KI-Proteindesign mit offenen Werkzeugen

Thomas Schüller · Syntagma Forschung · ORCID 0009-0003-9799-6747 · Fassung 0.3 · 06.11.2026 · Alle Rechte vorbehalten

Vorwort und Geltungsbereich

Arbeitstitel: KI-Protein-Design – Werkzeuge, Ketten, Daten und Grenzen
Autor: Thomas Schüller · ORCID 0009-0003-9799-6747 · Syntagma Forschung · Wien

Anlass

Ein Social-Media-Beitrag fasst einen verbreiteten Arbeitsablauf zusammen: Ein Modell entwirft Proteinrückgrate, ein zweites entwirft Sequenzen dazu, ein drittes sagt den Komplex aus Ziel und Binder voraus, anschließend wird im Labor geprüft. Dieses Werk arbeitet den Ablauf auf, ordnet die beteiligten Werkzeuge ein, benennt Lizenz- und Hardwaregrenzen und legt offen, was Vorhersagen leisten und was nicht.

Was das Werk ist

- ein Überblick über offene und geschlossene Werkzeuge für Strukturvorhersage, Binder-Design, Sequenzdesign, Docking und Strukturdatenbanken (Kapitel 2 und 7),
- eine Beschreibung der Werkzeugkette mit Prüfpunkten (Kapitel 3),
- ein Kapitel zu Daten, Sequenzierung und Arbeitsumgebungen (Kapitel 4 bis 6),
- ein Kapitel zu Validierung, Grenzen und Biosicherheit (Kapitel 8 und 9),
- ein Planungswerkzeug (KETTENBAU), das aus Aufgabe, Nutzungsart und Hardware einen Kettenplan erzeugt (Kapitel 10),
- eine Lernreihenfolge, Evidenzstufen und eine Vorlage für technische Rückfragen (Kapitel 11).

Was das Werk nicht ist

Es ist keine Laboranleitung und keine Bauanleitung für bestimmte Proteine. Es enthält keine Entwurfsparameter für konkrete Ziele. Lizenzangaben sind Arbeitsstand und kein Rechtsrat.

Kodex

1. **Keine militärische oder waffenbezogene Nutzung.** Das Werk und das Werkzeug dienen ausschließlich zivilen Zwecken.
2. **Keine Schadensziele.** Es werden keine Binder oder Enzyme gegen Toxine, Krankheitserreger-Faktoren oder andere Ziele mit Missbrauchspotenzial geplant.
3. **Vorhersage ist kein Nachweis.** Bindung, Aktivität und Funktion werden nur durch Experimente belegt.
4. **Quellenstatus offen legen.** Jede Angabe im Werkzeugregister trägt den Prüfstatus: „geprüft“ (in dieser Fassung gegen die Projektseite oder Veröffentlichung gelesen) oder „nicht geprüft“ (aus Vorwissen übernommen).
5. **Offenlegung zum KI-Einsatz.** Texte, Register und Werkzeug wurden mit Unterstützung von KI-Werkzeugen erstellt und redaktionell geprüft.

Lesehinweis zum Prüfstatus

Die Recherche zu diesem Werk lief am 02.11.2026; die Lizenzangaben der Werkzeuge wurden am 04.11.2026 nachgeprüft. Alles, was sich danach ändert (Lizenzen, Versionen, Gewichte), ist vor Nutzung neu zu prüfen. Bei Widersprüchen zwischen Quellen steht das im Text; die LICENSE-Datei im jeweiligen Projekt ist maßgeblich.

1 Von der Struktur zur Funktion

1.1 Drei Fragen, drei Modellklassen

Proteinentwurf mit KI beantwortet drei verschiedene Fragen, die oft vermischt werden:

Frage	Modellklasse	Beispiele (Stand 02.11.2026)
Wie faltet eine gegebene Sequenz, allein oder im Komplex?	Strukturvorhersage, Ko-Faltung	AlphaFold 2 und 3, OpenFold3, Boltz-2, Chai-1, RoseTTAFold3
Welche Form könnte an ein Ziel binden?	Generatives Design (Rückgrat oder All-Atom)	BoltzGen, RFdiffusion (1 und 3)
Welche Sequenz faltet in diese Form?	Inverse Faltung	ProteinMPNN, LigandMPNN

Dazu kommen Werkzeuge für Docking kleiner Moleküle (AutoDock Vina, GNINA, DiffDock-L), für Strukturvergleich (Foldseek) und für die Orchestrierung ganzer Läufe (nf-core/proteinfold).

1.2 Was sich 2025 und 2026 verschoben hat

- **Offene Ko-Faltungsmodelle.** AlphaFold 3 erschien mit Code unter einer nichtkommerziellen Lizenz und Gewichten nur auf Antrag. Offene Alternativen folgten: OpenFold3 (Apache-2.0, Vorabversion seit Oktober 2025), Boltz-1 und Boltz-2 (MIT), Chai-1 sowie RoseTTAFold3 im Foundry-Paket der Rosetta Commons.
- **Design und Vorhersage in einem Modell.** BoltzGen (MIT Jameel Clinic, Preprint November 2025) vereint beides und steht unter MIT-Lizenz mit Daten, Gewichten und Trainingscode.
- **All-Atom-Design.** RFdiffusion3 wurde am 03.12.2025 als Open Source veröffentlicht und modelliert alle Atome des Polymers, einschließlich Liganden und Nukleinsäuren.
- **Geschlossene Spitzenmodelle.** Chai-2 (Juni 2025) und Chai-3 (Mai 2026, laut Anbieter früher Zugang für Partner) sind nicht öffentlich verfügbar. Sie dienen hier nur als Literaturreferenz.

1.3 Warum eine Kette

Kein einzelnes Modell gilt als Beleg für einen funktionierenden Binder. Üblich ist deshalb ein gestaffeltes Verfahren: viele Entwürfe erzeugen, mit einem zweiten, unabhängigen Modell neu falten, nach vorab festgelegten Schwellen filtern, wenige Kandidaten ins Labor schicken. Jede Stufe entfernt Fehlalarme der vorherigen. Kapitel 3 beschreibt die Stufen, Kapitel 8 die Grenzen.

2 Werkzeugkarte

Die vollständige, maschinenlesbare Fassung steht im Werkzeugregister (`werkzeuge.json`) und im Werkzeug KETTENBAU. Die Tabellen zeigen den Stand vom 04.11.2026 (Erstfassung 02.11.2026, Lizenznachprüfung 04.11.2026). „Geprüft“ heißt: Angabe in dieser Fassung gegen Projektseite oder Veröffentlichung gelesen.

2.1 Strukturvorhersage und Ko-Faltung

Werkzeug	Ziele	Lizenzstatus	Gewerblich	Geprüft
AlphaFold 3	Protein, Ligand, Nukleinsäure	Code CC-BY-NC-SA 4.0; Gewichte und Output nur für nichtkommerzielle Organisationen	nein	ja
OpenFold3 (preview)	Protein, Ligand, Nukleinsäure	Apache-2.0, Code und Gewichte	ja	ja
Boltz-2	Protein, Ligand, Nukleinsäure; Affinität	MIT	ja	ja
Chai-1	Protein, Ligand, Nukleinsäure	Apache-2.0 laut README und Modellkarte; ältere nichtkommerzielle Angaben überholt	ja	ja
RoseTTAFold3	Protein, Ligand, Nukleinsäure	BSD-3-Clause laut RosettaCommons (Foundry); Lizenz der Gewichte nicht gesondert gelesen	ja	ja
AlphaFold 2	Protein	Code Apache-2.0, Parameter CC-BY-4.0 (Vorwissen)	prüfen	nein
ESMFold	Protein	MIT laut Hugging-Face-Modellkarte	ja	ja
ColabFold	Protein	MIT (Vorwissen)	prüfen	nein

Zu AlphaFold 3. Die Nutzungsbedingungen der Gewichte beschränken Nutzung und Output auf nichtkommerzielle Organisationen (Universitäten, gemeinnützige Einrichtungen, Forschungsinstitute, Bildungs-, Journalismus- und Regierungsstellen). Output darf nicht zum Training ähnlicher Strukturmodelle dienen. Wer als Einzelunternehmen oder gewerblich ausgerichtete Stelle arbeitet, fällt nach dem Wortlaut nicht unter die erlaubten Nutzer. Auch der AlphaFold Server ist nur für nichtkommerzielle Nutzung vorgesehen. KETTENBAU schließt AlphaFold 3 deshalb bei Nutzungsart „gewerblich“ aus.

2.2 Generatives Design und Sequenzdesign

Werkzeug	Aufgabe	Lizenzstatus	Gewerblich	Geprüft
BoltzGen	Binder für Protein, Peptid, kleine Moleküle, Nukleinsäuren; Design und Faltung in einem Modell	MIT, einschließlich Training	ja	ja

Werkzeug	Aufgabe	Lizenzstatus	Gewerblich	Geprüft
RFdiffusion3	All-Atom-Design (Binder, Enzyme)	Open Source seit 03.12.2025; BSD-3-Clause laut RosettaCommons (Foundry); Lizenz der Gewichte nicht gesondert gelesen	ja	ja
RFdiffusion (1)	Rückgrat-Design	BSD laut README, auch gewerblich frei	ja	ja
ProteinMPNN	Sequenz zu gegebenem Rückgrat	MIT laut Repository und RosettaCommons	ja	ja
LigandMPNN	Sequenz mit Ligandenkontext	MIT, Code und Modellparameter laut README	ja	ja
Chai-2 und Chai-3	Antikörper- und Binder-Design	geschlossen	nein	ja

2.3 Docking, Suche, Orchestrierung

Werkzeug	Aufgabe	Anmerkung	Geprüft
AutoDock Vina	Docking kleiner Moleküle	CPU, schnell, verbreitete Basislinie; Apache-2.0	ja
GNINA	Docking mit CNN-Bewertung	in Benchmarks starke offene Basislinie und Nachbewerter; Apache-2.0 laut Repository-Metadaten	ja
DiffDock-L	generatives Docking, Blind-Docking möglich	bei bekannter Tasche in neueren Benchmarks hinter Vina und GNINA; Posen auf Plausibilität prüfen	teilweise
Foldseek	Strukturvergleich gegen PDB und AlphaFold DB	GPU-Unterstützung für Monomer- und Multimersuche; GPL-3.0, Copyleft bei Weitergabe	ja
nf-core/proteinfold	Nextflow-Pipeline für AlphaFold 2, ColabFold, ESMFold	Docker, Singularity, Conda; MIT (Pflicht für nf-core-Pipelines), aufgerufene Modelle mit eigenen Lizenzen	ja

2.4 Was die Benchmarks hergeben

Zwei Beispiele aus 2026 zeigen, wie vorsichtig Rangfolgen zu lesen sind. Ein Vergleich von 15 offenen Affinitätsverfahren an 853 antiviralen Verbindungen und 16 Zielproteinen sah Boltz-2 und DrugFormDTA unter den lernbasierten Verfahren vorn und GNINA unter den Docking-Verfahren, mit deutlichen Schwankungen je Zielprotein (arXiv 2605.04265). Ein Vergleich auf PDBBind-Daten ergab nach einer Zusammenfassung der Arbeit für Vina und GNINA bei Redocking mit bekannter Tasche eine um etwa 20 bis 25 Prozentpunkte höhere Trefferquote als für DiffDock. Beides gilt für die jeweiligen Datensätze und nicht für beliebige Ziele.

3 Die Werkzeugkette

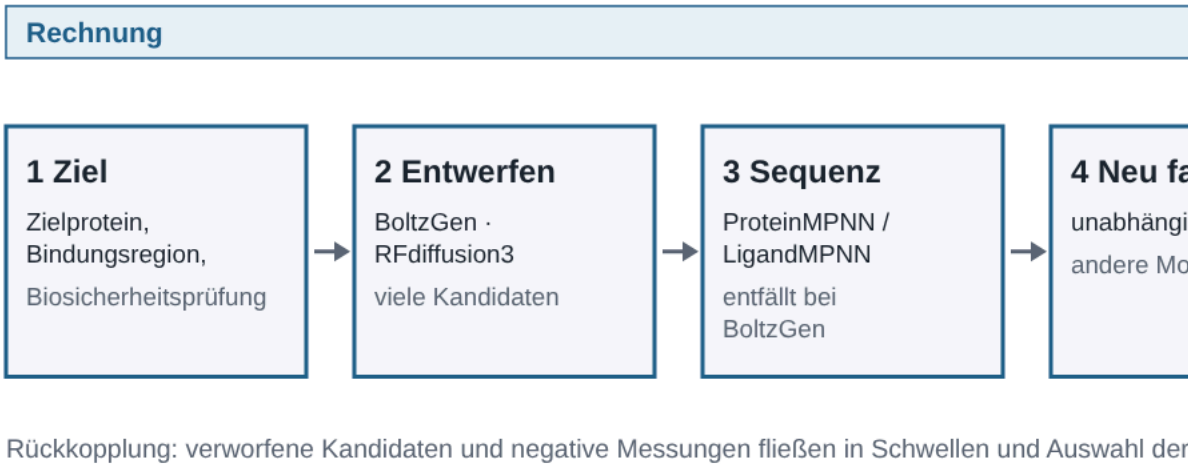


Abbildung 1: Werkzeugkette vom Ziel bis zur Laborprüfung

3.1 Stufen

1. **Ziel festlegen.** Zielprotein, Bindungsregion (soweit die Aufgabe eine verlangt) und Bezugsstruktur aus PDB oder AlphaFold DB. Eine Biosicherheitsprüfung des Ziels steht vor dem ersten Rechenlauf (Kapitel 9).
2. **Entwerfen.** Ein generatives Modell erzeugt viele Kandidaten (BoltzGen, RFdiffusion3 oder RFdiffusion 1). Die Zahl der Entwürfe ist ein Planungsparameter, keine Qualitätsangabe.
3. **Sequenz entwerfen.** Erzeugt das Designmodell nur ein Rückgrat, folgt Inverse Faltung mit ProteinMPNN oder LigandMPNN. BoltzGen entwirft Sequenz und Struktur gemeinsam, die Stufe entfällt.
4. **Neu falten mit einem unabhängigen Modell.** Jeder Kandidat wird im Komplex mit dem Ziel durch ein Modell vorhergesagt, das nicht dieselbe Familie wie das Designmodell ist. KETTENBAU ordnet deshalb Modellfamilien zu (Boltz, Foundry, Chai, OpenFold, AlphaFold) und wählt als Bewertungsmodell möglichst eine andere Familie.
5. **Filtern.** Schwellen werden vor dem Lauf festgelegt: Grenzflächen-ipTM, pLDDT an der Grenzfläche, PAE der Grenzfläche, Zahl sterischer Konflikte. Nachträgliches Verschieben der Schwellen ist Stichprobenfischen.
6. **Gen bestellen und Sequenz prüfen.** Synthese über einen Anbieter mit dokumentiertem Screening; nach der Klonierung wird der Klon sequenziert und mit der Entwurfssequenz verglichen (Kapitel 5).
7. **Bindung messen.** Erst nach Identitäts- und Reinheitsnachweis des Proteins (Kapitel 5).

3.2 Kennzahlen und ihre Lesart

Kennzahl	Bedeutung	Typischer Fehlschluss
pLDDT	lokale Konfidenz je Rest	niedrige Werte als Fehler lesen; sie markieren oft ungeordnete Bereiche
PAE	erwarteter Positionsfehler zwischen Restpaaren	nur Mittelwerte lesen; die Grenzflächenblöcke zählen
pTM, ipTM	globale Konfidenz, Konfidenz der Grenzfläche	ipTM als Bindungsstärke deuten
RMSD zum Entwurf	Abweichung zwischen Entwurf und Neufaltung	Übereinstimmung zweier verwandter Modelle als unabhängige Bestätigung deuten
Sterische Konflikte	Atomüberlappungen	wegretuschieren statt Kandidat verwerfen

3.3 Typische Fehlerquellen

- **Gleiches Modell entwirft und bewertet.** Das Ergebnis ist eine Selbstbestätigung. Auch verwandte Modelle (BoltzGen und Boltz-2 teilen Architektur und Trainingsdaten) liefern nur eine abgeschwächte Unabhängigkeit.
- **Zielstruktur aus einer Vorhersage.** Die AlphaFold DB enthält Vorhersagen; eine Bindungsregion, die nur dort sichtbar ist, trägt die Unsicherheit der Vorhersage mit.
- **Zu enge Datenbasis.** Ziele, die Trainingskomplexen stark ähneln, schneiden besser ab als neuartige. Die BoltzGen-Entwickler berichten deshalb ausdrücklich über Ziele ohne bekannte gebundene Strukturen.
- **Nicht dokumentierte Läufe.** Ohne Versionen, Gewichte-Prüfsummen und Zufallsstartwerte lässt sich ein Ergebnis nicht wiederholen (Kapitel 6).

4 Datenbanken und Daten

Quelle	Inhalt	Zugang	Prüfstand
AlphaFold Protein Structure Database (EMBL-EBI und Google DeepMind)	241.070.489 Strukturvorhersagen in Version 6, synchron mit UniProt 2025_03, darunter 40.054 Isoformen	Web, API, FTP, Google Cloud	geprüft
Protein Data Bank (RCSB, wwPDB)	experimentelle Strukturen	Web, API, FTP	nicht geprüft
UniProt	Sequenzen und Annotation	Web, REST, FTP	nicht geprüft
MGnify	metagenomische Sequenzen	Web, API	nicht geprüft

4.1 Hinweise aus den Release Notes der AlphaFold DB

Version 6 ergänzte netto über 26 Millionen Vorhersagen und umfasst damit mehr als 241 Millionen Strukturen. Gegenüber Version 4 wurden rund 65,7 Millionen Einträge hinzugefügt und rund 39,3 Millionen entfernt; die Koordinaten unveränderter Einträge stammen aus Version 4, nur die Metadaten wurden aktualisiert. Alte API-Felder liefen laut Hinweis am 25.06.2026 aus. Skripte, die ältere Felder lesen, sind zu prüfen. Ältere Versionen bleiben als vollständige Proteome im FTP-Bereich erhältlich. Seit März 2026 zeigt die Seite zusätzlich Proteinkomplexe.

4.2 Regeln für den Umgang

1. Datenbankstand (Version, Datum) im Laufprotokoll festhalten.
2. Vorhergesagte Strukturen als solche kennzeichnen und nicht mit experimentellen mischen.
3. Für Strukturvergleiche Foldseek gegen PDB und AlphaFold DB nutzen und Treffer nach Abdeckung und Score filtern.
4. Lokale Kopien (Mini-Datenbanken, nach nf-core rund 340 GB, volle Datenbanken bis etwa 3 TB) einplanen, bevor die Hardware angeschafft wird.

5 Sequenzierung und Laborschnittstelle

Dieses Kapitel beschreibt die Schnittstelle zwischen Rechnung und Labor auf Planungsebene. Es enthält keine Laborprotokolle.

5.1 Von der Entwurfssequenz zum geprüften Klon

Schritt	Zweck	Nachweis
Gen-Synthese bestellen	Entwurfssequenz als DNA erhalten	Anbieter mit dokumentiertem Biosicherheits-Screening; Auftragsbestätigung bewahren
Klonierung	Gen in Expressionsvektor bringen	Plasmidkarte mit Version
Sequenzprüfung des Klons	Identität mit der Entwurfssequenz belegen	Sanger-Sequenzierung für kurze Konstrukte; Nanopore-Sequenzierung für ganze Plasmide; Abgleich gegen die Entwurfssequenz
Expression und Reinigung	Protein gewinnen	Identität (z. B. Massenspektrometrie) und Reinheit
Bindungsmessung	Bindung belegen oder widerlegen	z. B. SPR oder BLI mit Negativkontrolle

5.2 Warum die Sequenzprüfung vor der Messung steht

Ein Messwert gehört nur dann zu einem Entwurf, wenn der gemessene Klon der Entwurfssequenz entspricht. Fehler bei Synthese oder Klonierung erzeugen sonst scheinbare Treffer oder scheinbare Fehlschläge, die dem Modell zugeschrieben werden. Die Sequenzdaten gehören zusammen mit Entwurf, Modellversion und Messwert in dieselbe Aufzeichnung.

5.3 Datenhaltung

Je Kandidat eine Kennung, die durch alle Stufen mitläuft: Entwurfsdatei, Sequenz, Neufaltung, Kennzahlen, Bestelldatum, Klon-Sequenzdatei, Messdaten. Negative Ergebnisse werden gleichrangig gespeichert, weil nur sie die Trefferquote eines Verfahrens ehrlich abbilden.

6 Arbeitsumgebungen und Reproduzierbarkeit

6.1 Grundsatz

Eine Umgebung je Werkzeug, getrennt von den anderen. Die Werkzeuge haben unterschiedliche Abhängigkeiten, und gemeinsame Umgebungen brechen bei Aktualisierungen.