

# 技術動向サーベイ

- 作成者：熊谷 渉（O&M Design Gr., Project Design Div.）
- 担当日：2021年11月12日

## 対象論文

- タイトル : AlphaDesign: A *de novo* protein design framework based on AlphaFold [【URL】](#)
- 発行日 : 2021年10月12日
- 雑誌名 : Biorxiv (The preprint server for biology)
- 著者 : Michael Jendrusch, Jan O. Korbel, and S. Kashif Sadiq
- 所属 : Genome Biology Unit, European Molecular Biology Laboratory (EMBL)<sup>▲1</sup>, Germany
- 引用数 : 0
- キーワード : 計算機、*de novo*タンパク質設計、深層学習

## Abstract

*De novo* タンパク質設計は、合成生物学の長年の基本目標だが、アミノ酸配列から正確な高解像度のタンパク質構造を確実に予測することが困難であることが障害となっていた。近年、AlphaFold (AF) に代表されるタンパク質構造予測手法の精度が向上し、プロテオームスケール<sup>▲2</sup>での単量体タンパク質の構造予測が容易となった。本研究では、AlphaDesignを開発した。AlphaDesignは、最適化された設計プロセスの中にAFをオラクル<sup>▲3</sup>として組み込むことで、新規タンパク質設計のための計算フレームワークである。このフレームワークでは、ランダムな配列から新規のタンパク質モノマーを迅速に予測することができ、そのモノマーは既知のタンパク質空間内で多様なフォールドに適合することが示されている。また、AFを用いてタンパク質複合体の構造を予測することで、より高次の複合体を設計することが可能となった。単量体、ホモ二量体、ヘテロ二量体、高次のホモオリゴマー（三量体から六量体まで）についても予測している。さらに、事前に指定された標的タンパク質に結合するタンパク質を設計する可能性も示されている。予測された構造の整合性は、標準的な第一原理フォールディングや構造解析法、厳密な全原子分子動力学シミュレーションを実行し、対応する構造の柔軟性、モノマー内および界面のアミノ酸接触部を分析することによって、検証・確認されている。これらの分析により、構造的な整合性が広く維持されていることが示され、我々のフレームワークがかなり正確なタンパク質設計を可能にすることが示唆された。驚くべきことに、AFは、アミロイド纖維形成時の  $\alpha$ -ヘリックスから  $\beta$ -シートへの切り替えのように、複合体（四次構造）形成時にコンフォメーションを切り替えるタンパク質を予測する能力があることも明らかにした。また、我々の設計フレームワークに統合すると、モノマー状態とオリゴマー状態の間でコンフォメーションを切り替えるタンパク質のサブセットを新規に設計することができる。

## メモ

- 重合体の名称

eng	jpn	意味
monomer	モノマー	単量体
oligomer	オリゴマー	重合体（低分子）
dimer	ダイマー	二量体

eng	jpn	意味
polymer	ポリマー	重合体（高分子）
homo-	ホモ-	構成分子が同じ
hetero-	ヘテロ-	構成分子が異なる

- タンパク質の名称 アミロイド：タンパク質が間違って折りたたみ、凝集することによってできた不溶性の線維。例えば、アミロイド $\beta$ には、比較的高温の条件では、モノマーとして  $\alpha$ -ヘリックスを形成するが、温度低下により分子間会合し、 $\beta$ -シートを形成するものもある。

## 背景

---

- タンパク質構造予測（Folding問題）は、AlphaFold（AF）の登場により大きく進歩した。
  - Folding問題は、アミノ酸配列とタンパク質の立体構造・機能の関係を明らかにする上で、本分野で主要な課題であった。
  - 2018年CASPでAF1が優勝し、2020年CASPでAF2が優勝した。AFは深層学習ベースの手法で、データベースを学習することで結晶構造と同等の原子精度を達成した。AFは一本鎖構造の予測を目的としていたが、複合体の予測が可能なAlphaFold-Multimerもリリースされた。
- 一方、現実のFolding問題は立体構造予測だけでは済まないほど複雑であり、複数の安定状態が存在することがしばしばみられる。
  - 例えば、タンパク質は生理的な条件下では硬い構造体ではない。多くのタンパク質は、Folding前後で熱力学的平衡を示したり、構造変化を起こすことで機能を発揮する。さらに、立体構造が不安定で、一時的に無秩序にFoldingするタンパク質（IDPs、アミロイド遷移のMiss-Folding）や、Foldingの安定状態が複数ある変性タンパク質もあり、これらのFolding問題も生物医学的に重要である。
  - また、分子動力学（MD）シミュレーションなどの計算物理学的手法は、タンパク質間やタンパク質とリガンド間の結合だけでなく、コンフォメーションの遷移、第一原理Folding、無秩序遷移などの、ダイナミクス、熱力学、動力学を明らかにする方法を提供している。
- タンパク質のFoldingの基本的な原理を解き明かすることで、新規タンパク質の工学的設計が期待できる。
  - 従来のアプローチは、自然のタンパク質を反復的な実験的スクリーニングによって改変する方法（directed evolution）が中心であった。
  - 最近の計算機を用いた*De novo*タンパク質設計手法は、トポロジー選択のルール、タンパク質骨格の構築、配列の最適化など、一連の機能を持っている。

## 先行研究の課題

---

- Rosetta Remodelが計算機によるタンパク質設計の*De Facto*スタンダードであり、成果を挙げてきたが、Markov-Chain Monte Carlo（MCMC）ベースなので、計算量が多いことが課題。
  - Rosetta Remodelは、ターゲットタンパク質のトポロジー選択から、Foldingタンパク質の配列設計・検証まで、すべてのステップに対応するツールを組み合わせている。具体的には、トポロジーの指定、骨格の生成、固定骨格の設計というタスクで構成されている。Rosettaベースのタンパク質設計の一般的な戦略は、さまざまな設計問題に適用してきたが、構造空間におけるMCMCは、時間と計算量がかかる。
- この古典的なタンパク質設計の課題を解決するために、様々な設計問題にニューラルネットワーク（NN）を適用するアプローチが増えてきたが、構造情報を明示的に考慮していない、予測された骨格の設計可能性については保証されていないなどの課題があった。

- 配列アプローチ : UniProt配列データベース上で言語モデルや生成モデルを学習し、目的の機能を持つタンパク質配列を直接生成する研究がいくつかあり、与えられた機能に関連するタンパク質配列の生成には成功している。しかし、構造情報を明示的に考慮していないため、三次構造に制約があるタンパク質設計タスクには適用できない。
  - 構造アプローチ : 生成モデルで骨格の距離マップや座標を学習し、タンパク質構造を生成する研究があり、潜在変数を操作することで、設計された骨格構造を細かく制御できる。しかし、予測された骨格の設計可能性については保証されていない。
- また、構造アプローチと配列アプローチのギャップを埋めるために、PDBのタンパク質構造を用いてNNを学習させ、固定された骨格構造が与えられたときのタンパク質の配列を予測する研究もある。しかし、これらのアプローチは、特定のタンパク質設計タスク（その骨格形状を考慮に入れる）のために学習したネットワークが必要となり、再学習無しで他の設計アプリケーションに容易に拡張することができない。
  - 学習したタンパク質の構造や機能の予測値を、*in silico*スクリーニングの枠組みの一部として再利用しようとする研究が数多くある。これらのアプローチでは、NNをスコア関数として扱ってタンパク質設計の品質を評価し、勾配ベース、直接探索法、NNベースの最適化を用いて設計案を改善させていく。
    - trRosetta + NN : trRosettaを構造予測に使用し、NNを埋め込んだ最適化ループと組み合わせて、固定骨格設計、タンパク質モチーフを安定化させるためのタンパク質足場生成に適用されている研究がある。
    - AF + 貪欲法 : AFを構造予測に使用し、学習済みモデルから初期配列を用いた貪欲法による固定骨格タンパク質設計に使用している研究もある。

## 本研究の特徴

- AF、Rosetta、MDシミュレーションなど、あらゆる最先端の計算技術を組み合わせた設計ループを確立したこと、探索範囲を広げつつも、正確かつ現実的かつ新規のタンパク質が設計可能となった。
  - AFを構造予測オラクルとして設計ループに組み込むことで、構造上設計可能で、AFの下で高い信頼性を持つ骨格と配列を生成する。
  - AFを用いた最適化のために、様々なタンパク質設計タスクをコード化する、柔軟なターゲット関数群を定義することで、構造予測器を用いたタンパク質設計の先行アプローチを拡張する。
  - この設計ループとRosettaやMDシミュレーションと統合することで、可能な設計タスクの範囲を拡大する。

## 成果

- 新規にタンパク質のモノマー、ダイマー、オリゴマーを設計した。さらに、標的タンパク質の配列のみを用いてその結合体や、複合体形成時にコンフォメーションを変化させるタンパク質を設計した。

## Ideas & Findings

### De novo 設計フレームワーク

#### 特徴

- タンパク質設計を、ある目標機能が一定の閾値を超えるようなタンパク質配列の集合を見つける探索問題として捉える。
- 配列の特性とタンパク質の全原子構造の両方を考慮するために、AFをターゲット機能に統合し、高品質な構造予測と予測の信頼性の測定を行う。
- また、Rosettaによる第一原理構造予測とMDシミュレーションを用いた最先端の検証を行う。

## 設計ループ

- Fig.1 A : タンパク質の配列をAFに入力し、AFからその配列に対応する、全原子構造と予測された信頼度を得る（入力情報）。
  - AFの信頼度は、予測局所距離差検定（pLDDT）と予測整列誤差（PAE）の組み合わせで表される。
  - pLDDTは、局所的なモデルの品質を測定し、PAEは、各アミノ酸ペアの信頼度を測定する。
- Fig.1 B : AFの出力から任意の目的関数Lを最大化するように配列を最適化します（出力情報）。
  - 最適化問題には、ターゲット構造に対するRMSDの最小化や、AF予測の信頼性の最大化などが含まれる。
  - 目的関数には、オリゴマーの状態やタンパク質の結合時のコンフォメーション変化を制約することもできる。
  - 一般的に、タンパク質の構造、配列、予測信頼度に関するあらゆる機能を最適化に使用することができる。
- Fig.1 C : シンプルなEvolutionary Algorithmを用い、次の配列をサンプリングする（最適化ループ）。
  - 配列プールの各配列について、対象となる関数の値を計算する。配列はさらに組み換えや変異を行い、配列空間を探索する。配列プールは、変異した配列で更新される。
  - 他の勾配のない、あるいは勾配に基づく最適化手法は、必要に応じて代用できる。
  - 最適化の過程で、タンパク質の構造が変化し、ローカルな信頼度とグローバルな信頼度の両方が増加する。
- Fig.1 "Validation" : 閾値以上の配列を返し、そのうちのサブセットをMDシミュレーションとRosettaの第一原理構造予測を用いて検証する。
  - あるタンパク質設計タスクでは、ある配列が完全に最適化されたとみなされる閾値が設定される。

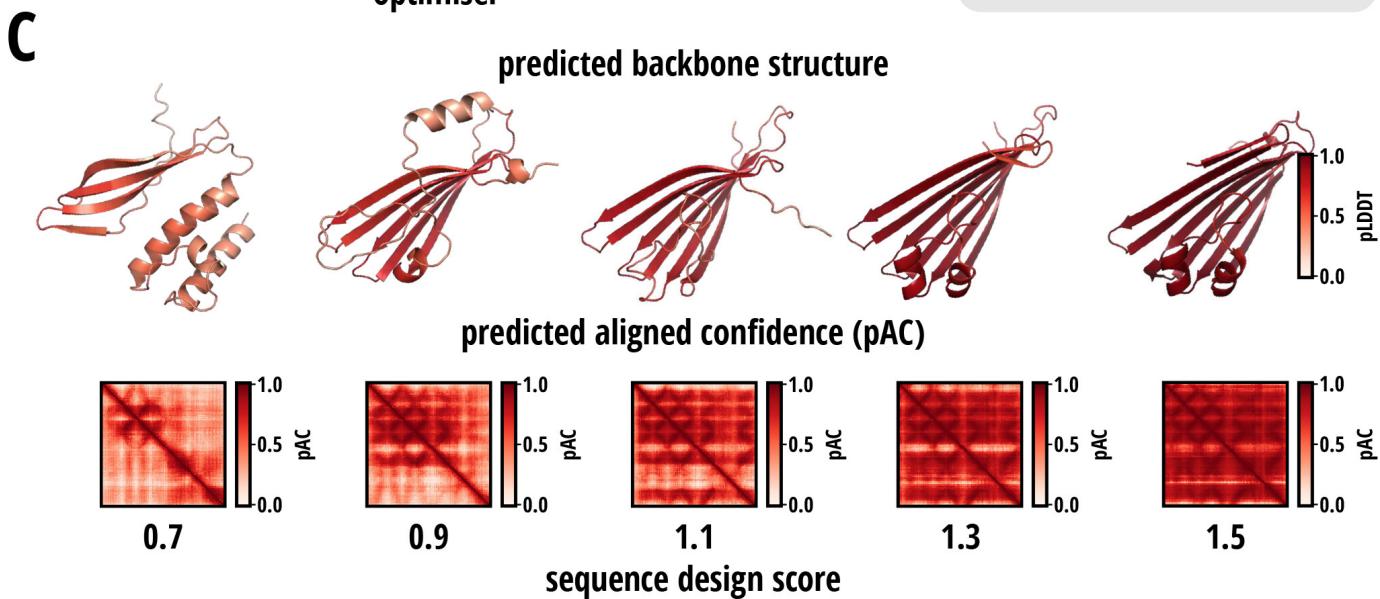
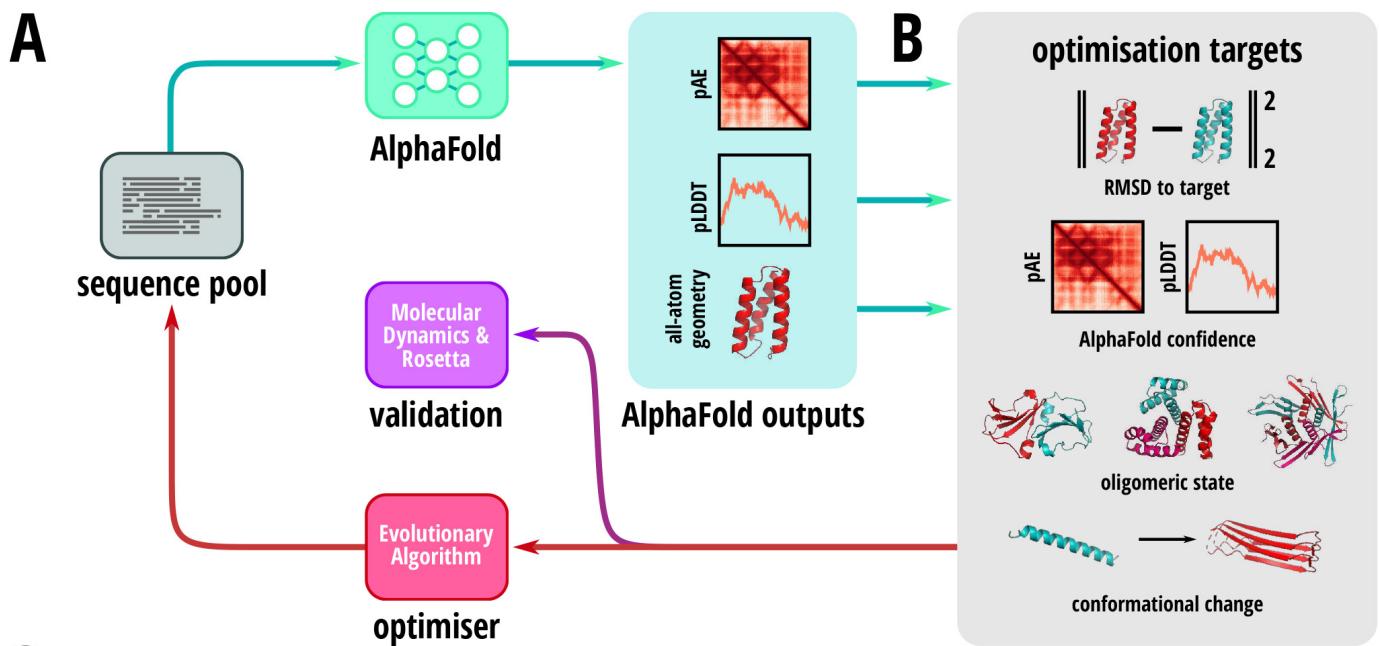


Fig.2 : AlphaFold prediction of protein complexes

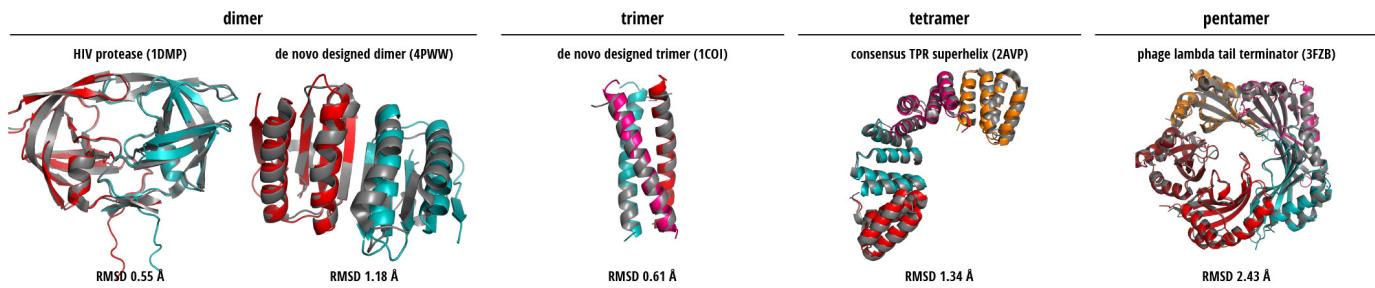


Fig.3 : De novo monomer design

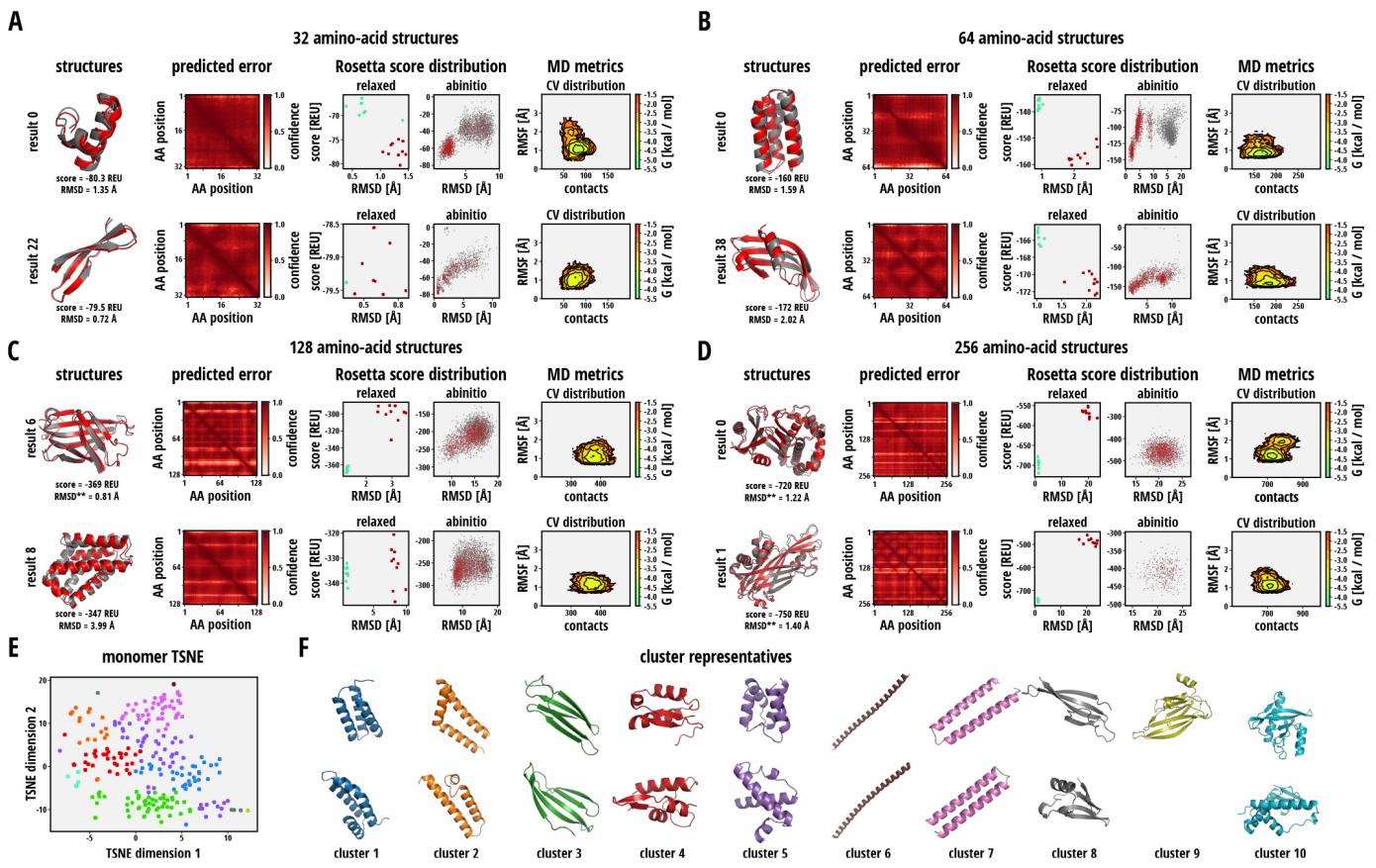


Fig.4 : De novo dimer design

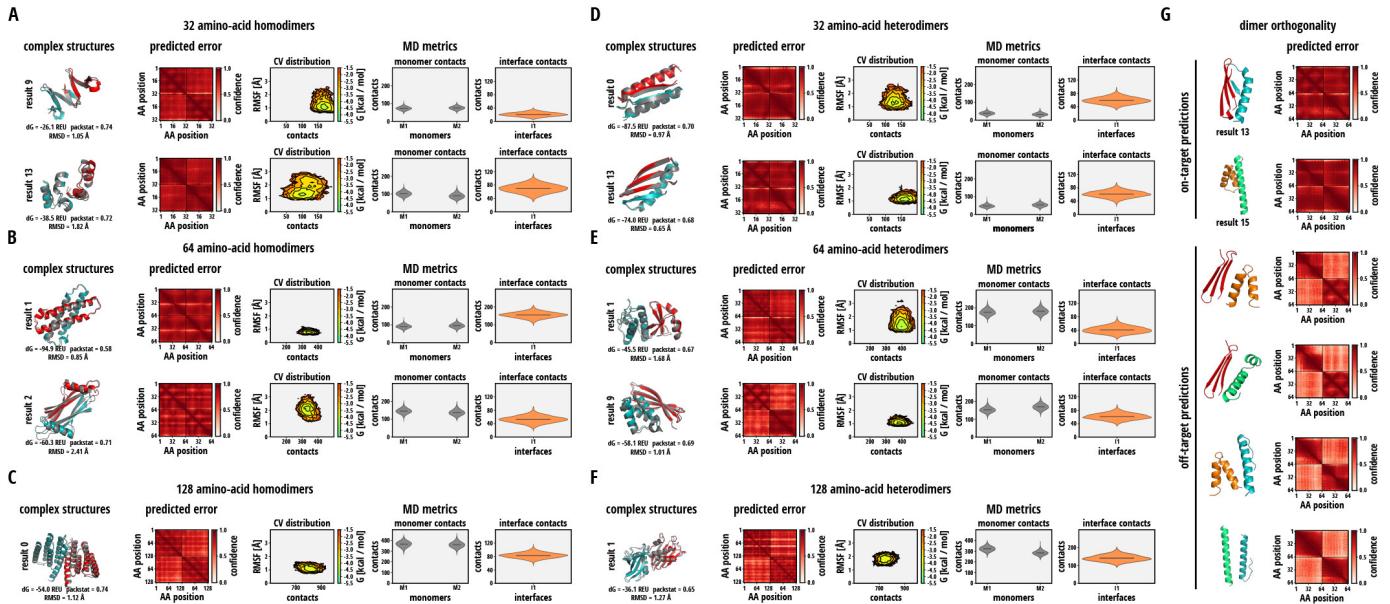


Fig.5 : De novo oligomer design

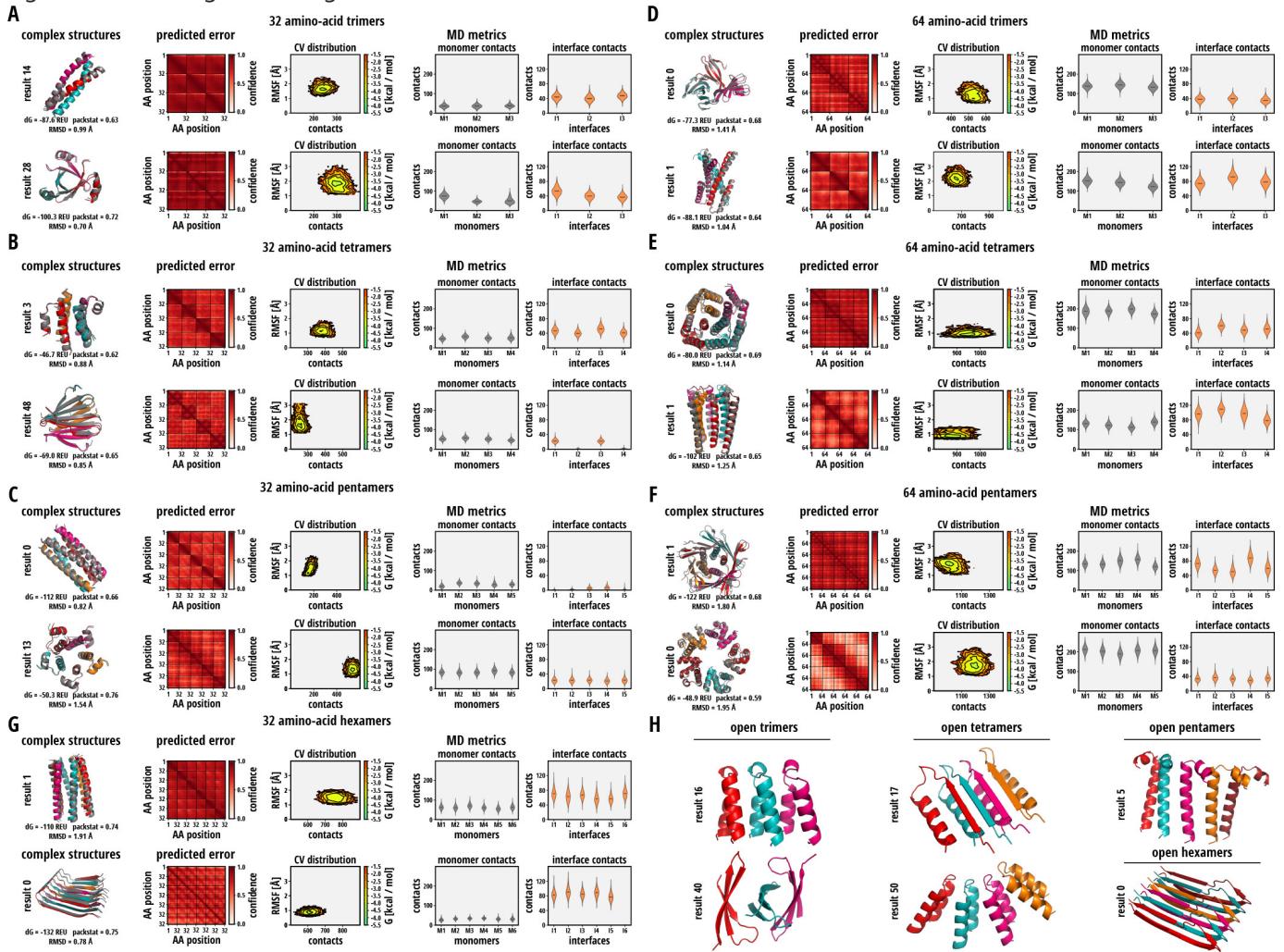


Fig.6 : De novo binder design

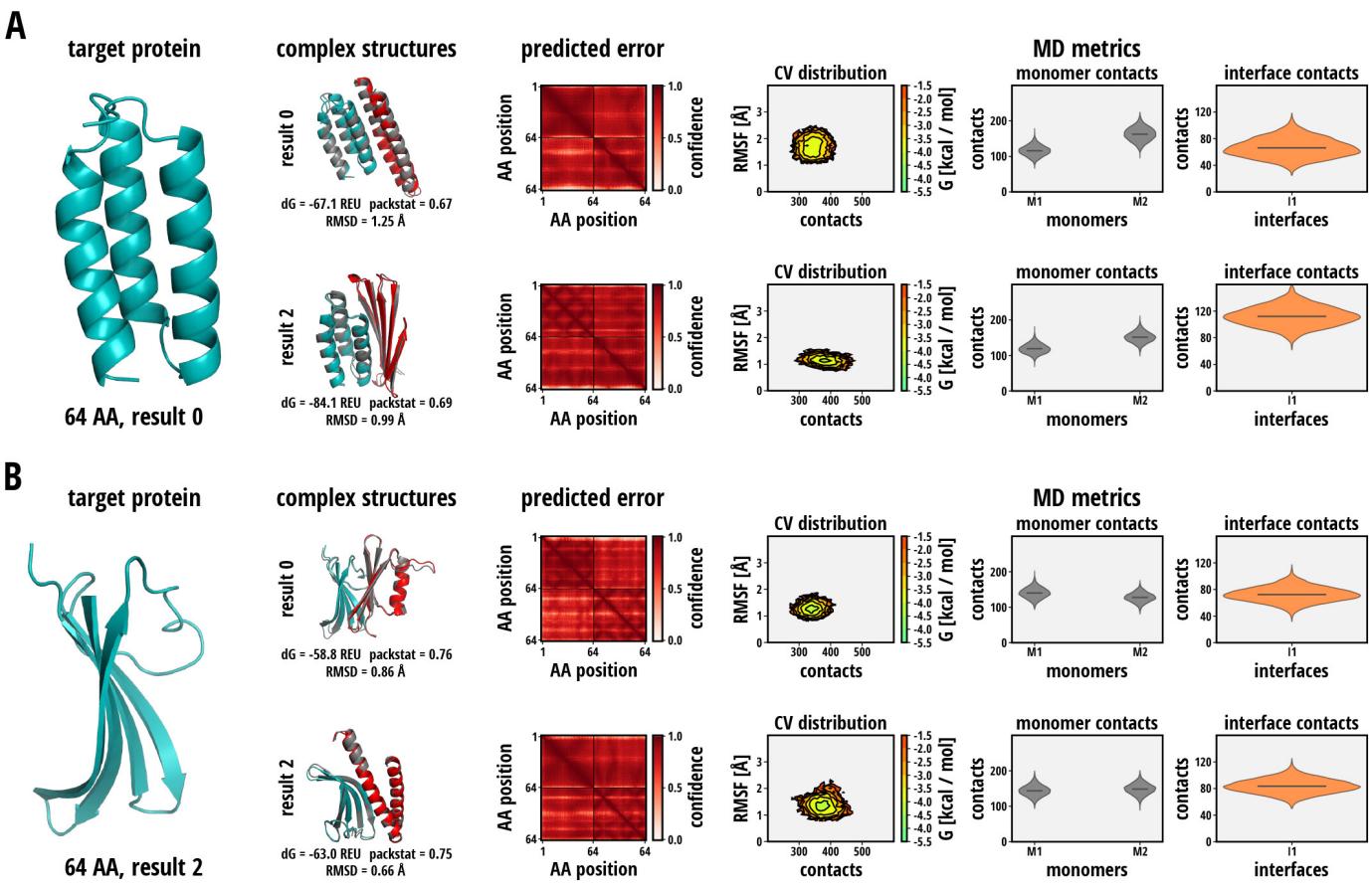
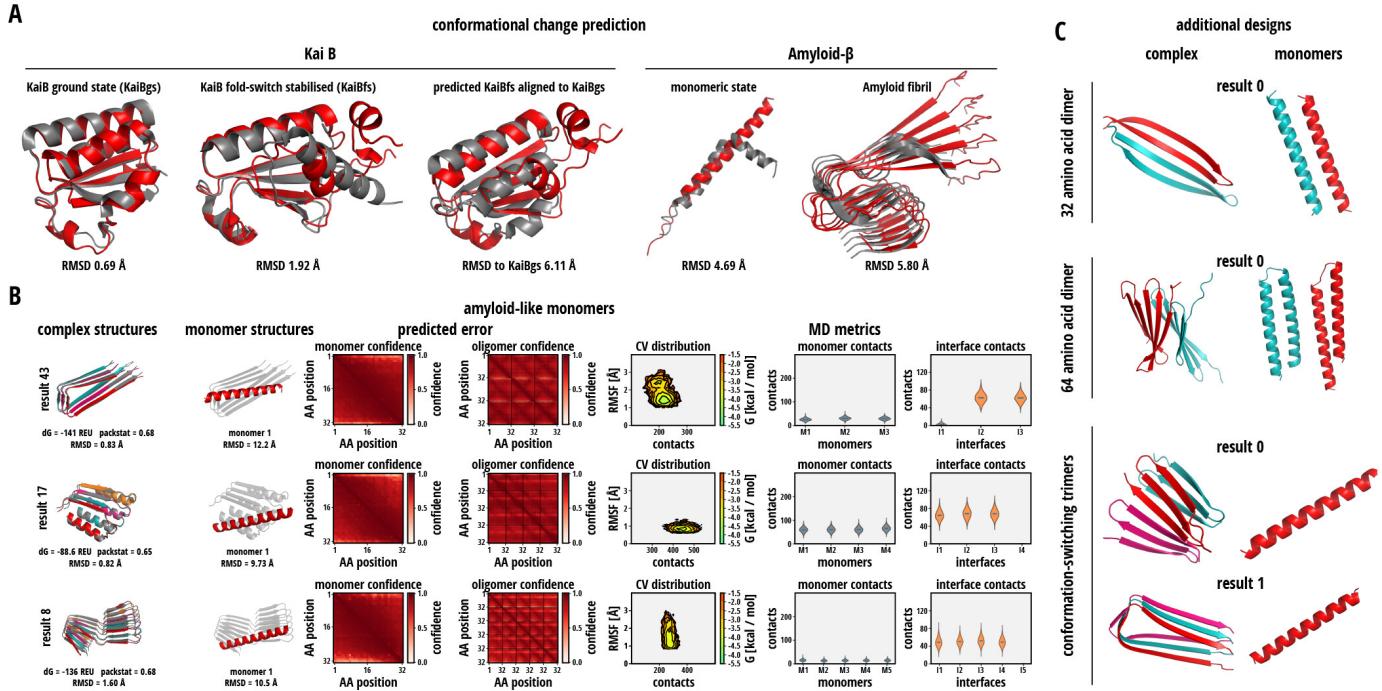


Fig.7 : Conformational change in AlphaFold sequence space



## 疑問点

## 参考になる文献

