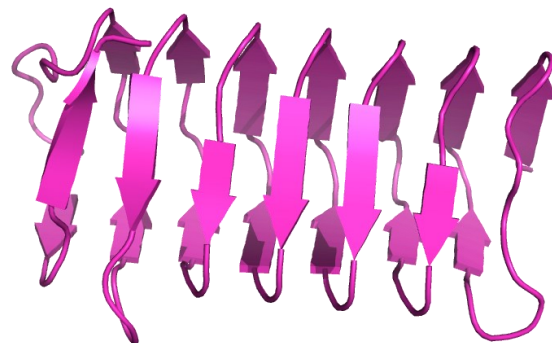


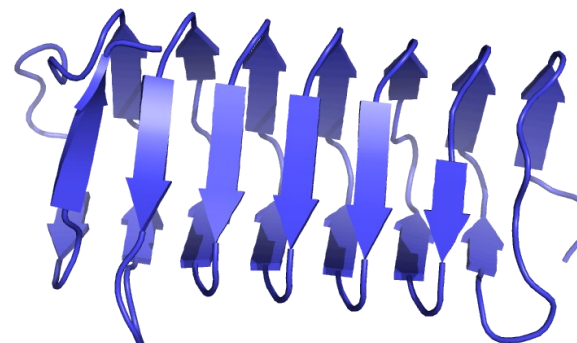
AlphaFold2 と ProteinMPNN を使用した 蛍光タンパク質の作成

AlphaFold2

タンパク質構造予測のためのディープラーニングモデル



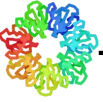
Experimental Structure

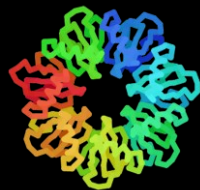


AlphaFold2 Structure

結晶構造解析・NMR構造解析などの従来法と比較して安価で高速

正確かつ結果には信頼性の指標があり予測が正確かどうかを推測するのに役立つ

- 
- 大量のVRAMを搭載したハイエンドグラフィック カードなどの高価なハードウェアが必要
 - コンピューターに適切にセットアップするには技術的な専門知識が必要
 - メンテナンスが面倒
 - 複数のタンパク質からなる複合体やタンパクと小分子（DNAやRNA）複合体、動的な構造変化など苦手なものがある



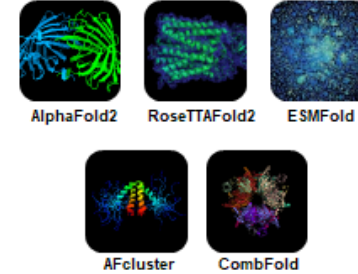
Neurosnap Platformで解決！

AlphaFold2の他、
30種類以上のAI予測ツールを搭載

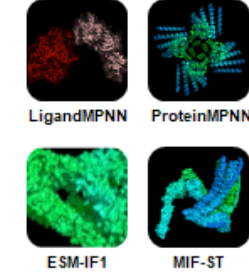
- タンパク質のフォールディング
- 分子ドッキング
- 創薬/ドラッグディスカバリー
- タンパク質工学
- In silico Mutagenesis

など

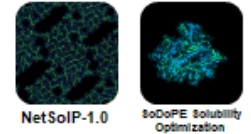
Protein Folding



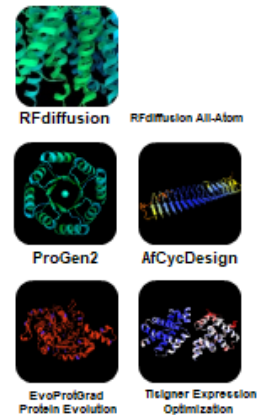
Inverse Folding



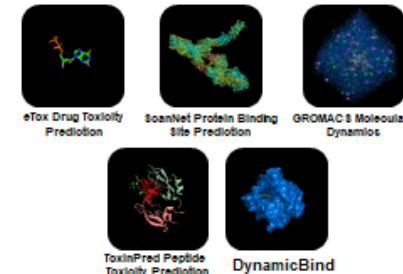
Protein Solubility



Protein Design



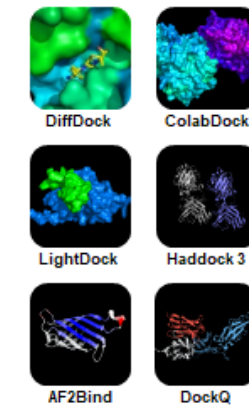
Drug Design



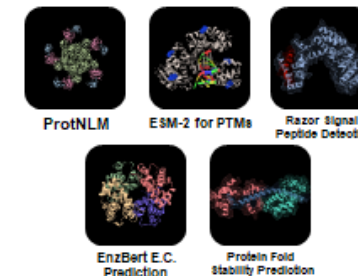
Protein Localization



Molecular Docking



Protein Annotation



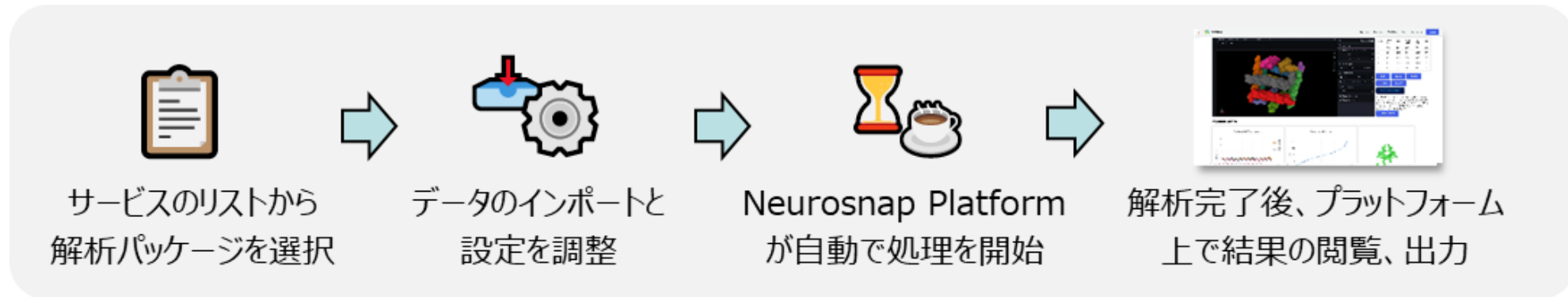
Protein Clustering



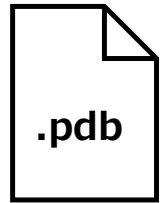
Multiple Sequence Alignment



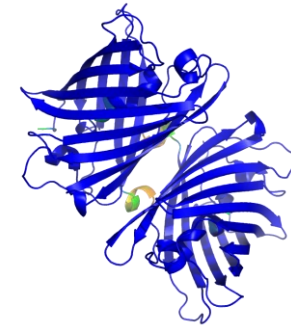
- 解析がパイプライン化



- ゼロコードで解析
- クラウド上で高速計算



Rank	Score	Sequence Identity	Sequence
1	0.593	0.585	LFREKIPYVVEMEGDVEGMI
2	0.596	0.599	LFREKIPYVVEMEGDVEGMI
3	0.603	0.603	LFREKIPYVVEMEGDVEGMI



データベースから
pdbファイルをダウンロード

ProteinMPNN
配列生成し変異体を作成

AlphaFold2
立体構造を予測



今回は蛍光タンパク質データベースFPbase.orgから、最も明るいGFPであるAequorea australis由来のAausFP1を選択し、ファイルをPDBからダウンロードしました。

ProteinMPNN

入力したタンパク構造からアミノ酸を予測（AlphaFold2 と正反対のことを行う）
タンパク質の機能的な相同体や変異体を作成できる

Configuration & Options

Model Inputs

Input Structure 選択されていません

Allowed Types: pdb

The input protein structure to predict the amino acid sequence of. Acceptable input file formats include PDB.

Design Options

Homo-oligomer ☐

Specify whether the structure is a homo-oligomer (homomer). Lengths of chains should be the same for correct symmetric tying.

Fixed Positions

Fixed positions are positions that will not be inverse folded or affected in anyway by ProteinMPNN. This option allows you to specify which chains, residues, and residue ranges should be fixed. To fix an entire chain simply enter the chain's ID (e.g., "C" to fix all residues of chain C). To fix specific positions use <chain ID><residue ID> (e.g., A10 to fix residue 10 on chain A). To fix a range of residues use <chain ID><start residue ID>-<end residue ID> (e.g., A10-20 will fix all residues between 10 and 20 on chain A). Multiple positions, chains, and ranges can be fixed all at once by comma delimiting your options (e.g., A15,A20-23,B will fix residues 15, 20, 21, 22, 23, and all residues on chain B).

Invert Selection ☐

Invert the selected fixed positions above. Basically if you decide to fix a position like A1-10 above and enable this mode then instead of fixing A1-10, everything will be fixed except for A1-10.

タンパク構造pdbファイルを入力します。

ホモオリゴマー（ホモマー）であるかどうかを指定します。

ProteinMPNNによって解析されない
（保持される）鎖、残基を指定します。

その他の設定も用意されていますが、
Neurosnap Platformは最適なパラメータ設定が
デフォルトとして設定されています。

ProteinMPNN結果

Predicted Sequences

NOTE: Lower ProteinMPNN scores are better.

Rank	Score	Sequence Identity	Sequence
1	0.593	0.585	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...
2	0.596	0.599	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...
3	0.603	0.603	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...
4	0.606	0.63	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...
5	0.606	0.592	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...
6	0.606	0.601	LFREKIPYVVEMEGDVEGMKFSVRGKGHDANTGKIEASFICTTGELPVPWSSILTTVAQCFAYPNNDIKDYPKSAMPEGYVQERTITFENDGVYKTRAEVTYEKGSVYNRVTLNGSGFKKGGNIGKKLEFNYNPHCIYVLPDVQNNGIK...

1行ずつ予測AIで作成された変異体の配列、信頼性のスコアが記載されたテーブルが得られます。

Output Files

Download all as a zip file:

Download All

File Name	File Size
pssm_conditional.json	401.50 KB
pssm_unconditional.json	397.62 KB
results.csv	88.95 KB

配列データのCSVファイルをダウンロードしAlphaFold2に使用します。

わかりやすい設定と結果の閲覧。（本解析では解析時間約30秒）

AlphaFold2設定画面

Configuration & Options

Model Inputs

Target Sequence(s)

↑ Select Proteins

Amino acid sequences of the structures you want to predict in FASTA format. Note that each sequence must have a unique alphanumeric name. To predict complexes / multimers use ":" to specify inter-protein chain-breaks. For example PI...SK:PI...SK for a homodimer.

Custom Template (optional)

ファイルを選択 選択されていません

Allowed Types: pdb, mmcif, or cif

Optionally provide a custom template structure (pdb or mmCIF formats) for AlphaFold2 to use during prediction. Note that custom templates can only be used if you are predicting the structure of a single protein.

Custom MSA (optional)

ファイルを選択 選択されていません

Allowed Types: fasta or a3m

Optionally provide a custom MSA (fasta or a3m format) for AlphaFold2 to use during the prediction. Note that custom MSAs can only be used if you are predicting the structure of a single protein. Make sure the first sequence in the MSA is the one you want to predict the structure of. Sequences will be truncated to the length of the first sequence (sequence that gets folded).

Core Settings

Use Amber ☐

Check this option to utilize amber force fields to relax the predicted structure. This doesn't improve the prediction accuracy but does help remove otherwise distracting stereochemical violations.

Model Type

auto

Select the model you want to use to predict your structure. For monomers we recommend alphafold2_ptm, for complexes we recommend alphafold2_multimer_v3. If you leave this as auto alphafold2_ptm will be used for monomers and alphafold2_multimer_v3 will be used for complexes.

Advanced Settings

Template Mode

none

A template is an optional input structures that should be similar, but different to the structure you are trying to predict. AlphaFold2 uses templates as a sort of "guide" to help it improve its prediction. Using the "none" option will result in no template being used, the "pdb70" option results in a similar structure from the pdb70 being used. Note if you upload a custom template then your selection here will be ignored. Warning: We query a 3rd party service to perform template searches, if you do not wish to share your sequence do not enable this option.

MSA Mode

mmseqs2_uniref_env

Choose the MSA database you want to use. Selecting "single_sequence" will result in the MSA containing only your target sequence. Note if you upload a custom MSA then your selection here will be ignored. In most cases mmseqs2_uniref_env tends to produce the best results. Warning: We query a 3rd party service to perform sequence searches, if you do not wish to share your sequence set this to "single_sequence".

先ほど作成した配列データを入力します。

カスタムテンプレート構造を指定したい場合入力します。

カスタムMSAを指定したい場合入力します。
入力しない場合は、適切なものを自動的に生成できます。

AMBER分子力場を使用した構造緩和を行うオプション。
側鎖が重要な場合は、このオプションを有効にすることをお勧めします。

使用するモデル タイプを選択できます。

「pdb70」オプションを使用すると、pdb70データベースの類似構造が使用されます。

使用する MSA データベースを選択します。
デフォルトのmmseqs2_uniref_env が最良の結果を生成する傾向があります。

AlphaFold2設定画面

Pair Mode

This option applies to the input MSA. "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" - only use paired sequences. Pairing the MSA can help AlphaFold2 create better predictions for complexes.

Number Recycles

Increased recycling steps tend to produce more accurate predictions but will also greatly increase prediction time. For smaller proteins and monomers we recommend 5 recycles, for bigger proteins we recommend 10 recycles. Maximum of 8 recycles for free users. Upgrade your plan to increase recycling values.

Recycles Early Stop Tolerance

Ends the protein folding process if the improvement between recycling steps doesn't surpass the tolerance value. The tolerance is defined as the RMSD (difference in distance matrices, angstrom units) between recycles. If it drops below the specified value, the recycling will terminate. We recommend a tolerance of 0.3 for smaller proteins and 0.5 for larger proteins and complexes.

Number Ensembles

The trunk of the network is run multiple times with different random choices for the MSA cluster centers (1 = default, 8 = casp14 setting). Increasing ensembling can occasionally improve prediction quality but will GREATLY increase prediction time. Maximum of 1 ensembles for free users. Upgrade your plan to increase ensembles values. If you need to increase this value beyond 3 contact customer support. CURRENTLY DISABLED DUE TO HIGH DEMAND. IF YOU NEED THIS VALUE INCREASED PLEASE CONTACT SUPPORT.

Training Mode ☐

Check this option to enable "Training Mode" which activates drop out layers within the model. This can be used to force the model to create more diverse predictions.

MSAのペアリングを制御します。

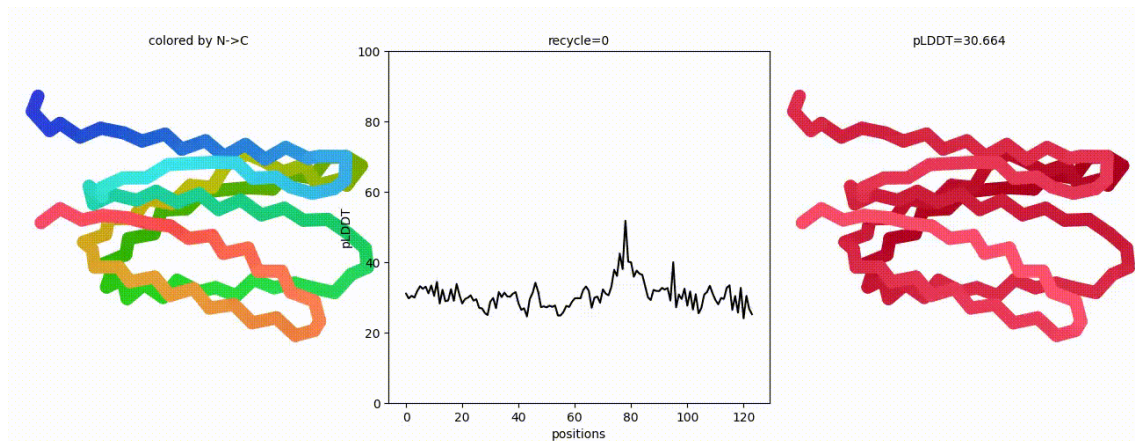
リサイクルステップ（出力を入力として受け取り、予測を構築する処理の回数）の数が増えると、ほとんどの場合、予測精度が向上します。

許容値が指定された値を下回ると、リサイクルは終了します。

入力MSAとして使用するシーケンスをランダムにサンプリングされるアンサンブルの数が多いほど、（特にMSAが非常に大きい場合に）予測の精度が高まります。

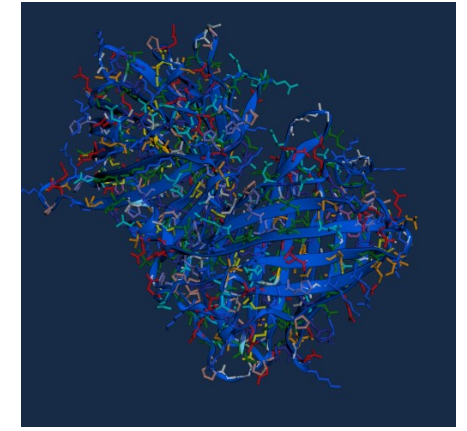
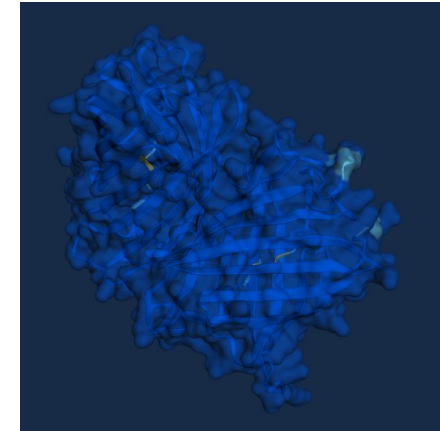
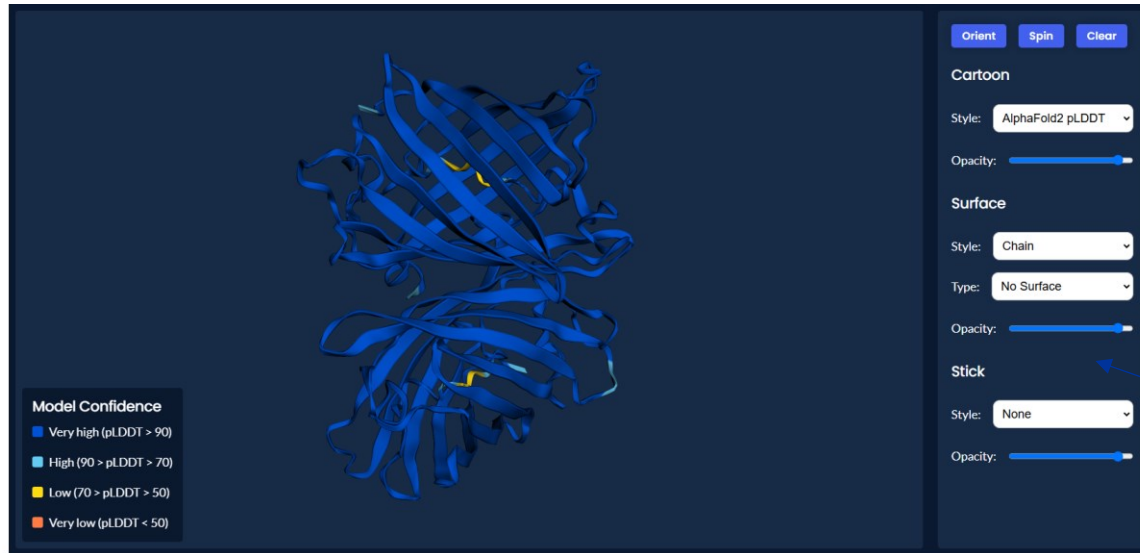
AlphaFold2の設定画面②

予測された構造が不明な場合に、構造空間をサンプリングするのに最適な方法です。



左の例（膜貫通タンパク質）の予測では、正しいフォールディングを得るために12を超えるリサイクルステップが必要です。Credit: Dr. Sergey Ovchinnikov

AlphaFold2結果



予測された構造が視覚化されます。
サイドパネルから表示を変更することも可能です。

Additional Controls

Change Protein Protein_1

Rank 1 Rank 2 Rank 3 Rank 4 Rank 5

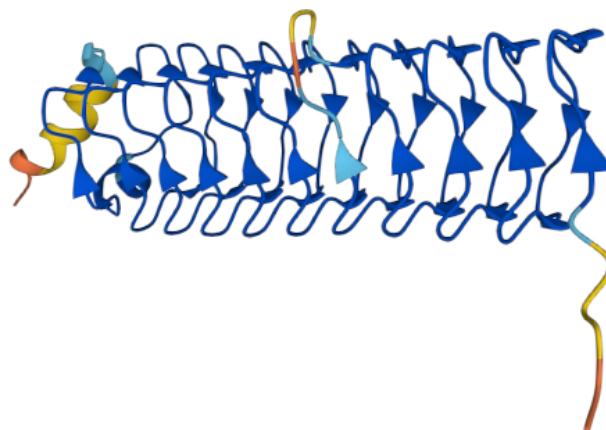
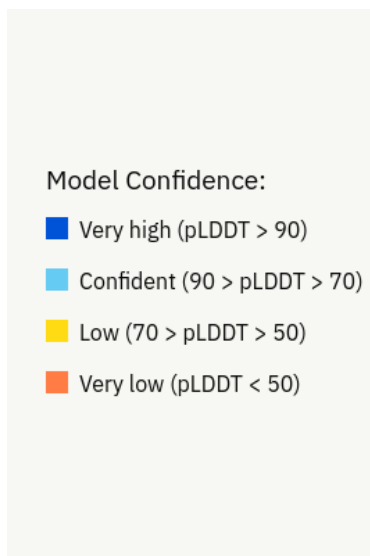
AlphaFold2 predicts 5 different structures using different weights and ranks them from best to worst by their mean pLDDT for monomers or $80 \times \text{ipTM} + 20 \times \text{ptm}$ for multimers. The **Uncertainty** metric ranges between 0 and 100 where 0 means high certainty and 100 means high uncertainty (lower is better). This AlphaFold2 implementation uses ColabFold as it is faster and more accurate.

Predicted Metrics

Model Rank	Mean pLDDT	Max PAE	pTM	ipTM	pDockQ	Uncertainty	Overall Quality
1	96.17	25.5	0.94	0.93	0.59	0.9%	Very High
2	96.08	27.64	0.94	0.93	0.58	1.23%	Very High
3	96.32	23.44	0.94	0.92	0.58	0.96%	Very High
4	95.68	29.67	0.94	0.92	0.60	2.64%	Very High
5	96.12	24.83	0.93	0.91	0.58	1.08%	Very High

5つの異なる構造を予測結果。平均pLDDT
によってランク付けされます。複数の配列を入力
した場合はChange Proteinから切り替えます。

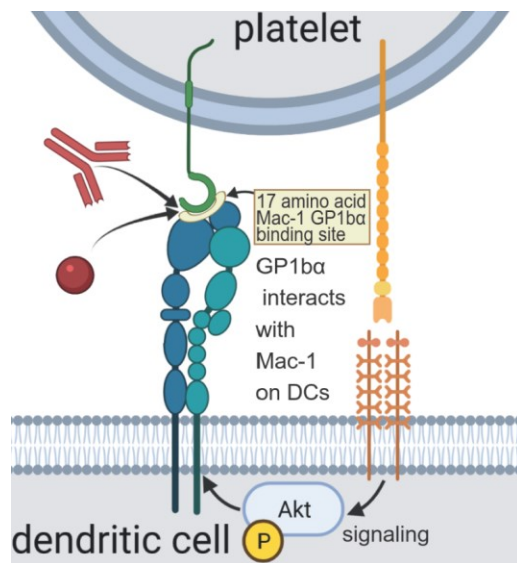
predicted Local Distance Difference Test (pLDDT)



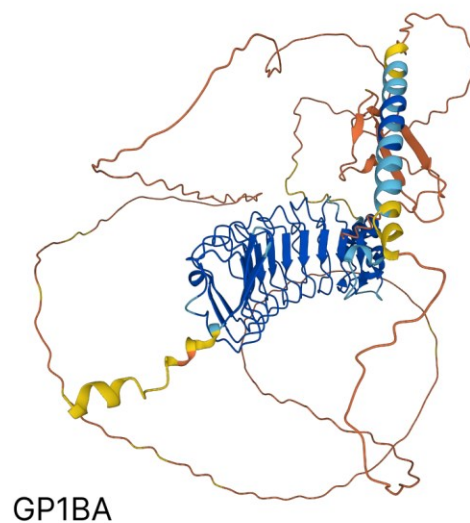
pLDDT は残基ごとの信頼性指標です。値が高いほど予測される残基の位置の信頼性が高くなります。

左上図の予測：

- ほとんどの残基の pLDDT スコアが高い
- 値の低い領域は主にタンパク質の末端付近に集中
→ 保存状態があまりよくない
- 中央に突き出ている、構造内のパターンから外れている小さなセクション
→ 変異が原因である可能性



[Credits for diagram on the left](#)



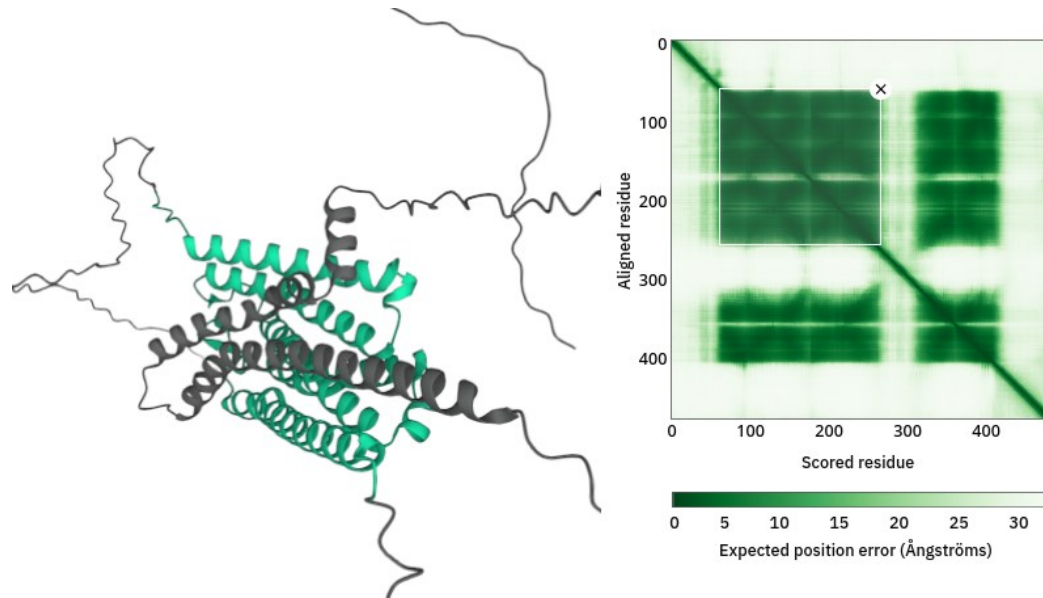
GP1BA

左下図：血小板の形成に関与する膜タンパク質であるヒトGP1BAの予測構造

- VWF-A1結合ドメインと膜貫通ドメインを正しく予測(青)
- 2つのドメイン間のリンカーとして機能する無秩序領域には、正確な位置が可変であるため、低いpLDDT値

→場合によっては、低いpLDDT値が必ずしも不正確な領域を意味するわけではない

Predicted Aligned Error (PAE)

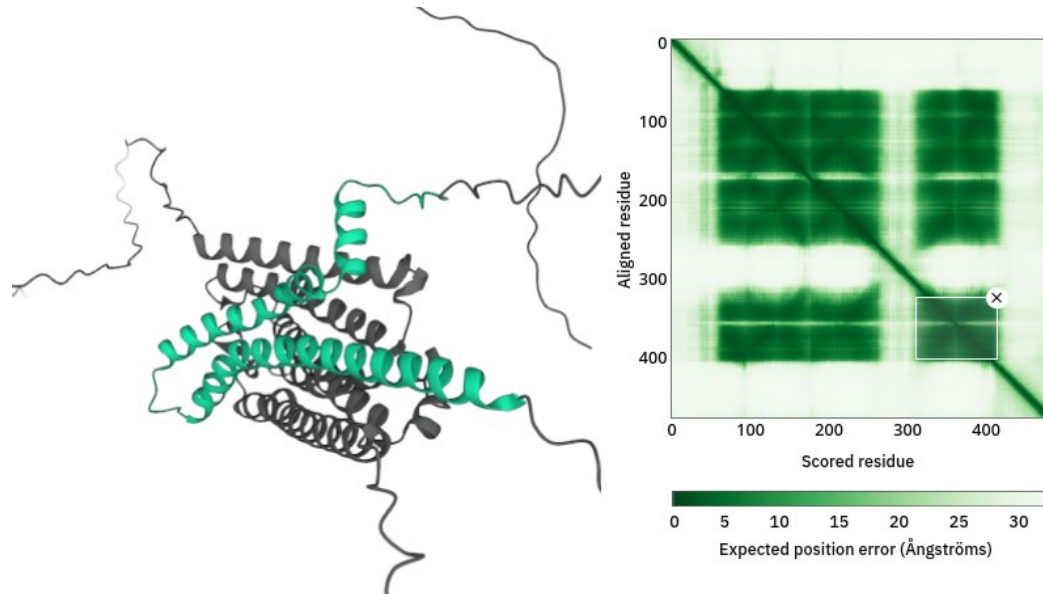


計算された二つの残基間の相対的な位置の誤差を2Dプロットにしたもの。

相対的な位置関係の精度を確認することやドメインごとの柔軟性の指標としても利用できます。（値が低いほど信頼性高）

上図

イヌザメの 5-ヒドロキシトリプタミン受容体 2A (A0A401SYI1) の予測構造について PAE によって強調表示された最初のドメイン。



下図

イヌザメの 5-ヒドロキシトリプタミン受容体 2A (A0A401SYI1) の予測構造において PAE によって強調表示された 2 番目のドメイン。

*Neurosnap Platformの現在のバージョンではカラーリングが異なります。

お問い合わせ先：フィルジェン株式会社

TEL 052-624-4388 (9:00～17：00)

FAX 052-624-4389

E-mail: support@filgen.jp