AlphaFold

Colin Zach – Ai Sommer School 23

1. Proteinfaltung

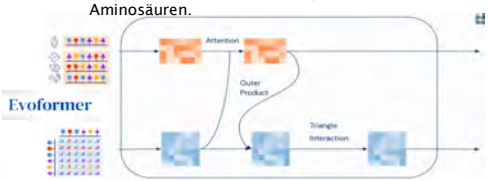
- Proteine sind wichtige **Körperbausteine** mit vielfältigen Funktionen.
- Als **Enzyme** katalysieren sie Reaktionen, **Antikörper** schützen vor Krankheiten, **Strukturproteine** formen Gewebe.
- Proteine bestehen aus **Aminosäureketten,** dessen Wechselwirkungen bestimmen **Struktur und Faltung**.
- Die Aminosäuren-Abfolge beeinflusst die Struktur.
- Die **3D-Anordnung beeinflusst Interaktionen und Funktionen**.

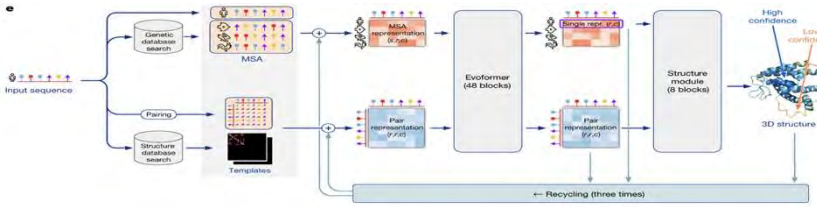
2. Schwierigkeit der Faltung

- Proteinfaltung, von Sequenz zur 3D-Struktur, bleibt eine wissenschaftliche Herausforderung.
- Es gibt bis zu **10^{50} - 10^{300} mögliche Anordnungen**, und meistens nur 1 passende Struktur.
- Experimentelle Bestimmung langwierig, nur **wenige Strukturen entschlüsselt** (200k in PDB).
- Wenige Daten für neuronales Netztraining zur Strukturvorhersage, kreative Ansätze nötig.

3. Aufbau des AlphaFold-Systems

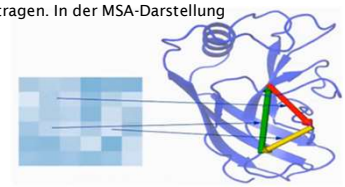
- AlphaFold nutzt zwei Hauptinformationsquellen: **„MSA-Repräsentation“** (Multiple Sequence Alignment) und **„Paar-Repräsentation“**. MSA zeigt **Aminosäure-Anordnung** durch **evolutionäre Ähnlichkeit**, Paar-Repräsentation enthält **Bindungs-, Distanz- und Interaktionsinfos** der Aminosäuren.





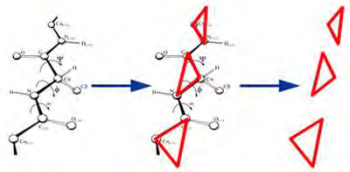
- Der **Evoformer** ist das **zentrale neuronale Netzwerk** in AlphaFold, das die MSA- und Paar-Informationsquellen optimiert. Durch ein **Attention-Modell** werden die Daten angepasst und Informationen zwischen den beiden **Repräsentationen übertragen**.
- Das **„Outer Product“** ermöglicht **Kommunikation** zwischen **MSA- und Paar-Repräsentation**. Infos von MSA zu Paardarstellungen übertragen. In der MSA-Darstellung **multipliziert** es **Elemente der Paardarstellung** mit **MSA-Sequenzdimension** und **summiert** die Ergebnisse.

Das **„Structure Module“** von AlphaFold nutzt **3D-Proteinstruktur** als **“Frames“** (siehe Bild unten links: rote Dreiecke). Frames iterativ verfeinert für **genauere Anordnung**. Seitenketteninfos nicht berücksichtigt, trotzdem hohe Genauigkeit

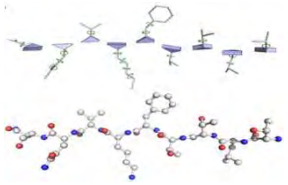


- Im Evoformer nutzen **„Triangle Interactions“** Distanzen zur Berechnung dritter Distanz (siehe Bild rechts). Erzeugt zusätzliche Infos und indirekte Aminosäurebeziehungen. Aus **Paaren AB und BC** folgt Distanz des **Paares AC**.

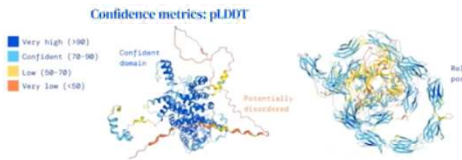
Protein backbone = gas of 3-D rigid bodies (chain is learned!)

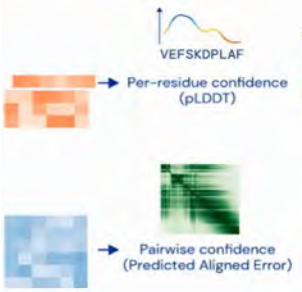


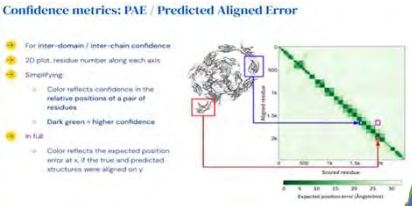
- Ein **Kernstück des Structure Modules** ist **„Invariant Point Attention“** (IPA). Erzeugt **Räumliche Ausrichtung** der Frames und fokussiert relevante Strukturmerkmale durch **Zuweisung von "Values" zu "Keys"**. IPA steigert Vorhersagegenauigkeit, erzeugt **Invarianz gegenüber Rotationen/Translationen** durch lokale Ausrichtungsinformationen anderer Frames. **3D-Struktur wird erstellt**, dann **Aminosäureseitenketten errechnet** und dargestellt.



- AlphaFold nutzt zwei Ansätze für **„Confidence Measures“** (**Vertrauenswürdigkeitsbestimmung**) der Proteinstrukturvorhersagen:
  „Predicted Alignment Error“ (PAE): quantifiziert Genauigkeit durch Fehler zwischen Strukturstellen, und
  „pLDDT“-Wert: gibt Genauigkeit für jeden Aminosäurerest an, basiert auf statistischen Modellen.







4. Bedeutung und Verwendung

- Durch neuronale Netzwerke wird der Proteinfaltungsprozess **in Minuten gemeistert** – ein **bahnbrechender Fortschritt für die Molekularbiologie**.
- Forscher erzeugen rasch Proteinstrukturen, was neue Einblicke ermöglicht.
- Dies eröffnet vielfältige Möglichkeiten, z.B. **gezielte Enzymherstellung** für bspw. Plastikabbau.
- AlphaFold optimiert **Elektronenmikroskop-Auflösung** mit **Atomkenntnissen**.
- Proteinrepräsentationen ausgestorbener Arten bieten **Stammbauminformationen**.
- Proteinstrukturanalyse via neuronale Netzwerke erweitert **Krankheitsforschungs- und Behandlungsoptionen**.

5. Leistung

- **Von 100 zu lösenden Proteinen wurden 70 mit experimenteller Präzision gelöst**.
- CASP 15 von Teams mit **AlphaFold-ähnlichen Softwares dominiert** (wegen open source).
- AlphaFold Leistung:
- Generiert **hochpräzise Proteinstrukturen** ohne **biologischen Kontext**
- Integriert **Bindestellen für Moleküle und komplexe Strukturen**, die sich aus Zusammenspiel mehrerer Sequenzen entfalten
- **Toleranz** gegenüber **fehlendem Kontext** (Bindenede Moleküle)
- Kann auch **Störungen** über **„Confidence Measures“** erkennen (Werte < 50)

* CASP (Critical Assessment of Protein Structure Prediction) ist ein wissenschaftlicher Wettbewerb, bei dem Teilnehmer Vorhersagen für Proteinstrukturen einreichen, die dann anhand experimenteller Daten bewertet werden, um die Genauigkeit und Leistung der Methoden zu bestimmen.