-novo-

Designs nach der Wahrscheinlichkeit ihrer SMILES-Strings zu ordnen.

Zur Untersuchung und Anwendung des Beam-Search-Sampling haben wir ein kürzlich veröffentlichtes CLM genutzt, das auf einem rekurrenten neuronalen Netzwerk mit sog. long short-term memory cells (LSTM) basiert, welche für Sequenzmodellierung geeignet sind.^[34] Das CLM wurde mit den SMILES-Strings von 365 063 Molekülen aus ChEMBL^[35] trainiert, iterativ das nächste Token jedes SMILES-Strings unter Berücksichtigung der vorhergehenden Token vorherzusagen (Abbildung 1b). Die Trainingsprozedur wurde über zehn Epochen durchgeführt, was bedeutet, dass jedes für das Training verwendete Molekül vom CLM zehnmal gesehen wurde. Durch Transfer-Lernen („Fine-Tuning“) mit Sets bekannter ROR-Liganden (Abbildung S1, Tabelle S1) wurde dann in das vortrainierte CLM ein Bias in Richtung des Design-Ziels, nämlich die Entwicklung neuer Moleküle mit biologischer Aktivität an den RORs, eingeführt. Die Open-Source-Codes für das CLM und den Beam-Search-Algorithmus sowie die in dieser Studie verwendeten Daten sind unter https://github.com/ETHmodlab/molecular_design_with_beam_search verfügbar.

Anwendung des Beam-Search-Samplings zum Design inverser ROR γ -Agonisten

In einer prospektiven Analyse wurde der Beam-Search-Algorithmus auf das Design Naturstoff-inspirierter ROR γ -Liganden angewendet. Als traditionelle Inspirationsquelle für die Arzneimittelentwicklung^[36,37] kann das Lernen von Naturstoffen gegenüber rein synthetischen Molekülen mehrere Vorteile bieten. Naturstoffe weisen insgesamt mehr strukturelle Vielfalt, größere Dreidimensionalität und häufig eine höhere Selektivität auf.^[38,39] Daher strebten wir an, De-novo-

Designs zu erhalten, die drei Eigenschaften besitzen: 1) eine von Naturstoffen inspirierte chemische Struktur, 2) chemische Synthetisierbarkeit und 3) biologische Aktivität an ROR γ . Um alle drei Ziele während des Transfer-Lernens zu erfüllen, wurde das zuvor mit biologisch aktiven Molekülen aus ChEMBL^[17] vortrainierte CLM mithilfe eines synthetischen ROR γ -Liganden und vier in der Literatur^[30] beschriebenen ROR γ -modulierenden Naturstoffen verfeinert (Abbildung S1). Von diesem CLM wurde ab der fünften Epoche des Fine-Tunings mit dem Beam-Search-Sampling begonnen, um sicherzustellen, dass das CLM die molekularen Merkmale des kleinen Fine-Tuning-Datensatzes ausreichend erfasst hatte.

Alle gültigen SMILES-Strings, die das CLM zwischen den Epochen 5 und 16 (letzte Epoche des Fine-Tunings) generierte, wurden anhand des Beam-Search-Scores eingestuft. Die fünf Designs mit dem höchsten Beam-Search-Score (Abbildung 2a) wurden jedoch von Medizinalchemikern als synthetisch unzugänglich eingestuft, und auch die Vorhersagen eines maschinellen Lernalgorithmus für retrosynthetische Analysen (IBM RXN)^[40] konnten für keines dieser Moleküle eine Syntheseroute finden. Während das CLM also die Ähnlichkeit zu Naturstoffen erfasste, erfüllte es nicht das generische Designkriterium der Synthetisierbarkeit. Diese Ergebnisse deuten einen Nutzen des Beam-Search-Samplings an, die wahrscheinlichsten Designs eines CLMs zu offenbaren und den Erfolg des Fine-Tunings hinsichtlich der Design-Ziele zu bewerten.

Um diese Ergebnisse zu verbessern, wurde ein zweites Experiment mit einer zweistufigen Fine-Tuning-Strategie durchgeführt, bei dem das vortrainierte CLM zunächst für 20 Epochen mit 255 synthetischen ROR γ -Liganden aus dem US-Patent-Subset der Protein Data Bank^[41] (255 Moleküle, Tabelle S1) trainiert wurde, um sowohl biologische Aktivität als auch Synthetisierbarkeit zu erfassen. Anschließend wurde

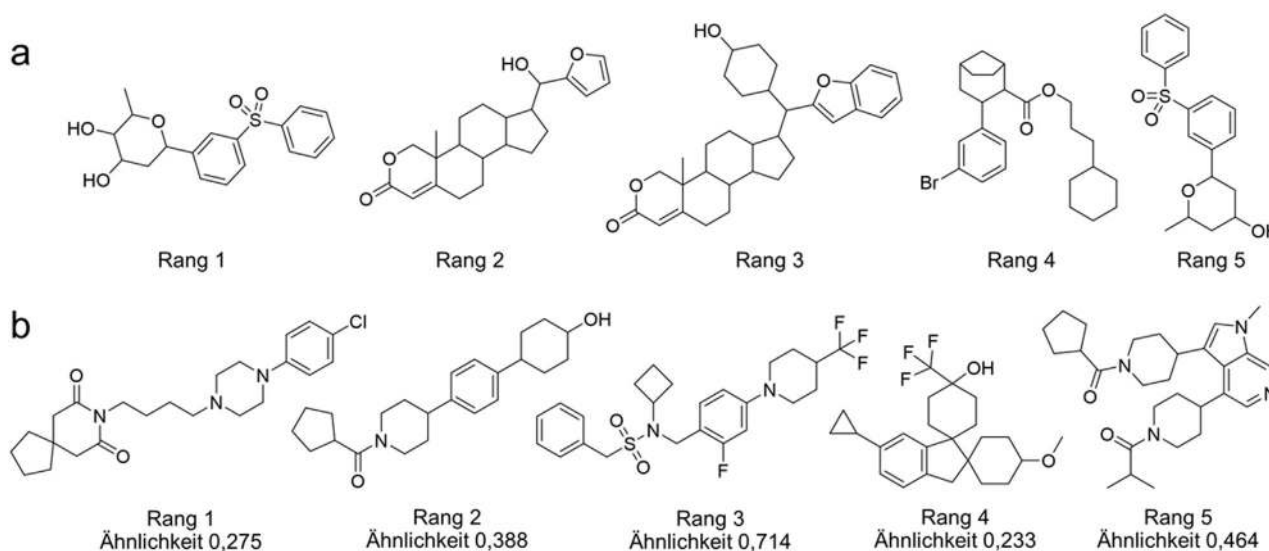


Abbildung 2. Höchstklassifizierte, durch Beam-Search-Sampling erhaltene Designs. a) Einfaches Fine-Tuning, b) zweistufiges Fine-Tuning. Ränge basieren auf dem Beam-Search-Score der Designs. Die Ähnlichkeitswerte der höchstklassifizierten Designs aus dem zweistufigen Fine-Tuning Experiment beziehen sich auf die auf Morgan-Fingerprints (Länge = 1024, Radius = 2 Bindungen) berechnete Tanimoto-Ähnlichkeit zum ähnlichsten bekannten aktiven Molekül mit einem IC₅₀-Wert an ROR γ in ChEMBL (Strukturen sind in Abbildung S2 gezeigt).

das CLM mit vier ROR γ -modulierenden Naturstoffen^[30] (Abbildung S1) für 16 Epochen mit dem Ziel weitertrainiert, das Modell in Richtung Naturstoff-Ähnlichkeit zu beeinflussen. Wie im ersten Experiment wurden dann alle gültigen SMILES-Strings untersucht, die das CLM durch Beam-Search-Sampling zwischen den Epochen 5 und 16 des (zweiten) Fine-Tuning-Schrittes generierte. Die fünf in diesem zweiten Ansatz designten Moleküle mit dem höchsten Beam-Search-Score (Abbildung 2b) waren gemäß IBM RXN^[40] synthetisch zugänglich; für jedes Design wurde eine Syntheseroute gefunden. Die computergenerierten Moleküle besaßen außerdem Naturstoffcharakteristika (Abbildung 3, Tabelle S2), was sich durch einen hohen Anteil an sp³-hybridisierten Kohlenstoffatomen (Fsp³) ausdrückte. Die Top-5-Designs wiesen Fsp³-Werte zwischen 50 % und 75 % auf, was mit den Werten für die MEGx-Naturstoffbibliothek (Analyticon Discovery GmbH, rel. 09-01-2018) vergleichbar war und die Fsp³-Werte der für das Pre-Training verwendeten ChEMBL-Moleküle (51 \pm 30 % bzw. 33 \pm 20 %) überstieg. Diese Ergebnisse deu-

teten also darauf hin, dass das zweistufige Fine-Tuning-Verfahren die Designziele erfüllte, sodass der zweistufige Ansatz für die prospektive Anwendung gewählt wurde.

Der Vergleich der Beam-Search Designs aus dem zweistufigen Fine-Tuning-Verfahren mit den Trainingsmolekülen und mit bekannten ROR γ -Modulatoren (Abbildung 3 a,b) zeigte, dass das Beam-Search-Sampling die Erkundung des chemischen Raums jenseits jener Regionen erlaubte, die von den Molekülen des Fine-Tunings besetzt sind, obwohl dieses Sampling-Verfahren die wahrscheinlichsten Token bei der Erzeugung neuer SMILES-Strings begünstigt und nur eine begrenzte Zahl an Möglichkeiten untersucht. Im Vergleich zu den in ChEMBL annotierten inversen ROR γ -Agonisten (IC₅₀ < 1 μ M) waren die Beam-Search-Designs außerdem strukturell vielfältiger im Hinblick auf die durch Morgan-Fingerprints^[42] dargestellten Substrukturfragmente (Abbildung 3 b). Gleichzeitig besaßen die Designs aber hinsichtlich ihrer dreidimensionalen Gestalt und Partialladungsverteilung (dargestellt durch die „Weighted Holistic Atom Localization and Entity Shape [WHALES]“-Deskriptoren^[43,44]) Ähnlichkeit zu den bekannten aktiven Molekülen. Offenbar lernte das CLM also zusätzlich zur SMILES-Syntax auch gewisse „semantische“ Strukturmerkmale, die für die Bindung an Makromoleküle relevant sind, wie z. B. molekulare Form und Partialladungsmuster.

Prospektive experimentelle Validierung

Auf Grundlage der Beam-Search-Scores wurden drei Designs zur Synthese und In-vitro-Charakterisierung ausgewählt. Von den fünf Designs mit höchstem Beam-Search-Score (Abbildung 2b) wählten wir Moleküle **1** und **2** vom ersten und dritten Rang aus. Verbindung **2** zeigte dabei die höchste Tanimoto-Ähnlichkeit (Morgan-Fingerprints) zu einem bekannten ROR γ -Modulator (Abbildung 2b). Die Grundgerüste der Verbindungen **1** und **2** waren auch unter Beam-Search-Designs jenseits der Top 5 verbreitet, was auf strukturelle Präferenzen hindeutete. Das „Scaffold“ (Molekülgerüst) von **1** fand sich im sechstplatzierten Design wieder, und die Moleküle auf den Rängen 10 und 13 wiesen eine hohe Ähnlichkeit zu Verbindung **2** auf, weshalb zusätzlich Verbindung **3** dieses prominenten Chemotyps von Rang 13 für eine prospektive Validierung ausgewählt wurde. Die Verbindungen **1–3** wurden gemäß Schema 1 synthetisiert.

Zur Herstellung von **1** wurden zunächst (4-Chlorphenyl)piperazin (**4**) und 4-Brombutylacetat (**5**) durch nukleophile Substitution zu **6** umgesetzt. Nach alkalischer Hydrolyse der Esterschutzgruppe in **6** wurde aus dem freien Alkohol **7** durch Mitsunobu-Reaktion mit 8-Azapiro[4.5]decan-7,9-dion (**8**) das Design **1** erhalten. Die Synthese von Design **2** begann ausgehend von 4-Brom-2-fluorbenzaldehyd (**9**), welcher durch reduktive Aminierung mit Cyclobutanamin (**10**) das sekundäre Amin **11** lieferte. Das Zwischenprodukt **11** wurde anschließend mit dem Sulfonylechlorid **12** zu **13** umgesetzt, bevor **13** im letzten Reaktionsschritt unter Buchwald-Hartwig-Bedingungen mit **14** das Design **2** ergab. Das strukturell verwandte Design **3** wurde, ausgehend von einer nukleophilen aromatischen Substitution mit 4-Trif-

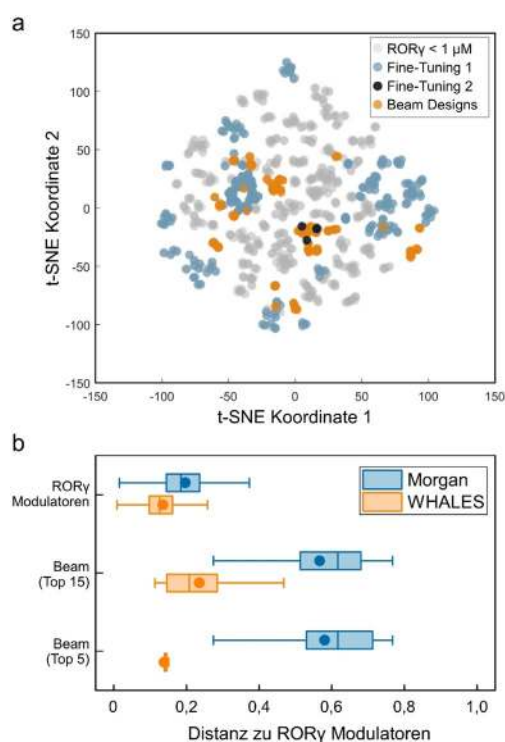
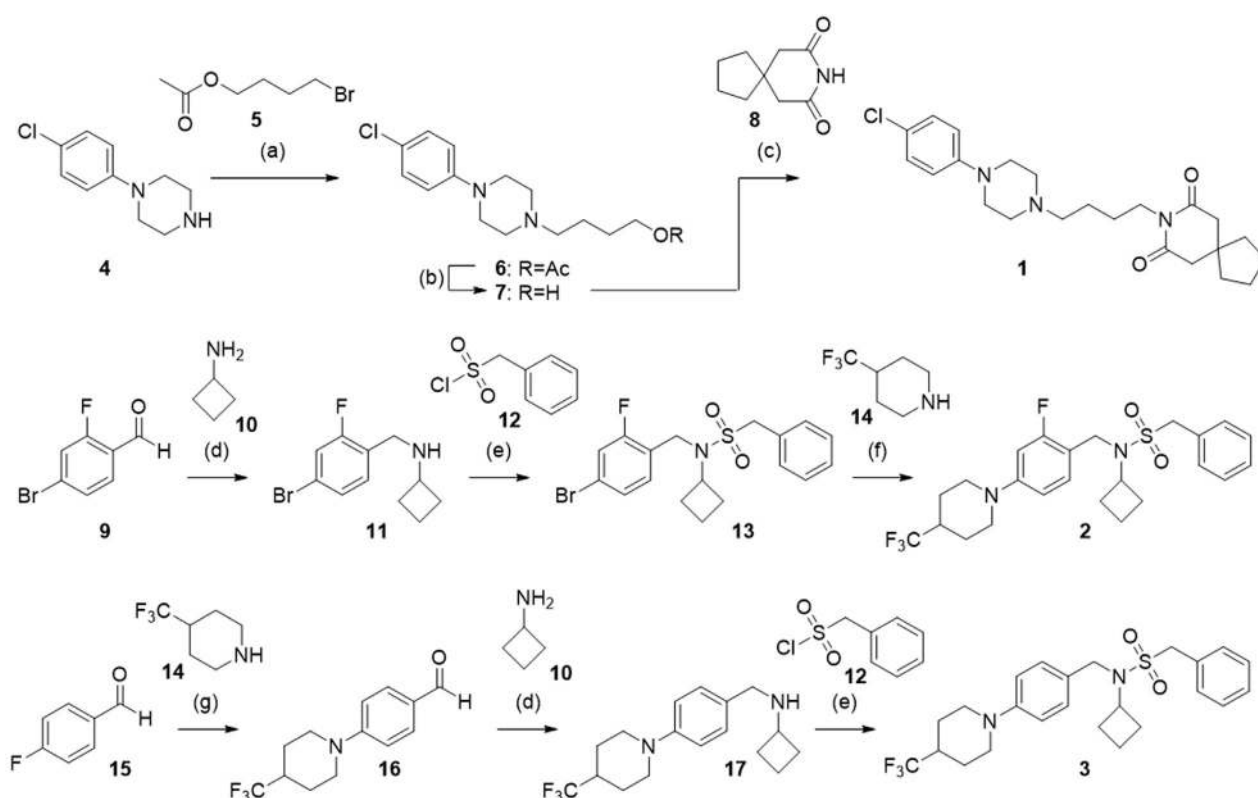


Abbildung 3. Charakteristika der CLM-Designs nach doppeltem Fine-Tuning. a) t-Verteilte stochastische Nachbareinbettung (t-SNE)^[45] der Molekülsets gemäß Morgan-Fragment-Fingerprints (Länge = 1024, Radius = 2 Bindungen, Tanimoto-Ähnlichkeit). Die beiden Fine-Tuning-Sets, die in ChEMBL enthaltenen ROR γ -Liganden (IC₅₀ < 1 μ M, 1091 Moleküle) und die Beam-Search-Designs sind gezeigt. b) Vergleich der Beam-Search-Designs mit bekannten ROR γ -Liganden (IC₅₀ < 1 μ M) hinsichtlich Morgan-Fragment-Fingerprints („Morgan“) sowie dreidimensionaler Gestalt und Partialladungsverteilung („WHALES“). Die paarweise Distanzverteilung zwischen in ChEMBL enthaltenen ROR γ -Liganden ist zum Vergleich gezeigt. Für Morgan-Fingerprints ist die Tanimoto-Distanz gezeigt, für WHALES die skalierte euklidische Distanz. „Beam (Top 15)“ und „Beam (Top 5)“ beziehen sich auf die 5 bzw. 15 höchstklassifizierten Beam-Search-Designs. Die Boxplots zeigen die 25., 50. und 75. Perzentile (Linien), Mittelwerte (Kreis) und Ausreißergrenzen (Whisker).

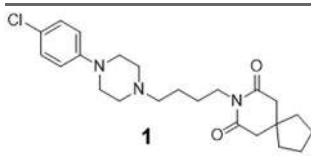
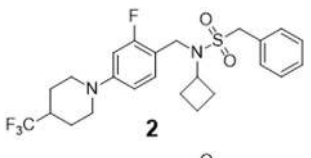
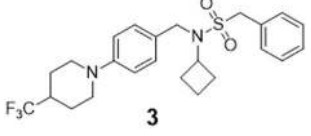


Scheme 1. Synthese der CLM-Designs **1**, **2** und **3**. Reagenzien und Bedingungen: a) DMF, 4-DMAP, 60 °C, 16 h, 48%; b) KOH, H₂O/THF/MeOH, MW, 100 °C, 30 min, 98%; c) DIAD, PPh₃, THF, 0 °C → RT, 16 h, 42%; d) NaB(OAc)₃H, HOAc, DCE, RT, 50 h, 73%; e) 4-DMAP, Pyridin, CH₂Cl₂, Reflux, 16 h, 37%; f) Pd₂(dba)₃, Xantphos, Cs₂CO₃, 1,4-Dioxan, Reflux, 16 h, 18%; g) K₂CO₃, DMSO, Reflux, 48 h, 82%.

luormethylpiperidin (**14**) und 4-Fluorbenzaldehyd (**15**), über eine alternative Route synthetisiert. Die nukleophile aromatische Substitution lieferte eine höhere Ausbeute (siehe Schema 1) als die Buchwald-Hartwig-Reaktion, konnte aber wegen der möglichen Bildung von Regioisomeren nicht für die Synthese von **2** angewendet werden. Reduktive Aminierung des Substitutionsproduktes **16** mit Cyclobutanamin (**10**) zu **17**, gefolgt von einer Sulfonamidbildung mit Phenylmethansulfonylchlorid (**12**), lieferte Design **3**.

Die In-vitro-Charakterisierung der Verbindungen **1**, **2** und **3** in Gal4-ROR-Hybrid-Reporter-Gen-Assays bestätigte den angestrebten inversen ROR γ -Agonismus mit mikromolaren bis submikromolaren IC₅₀-Werten (Tabelle 1). Die gemäß Beam-Search-Score ranghöchste Verbindung **1** wirkte der basalen ROR γ -Aktivität mit einem IC₅₀-Wert von 4,6 μ M entgegen. Sie war auch an ROR α und ROR β aktiv, genaue IC₅₀-Werte konnten aber aufgrund von Zytotoxizität nicht bestimmt werden. Die Verbindungen **2** und **3** zeigten inversen ROR γ -Agonismus mit IC₅₀-Werten von 0,37 μ M (**2**) bzw. 0,68 μ M (**3**). Neben dem angestrebten inversen ROR γ -Agonismus wiesen alle drei synthetisierten Designs eine ausgeprägte Präferenz für den ROR γ -Subtyp auf, wobei die Verbindungen **2** und **3** im Vergleich zu den verwandten ROR α - und ROR β -Isoformen eine mehr als zehnfach höhere Potenz an ROR γ besaßen. Diese Ergebnisse zeigen, dass das CLM mit Beam-Search-Sampling die biologische Aktivität der Trainingsmoleküle in den De-novo-Designs konservierte.

Tabelle 1: Biologische Aktivität der CLM-Designs **1**, **2** und **3** an den RORs in Gal4-Hybrid-Reporter-Gen-Assays. Daten sind als Mittelwert \pm S.E.M. dargestellt, $n \geq 4$.

Struktur und ID	IC ₅₀ [μ M]		
	ROR α	ROR β	ROR γ
	> 10	> 10	4,6 \pm 0,5
	23 \pm 3	22 \pm 1	0,37 \pm 0,05
	10 \pm 1	7,6 \pm 0,5	0,68 \pm 0,07

Fazit

Der Beam-Search-Algorithmus wurde zum De-novo-Design mit einem CLM angewendet. Dabei erzeugte und bewertete der Algorithmus die Designs automatisch, ohne dass zusätzliche Priorisierungsregeln erforderlich waren. Die

prospektive experimentelle Validierung der Methode lieferte neue Moleküle, die mit inversem Agonismus am Ligand-aktivierten Transkriptionsfaktor ROR γ die angestrebte biologische Aktivität besaßen und unterschiedliche Grade an Ähnlichkeit zu bekannten ROR γ -Modulatoren aufwiesen (0,28–0,71 Tanimoto-Ähnlichkeit auf Morgan-Fingerprints). Der Beam-Search-Algorithmus in Verbindung mit einem CLM bewahrte also offenbar strukturelle Merkmale, die für die angestrebte biologische Aktivität notwendig sind, entwarf dabei aber strukturell diverse Moleküle. Dieses Ergebnis bestätigt Beam-Search-Sampling als eine geeignete Technik zum De-novo-Design bioaktiver Moleküle durch ein CLM. Die rechnerischen und experimentellen Ergebnisse weisen außerdem auf zwei attraktive Eigenschaften des Beam-Search-Algorithmus in dieser Anwendung hin. Zum einen zeigt der Beam-Search-Algorithmus durch die Suche nach den wahrscheinlichsten Molekülen, die ein CLM generieren kann, die Eignung des Modells für die zu untersuchende Aufgabenstellung auf. Die Auswertung der Beam-Search Designs erlaubt die Überprüfung, ob die Moleküle mit den Designzielen übereinstimmen, und damit eine Bewertung des Fine-Tuning-Erfolgs. Dies steht im Gegensatz zum herkömmlichen Temperatur-Sampling, das Chemiker dazu verleiten könnte, Designs in Betracht zu ziehen, die gemäß dem Modell nicht wahrscheinlich sind. Zum anderen könnte der Beam-Search-Score, der eine intrinsische Klassifizierung ermöglicht, die Notwendigkeit der externen Priorisierung von Designs überwinden. Es ist jedoch zu beachten, dass die Anzahl der Designs, die durch Beam-Search erstellt werden können, begrenzt ist, während Temperatur-Sampling eine praktisch unendliche Anzahl von chemischen Strukturen generieren kann. Beide Techniken ergänzen sich gegenseitig und bieten jeweils Vorteile. Die angestrebte Anwendung sollte die Wahl der einen oder anderen Strategie leiten. Wenn zukünftige prospektive Studien diese Beobachtungen bestätigen, könnte das Beam-Search-Sampling dazu beitragen, die Anwendbarkeit von CLMs für das De-novo-Design in der medizinischen Chemie zu stärken.

Acknowledgements

Diese Forschung wurde durch den Schweizer Nationalfonds (grant no. 205321_182176 an G.S.), die RETHINK Initiative der ETH Zürich und die Novartis Forschungsstiftung (Free-Novation grant „AI in Drug Discovery“ an G.S.) gefördert. Open Access Veröffentlichung ermöglicht und organisiert durch Projekt DEAL.

Interessenkonflikt

G.S. erklärt einen möglichen finanziellen Interessenkonflikt als Gründer der inSili.com GmbH, Zürich, und in seiner Rolle als Berater der pharmazeutischen Industrie.

Stichwörter: De-novo-Design · Deep Learning · Kernrezeptor · Neuronale Netze · Wirkstoffforschung

- [1] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85–117.
- [2] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.
- [3] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241–1250.
- [4] W. P. Walters, R. Barzilay, *Acc. Chem. Res.* **2021**, *54*, 263–270.
- [5] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.
- [6] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268–276.
- [7] N. de Cao, T. Kipf, *arXiv* **2018**, <https://arxiv.org/abs/1805.11973>.
- [8] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700111.
- [9] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.
- [10] J. Bradshaw, B. Paige, M. J. Kusner, M. H. S. Segler, J. M. Hernández-Lobato, *arXiv* **2020**, <https://arxiv.org/abs/2012.11522>.
- [11] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- [12] M. A. Skinnider, R. G. Stacey, D. S. Wishart, L. J. Foster, *ChemRxiv* **2021**, <https://doi.org/10.26434/CHEMRXIV.13638347.V1>.
- [13] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2018**, *4*, 120–131.
- [14] W. Yuan, D. Jiang, D. K. Nambiar, L. P. Liew, M. P. Hay, J. Bloomstein, P. Lu, B. Turner, Q. T. Le, R. Tibshirani, P. Khatri, M. G. Moloney, A. C. Koong, *J. Chem. Inf. Model.* **2017**, *57*, 875–882.
- [15] D. Merk, F. Grisoni, L. Friedrich, G. Schneider, *Commun. Chem.* **2018**, *1*, 68.
- [16] Y. Yang, R. Zhang, Z. Li, L. Mei, S. Wan, H. Ding, Z. Chen, J. Xing, H. Feng, J. Han, H. Jiang, M. Zheng, C. Luo, B. Zhou, *J. Med. Chem.* **2020**, *63*, 1337–1360.
- [17] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, *Nat. Mach. Intell.* **2020**, *2*, 171–180.
- [18] M. Awale, F. Sirockin, N. Stiefl, J. L. Reymond, *J. Chem. Inf. Model.* **2019**, *59*, 1347–1356.
- [19] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3320–3328.
- [20] M. Peters, S. Ruder, N. A. Smith, *arXiv* **2019**, <https://arxiv.org/abs/1903.05987>.
- [21] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283–293.
- [22] Y. Wang, Q. Yao, J. Kwok, L. M. Ni, *arXiv* **2019**, <https://arxiv.org/abs/1904.05046>.
- [23] X. Yang, J. Zhang, K. Yoshizoe, K. Terayama, K. Tsuda, *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.
- [24] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583.
- [25] D. Grechishnikova, *Sci. Rep.* **2021**, *11*, 321.
- [26] G. Benoit, A. Cooney, V. Giguere, H. Ingraham, M. Lazar, G. Muscat, T. Perlmann, J. P. Renaud, J. Schwabe, F. Sladek, M. J. Tsai, V. Laudet, *Pharmacol. Rev.* **2006**, *58*, 798–836.
- [27] D. P. Marciano, M. R. Chang, C. A. Corzo, D. Goswami, V. Q. Lam, B. D. Pascal, P. R. Griffin, *Cell Metab.* **2014**, *19*, 193–208.
- [28] Y. Hoon Kim, M. A. Lazar, *Endocr. Rev.* **2020**, *41*, 707–732.
- [29] V. B. Pandya, S. Kumar, Sachchidanand, R. Sharma, R. C. Desai, *J. Med. Chem.* **2018**, *61*, 10976–10995.
- [30] L. A. Solt, T. P. Burris, *Trends Endocrinol. Metab.* **2012**, *23*, 619–627.
- [31] S. Asimus, R. Palmér, M. Albayaty, H. Forsman, C. Lundin, M. Olsson, R. Pehrson, J. Mo, M. Russell, S. Carlert, D. Close, D. Keeling, *Br. J. Clin. Pharmacol.* **2020**, *86*, 1398–1405.
- [32] D. J. Kojetin, T. P. Burris, *Nat. Rev. Drug Discovery* **2014**, *13*, 197–216.

- [33] B. T. Lowerre, PhD Thesis, Carnegie Mellon Univ. Pittsburgh, **1976**.
- [34] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735–1780.
- [35] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- [36] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2020**, *83*, 770–803.
- [37] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, *Nat. Chem.* **2016**, *8*, 531–541.
- [38] P. Ertl, A. Schuffenhauer, in *Prog. Drug Res.*, Birkhäuser, Basel, **2008**, S. 217–235.
- [39] P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
- [40] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci.* **2020**, *11*, 3316–3325.
- [41] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, C. Zardecki, *Acta Crystallogr. Sect. D* **2002**, *58*, 899–907.
- [42] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.
- [43] F. Grisoni, D. Merk, V. Consonni, J. A. Hiss, S. G. Tagliabue, R. Todeschini, G. Schneider, *Commun. Chem.* **2018**, *1*, 44.
- [44] F. Grisoni, G. Schneider, *Methods Mol. Biol.* **2021**, *2266*, 11–35.
- [45] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Manuskript erhalten: 30. März 2021

Veränderte Fassung erhalten: 2. Juni 2021

Akzeptierte Fassung online: 24. Juni 2021

Endgültige Fassung online: 19. Juli 2021