**研究成果報告書：AlphaFoldによる変異体評価 人工酵素設計PJT**



**Confidential**

目次

1. **はじめに2**

1.1　本書の位置づけ2

1.2　研究の背景2

1. **理論、システム構成2**

2.1　最適化問題2

2.2　最適化手法3

1. 実験結果**4**

3.1　変異体の獲得4

3.2　変異体の分析5

1. まとめ**6**
2. 参考文献**6**
3. 付録**6**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 2022/10/24 | Description | 新規作成 | | | | | | |
| Issue | 横河電機 MKHQ INVC | | | Author | 熊谷 | Check | 生田目 | Approval | 生田目 |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

1. **はじめに**
   1. **本書の位置づけ**

本書は、人工酵素設計PJT（2020年3月～2022年10月）の要素技術の一つである「AlphaFoldによる変異体評価」の調査研究の概要・実施結果について報告を行うものである。

* 1. **研究の背景**

本PJTでは、酵素の産業応用上の課題として、特許発明に基づく酵素利用制限や、天然物の機能利用制限だと捉え、その解決のために、酵素の探索だけでなく酵素の人工設計が鍵となると考えた。さらに、有用な酵素を人工的に設計する上で、天然酵素のアミノ酸残基の改変によって変異体を生成するアプローチが期待できる。本PJTでは、このアプローチに従って、天然酵素の特徴を継承しながら、アミノ酸配列の類似度が一定以下の新規の変異体を生成することを目標としていた。

そこで、本PJTでは、変異数を制限しながら天然タンパク質（Wild-Type）の変異体をサンプリングする「変異体探索」技術を開発し、実際に数万個のセルロース結合性タンパク質（CBD）の変異体（最大5点変異）を机上で生成した。一方、一般的に、Wet実験を通じて評価するとき、全ての候補を対象とするのは非効率であるため、候補から有望な変異体を選定する「計算機スクリーニング」が実施される。本研究においても、セルロース結合性評価実験に向けた選定のために、CBD変異体候補の計算機スクリーニングが必要である。

結合性タンパク質の一般的な計算機スクリーニングとしては、ドッキングシミュレーションや分子動力学シミュレーションなどの分子シミュレーションを用いた机上評価が代表的である。結合性タンパク質は特定の基質（セルロース）を認識・相互作用することでその機能を発現し、このメカニズムは結合部位付近の立体構造や動的な振る舞いが密接に関係している。ドッキングシミュレーションは、基質と結合部位（ポケット）の形状の関係性から、結合親和性スコアや結合ポーズをシミュレートする。分子動力学シミュレーションは、計算機上で水溶液中の基質と結合性タンパク質を仮定し、原子間の相互作用が与えられた力場に基づいて物理的な振る舞いをシミュレートする。このため、分子シミュレーションは計算機上で結合性タンパク質候補の良し悪しを評価するのに不可欠であるといえる。

図1にタンパク質のアミノ酸配列と立体構造の関係を示す。アミノ酸配列は、自由エネルギーが最小となる（安定化する）ように折り畳まれていき立体構造（天然状態）を形成するが、それが安定しない場合は立体構造を形成せずに変性状態となる。つまり、立体構造形成が可能なアミノ酸配列のパターンは極一部に限定されている。配列から立体構造を形成することをFoldingと呼ぶことに対して、指定の立体構造を実現可能なアミノ酸配列を特定することをInverse-Folding Problemと呼ぶ。Inverse-Folding Problemは新規タンパク質設計（De Novo Design）において重要な課題と言われている。CBD変異体設計においても、天然タンパク質から1割以上の残基を変異させる場合、配列空間の自由度が広がるため、Inverse-Folding Problemが課題となると考えられる。また、本研究の変異体生成では、Rosettaのエネルギースコアが最小とする最適化問題を解いているが、あくまで計算機上のエネルギー関数を用いているに過ぎず、立体構造へ折り畳む妥当性「Fold可能性（Foldability）」を十分保証しているとは限らない。同様に、分子シミュレーションもFoldabilityを直接評価していない。このため、Foldabilityに焦点を当てた変異体の評価が可能ならば、別の観点によるCBD変異体の計算機スクリーニングが期待できる。

一方、AlphaFold2[1]は、2021年に登場した、アミノ酸配列からその立体構造を高精度で予測する技術である。具体的には、PDBデータを用いて、予測構造の信頼度スコアを計算し、それを高めるように学習する。AlphaFold2の信頼度スコアは、熱・統計力学に基づくものではなく、学習データに基づいて統計的な観点で計算されるため、Rosettaや分子シミュレーションのスコアとはある程度独立的な観点で、アミノ酸配列のFoldabilityを評価できる。

しかしながら、AlphaFold2はPDBに登録されている天然タンパク質を中心に学習・適用されており、設計した変異体への適用例は未だに少ない。ペプチド配列設計に関する研究成果において、アミノ酸の溶解度指標以外に、AlphaFold2の信頼度スコアも同時に最大化することで、標的タンパク質に対する結合親和性が高く、かつ脂溶性が高すぎない候補を効率良く生成できた事例[2]があり、候補設計における信頼度スコアの有用性が期待できるが、CBD変異体に対する事例は確認されていない。以上から、本研究ではトライアルとして、本PJTのCBD変異体について、信頼度スコアに基づく計算機スクリーニングの可能性について簡易的に検証した。

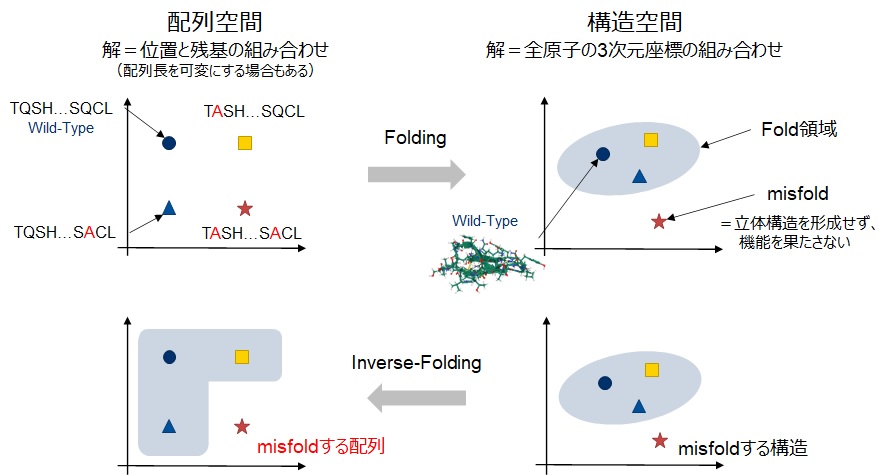


図1：Inverse-Folding ProblemとFoldability

1. **理論、システム構成**
   1. **AlphaFold2**

AlphaFold2の概要を簡潔に説明する。AlphaFold2は、2020年のタンパク質構造予測コンペティション（CASP）[4]は類似度が一定以下の変異体生成のために検討した技術について説明する。図1に今回検討した技術の概要を示す。指定した変異数内の変異位置・残基の組を、上位最適化の最適化変数に設定し、目的関数値であるエネルギースコア（Rosettaスコア）を最小化するような、最適化問題を構築した。この最適化問題を全探索やランダム探索よりも効率的なアルゴリズムで解くことで、類似度が一定以下かつエネルギースコアが良い変異体を得ることが期待できる。予測構造の信頼度スコア（predicted local distance difference test；pLDDT）

以降は、最適化問題と最適化手法について説明する。構築した組合せ最適化問題は式として定式化される。

ただし、は目的関数、は離散的な実行可能解の集合、は組合せ的な構造を有する最適化変数（解）である。本研究では、アミノ酸残基の20種・配列長から構成される各アミノ酸配列が解である。また、は、任意のアミノ酸配列が、基準とする天然アミノ酸配列と異なる残基数を出力する関数であるため、式の制約条件は、基準とする天然アミノ酸配列から変異以下のアミノ酸配列に制限することを意味する。したがって、全通りは通りだが、制約条件を満たす組合せのみで構成したのが探索空間である。また、目的関数は、基準とする天然アミノ酸配列が任意のアミノ酸配列に変異したとき、Cartesian DDGで計算したときに得るエネルギースコアである。

　なお、天然配列からの変異に着目した表現として、変異位置と残基の組で表現する方法がある。例えば、2番目の残基をセリンSに置換、10番目の残基をアラニンAに置換する変異はと表現できる。以降ではこの変異表現を解の代わりに使用する。

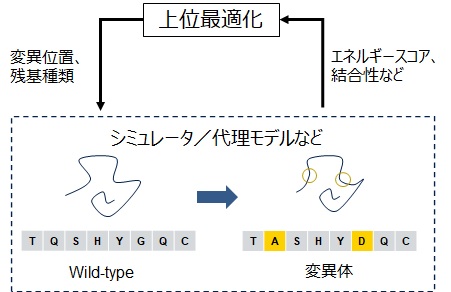


図1：変異体探索プロトコルの概要

* 1. **最適化手法**

図2に、配列空間における変異体探索のイメージを示す。最適化手法はLocal Searchを使用する。Local Searchは、単点探索アルゴリズムで、探索空間（配列空間）内の暫定解の近傍を生成し、その近傍内の目的関数値が改善する解に移動していく。近傍内の改善位置にしか移動しないため、初期位置に依存して、いずれかの局所解に収束する。本研究では、近傍で最も目的関数値が優れた解を移動先として選択するbest改善を使用した。初期解は一様乱数で与え、複数並列のLocal Searchを適用した。

近傍と問題規模について説明する。本研究で使用した近傍は、Hamming距離が1以内かつ変異数が以下を満たす解の集合とした。これにより、近傍解は制約条件を満たす解に限定することを保証する。このとき、近傍のサイズ（サンプリング数）は最適化問題の次元数に一致することが知られており、摂動を加える解自身を除くと、となる。例えば、配列長36、最大4点変異の場合、次元である。なお、反復回数、並列数のとき、探索過程で全サンプル数は個、制約を満たす範囲の全通りは通りである。実際の近傍生成では、上記に加えて実応用上好ましくない解を除外することで、問題規模を削減することが可能である。

　本研究では、TrCel7A-CBM1（セルロース結合性タンパク質）である1CBHを対象とした。1CBHの配列長は36残基である。さらに、変異位置候補から、立体構造中のジスルフィド結合を形成しているシステインCYSの8,19,25,35番目と、結合部位に該当する5,29,31,32,34番目を除外した。表1に、これらを前提とし、最大2点・3点・4点・5点変異の場合の配列空間の問題規模を示す。次元数、全通り、探索で得られる全サンプル数（並列数は10、反復回数30）は、それぞれ、、となる。変異数に対して全通りは指数増加する一方で、サンプル数は線形増加に抑えている。したがって、変異体探索は全探索よりもリーズナブルにサンプリング可能である。

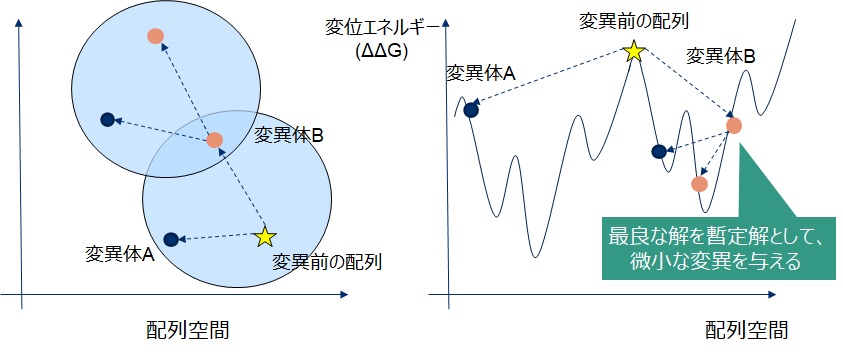


図2：配列空間における変異体探索のイメージ

〇〇項で検討した、結合能簡易評価系によって評価した変異体100種類（3点変異、4点変異、5点変異）のアミノ酸配列について、AlphaFold2（AF2）を適用し、その予測構造と各原子位置のpLDDTを計算した。AlphaFold2は、Webブラウザの計算環境（Google Colaboratory）で実行可能なColabFold[2]を使用した。

1. **実験結果**
   1. **変異体の獲得**

　1CBHのWild-TypeをRosettaのFastRelaxで構造緩和したPDBデータを初期構造として、変異体探索を実行した。探索条件は2.2節に示した通りである。図3に、最大3点・4点・5点変異探索の過程で得た、全変異体のスコア頻度分布を示す。図3の横軸はCartesian ddgスコア、縦軸は頻度である。Cartesian ddgスコアはRosettaスコアと線形関係であり、低ければスコアが良いことを示す。変異数を増やしていくと、スコアが改善していく様子が確認できる。また、変異体探索によって、エネルギースコアの下限付近を十分にサンプリングできていると思われることから、全探索せずにスコアを大きく改善する変異体が得られると期待できる。

また、図4に、最大3点変異探索で得た、変異体の立体構造例を示す。3点変異体のアミノ酸残基に変異した部位を紫色で示している。Wild-TypeのRosettaスコアは-127.9だが、3点変異体のRosettaスコアは-157.8であるため、全探索せずにエネルギースコアを大きく改善した変異体が得られていることが確認できる。

図3に、4つの3点変異体の予測構造とpLDDTを示す。緑色の構造がRosettaで生成した構造、白色の構造がAF2で予測した構造、水色が結合部位、紫色がWild-Typeから変異した部分である。予測構造の下の図は、各原子位置のpLDDTを色付けしており、青色が濃いほど信頼度が高いことを表す。図3から下記のことが確認できる。

* global pLDDT（全原子のpLDDTの平均値）は90%以上であるため、CBD変異体についてもAF2の予測信頼度は極めて高い。
* 各変異体のC-α原子のRMSDは0.5Åであるため、AF2予測構造は、Wild-Typeの構造やエネルギースコアを使用せずとも、Rosetta構造と似た結果を示している。
* 一方、変異体によっては、構造の一部のpLDDTが薄い水色（80%代）であるため、予測信頼度が低下する部位もある。

さらに、図4に、100個の変異体のスコアとglobal pLDDTの散布図を示す。左図の横軸がRosettaのエネルギー変位スコア（ΔΔGスコア）、右図の横軸がAutoDock VinaのAffinity（結合親和性）である。各データの色は、〇〇項で評価した、簡易評価によるセルロース結合能の強さで、赤色が強い、緑色が弱い、青色が無いことを示している。図4から下記のことが確認できる。

* 左図から、ΔΔGとpLDDTの間には関係性がみられない。
* 一方、右図から、Affinityが低く、pLDDTが高い領域に、結合能が強い変異体が多く位置する。

よって、Wet実験に移行する前に、机上でAffinity以外にもpLDDTを使用することで、CBD変異体候補のスクリーニング効率が改善する可能性を示唆している。これは、文献[3]の事例と通じる結果である。一方、今回の変異体の中で、結合性が強かった変異体は全て三点変異であり、四点・五点変異体は全て結合性が弱い、つまり変異数の影響を受けて、結果が偏っている可能性がある。したがって、CBD候補設計における信頼度スコアの有用性を詳細に検証するためには、より多様なサンプルに対して評価する必要がある。



図3：変異体の予測構造と信頼度スコア（三点変異体の例）

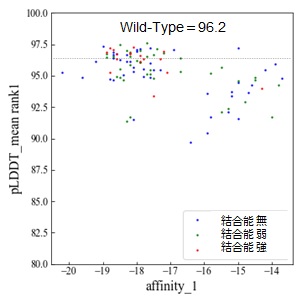
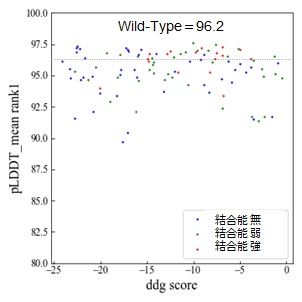


図4：各変異体のスコアと信頼度スコアの散布図（左図：ΔΔGスコア、右図：結合親和性）

* 1. **変異体の分析**

変異表現の偏りを分析した。図5に、3.1節で得た全変異体の単位変異表現の頻度分布を示す。単位変異表現とは、1単位の変異表現のことで、例えば、3点変異体は、の3種類の単位変異表現を有しているとみなす。図5の横軸は、アミノ酸配列の残基位置、縦軸は変異した残基種類（図の上に文字列がWild-Typeの配列）である。青色が強いほど、その単位変異表現の頻度が高いことを表す。図中に直接記載されている単位変異表現は特に頻度が高い表現で、中でも赤い表現はスコア上位が変異体に含まれている単位変異表現である。なお、変異表現が一致するサンプルは、最良スコアのサンプル以外は除外した。図5から、いずれの結果においても、高頻度の単位変異表現は、一部に偏りを持っており、その中に優れた変異体に共通していることが確認できる。

また、図6は、図5中の高頻度な単位変異表現に注目し、変異前後の残基を整理したものである。なお、正負の欄は、1点変異におけるΔΔGスコアの正負であり、正の場合は悪い変異であったことを表す。図6から確認できることは下記の通りである。

* 3点以上の変異体に含まれる高頻度な単位変異表現は、1点変異においてもエネルギースコアが良い
* 親水性かつ荷電無しのアミノ酸残基（赤色）を疎水性残基（水色）に置換するケースが多い

特に後者の結果は、エネルギースコアを最適化しただけにも関わらず、変異のパターンが偏っている点から興味深い。熱・統計力学的な観点から、エネルギー（タンパク質の立体構造安定性）は残基の親水性／疎水性や荷電の有無と関係することが知られている。例えば、タンパク質の疎水性残基は疎水基同士で集合し、内部に位置するほうが安定化する（疎水効果）などが挙げられる。しかしながら、今回の結果から、エネルギーと変異パターンに対して明示的な解釈を与えることはできなかった。

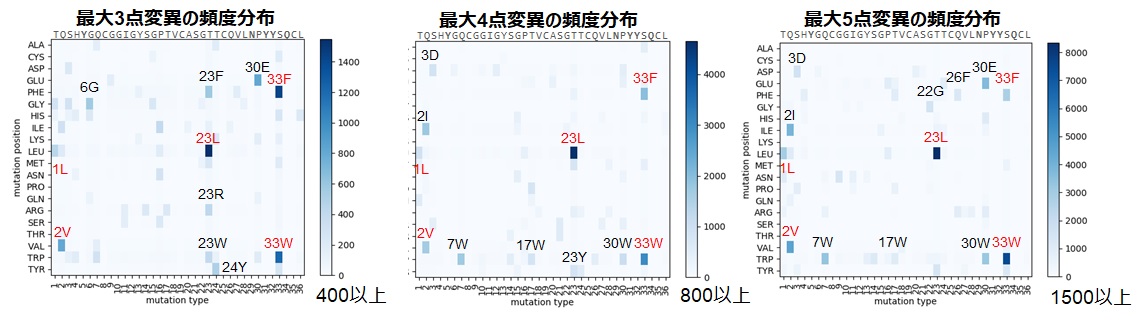


図5：最大3点・4点・5点変異体の単位変異表現の頻度分布

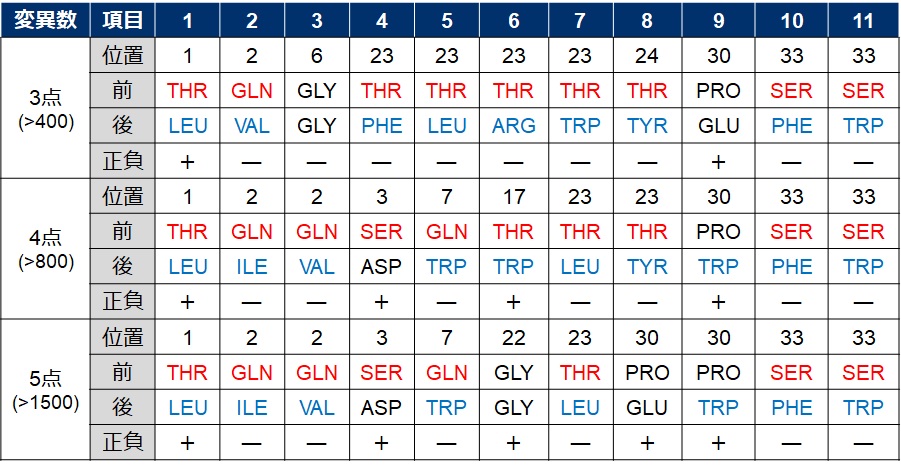


図6：高頻度な単位変異表現（正負は1点変異におけるΔΔGの正負を表す）

1. **まとめ**

本書では、人工酵素設計PJTの要素技術の一つである「変異体探索」の調査研究と実施結果について報告した。本研究の変異体探索は、変異数制約を満たすWild-Typeの変異体候補を探索・生成する方法である。数値実験を通じて、変異体探索が全探索・ランダム探索よりも効率的に優れたサンプルを獲得することが可能であることを示した。今回対象とした1CBH （36配列長）の4点変異体は、Wild-Typeの残基と全体の10%が異なることから、人工酵素設計に貢献することが期待できる。一方、変異体の残基傾向とスコアについては十分な考察ができなかったが、今後詳細について考察する場合は、変異位置が立体構造上の表面／内部に位置するかなどを含めて、総合的に分析することが必要だと考えられる。

1. **参考文献**
2. J. Jumper et al.: “Highly accurate protein structure prediction with AlphaFold”, Nature (2021)
3. T. Kosugi et al.:“Solubility-Aware Protein Binding Peptide Design Using AlphaFold”, Biomedicines, Vol.10, No.1626 (2022)
4. M. Mirdita et al.:“ColabFold: Making protein folding accessible to all”, Nature (2021)
5. https://predictioncenter.org (2022.10.28, Accessed)
6. <https://github.com/sokrypton/ColabFold> (2022.10.28, Accessed)
7. <https://doi.org/10.5281/zenodo.6683680> (2022.10.28, Accessed)
8. <https://github.com/YoshitakaMo/localcolabfold> (2022.10.28, Accessed)
9. **付録**
   1. **AlphaFold2のバージョン**

AlphaFold2は、オリジナルのバージョン以外に、いくつか派生版が提供されている[4]。図A1にバージョンを整理した表を示す。

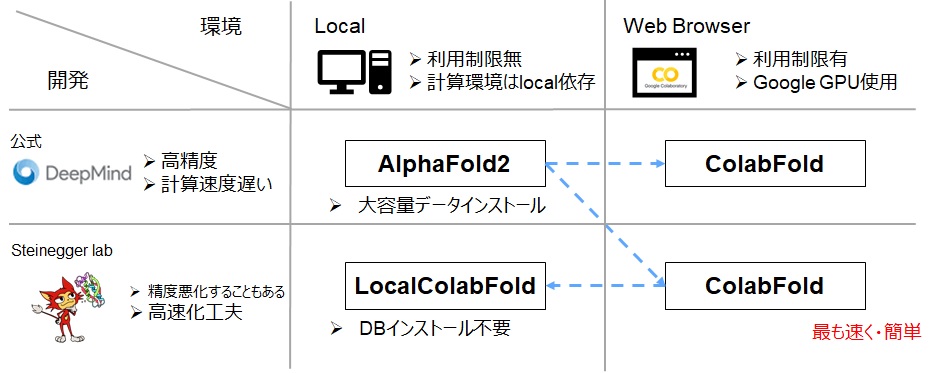
まず、ローカル版とWebブラウザ版に大別される。ローカル版は、Githubからローカル環境にインストールして計算するバージョンである。オリジナルは、DeepMind社が提供している学習済のモデルで、最新版がVersion 2.2.4（2022年10月28日時点）である。なお、Version2.2以上は複合体への適用が可能である。また、OpenFold[5]と呼ばれるローカル環境での学習が可能なモデルも存在している。

一方、Webブラウザ版は、Webブラウザ（Google Colaboratory）上でアミノ酸配列と条件を設定するだけで、構造予測が実行可能なバージョンである。Webブラウザ版は、オリジナルのAlphaFold2に準拠したモデルとColabFold[3]がある。ColabFoldは、Steineggerを中心とする有志の研究者が計算高速化のために開発・公開したバージョンで、一般ユーザがAlphaFold2を積極的に使用するきっかけとなった。

さらに、Webブラウザ版はGoogle Colaboratoryでの計算は利用制限がある、オリジナルはローカル環境へのインストールに巨大なディスク・メモリ容量を必要とするなどの問題があるため、LocalColabFold[6]は、オリジナルよりも軽い容量でローカル環境にインストール可能で、ColabFoldと同様に高速に計算が可能である。表A1にこれらのバージョンのスペックを示す。ColabFoldはオリジナルよりも予測精度が多少低下するものの、高速かつインストール不要で計算できることが大きな特徴である。多量体のように、ColabFoldでも計算時間が長期化することが見込まれる場合は、LocalColabFoldが適切だと考えられる。

表A1：AlphaFold2のスペック比較

|  |  |  |  |
| --- | --- | --- | --- |
| 項目 | オリジナル | ColabFold | LocalColabFold |
| 予測精度 | 高い | オリジナルよりも低い | オリジナルよりも低い |
| 計算速度 | 遅い | 速い | 速い |
| 利用制限 | 無し | 有り | 無し |
| ディスク容量 | 2.5TB | 無し（ブラウザ上） | 3.5GB |
| メモリ容量 | 32GB以上 | 無し | 無し |



図A1：AlphaFold2の派生版