単粒子解析におけるタンパク質構造分類のための深層学習アプローチの動向

Trends in Deep Learning Approaches for Protein Structure Classification in Single Particle Analysis

馬水　信弥[a, b]*，田中康太郎[c]，安永　卓生[c]
Nobuya Mamizu, Kotaro Tanaka and Takuo Yasunaga

[a]九州工業大学大学院情報工学府
[b]株式会社システムインフロンティア
[c]九州工業大学大学院情報工学研究院

要　旨　クライオ電子顕微鏡による単粒子解析では，試料中に含まれる複数のタンパク質構造を分類しながら解くことが出来る．ただし分類された構造間のダイナミクスの情報は類推するしかない．この問題について2020年に発表された三次元再構成およびクラス分類を行うための深層学習アプローチであるcryoDRGNは，離散的なデータ分割による構造分類を脱却し，連続的な構造分類を実現した．そこでは，オートエンコーダーをベースとし，入力粒子画像から投影パラメーターに依存する情報を分離して潜在空間を構築している．本稿では従来の構造分類と，cryoDRGNおよびその背景となる深層学習のトピックについて解説を行ったのち，構造分類のベンチマークとして6種類の複合体を有するGroEL/ESの実データについて三次元再構成とその分類を試みた．

キーワード：単粒子解析，クラス分類，深層学習，オートエンコーダー

1.　はじめに

クライオ電子顕微鏡による単粒子解析法はタンパク質分子構造解析の一手法として現在重要な位置を占めている．急速凍結されたタンパク質粒子の透過型電子顕微鏡による二次元投影像を計算処理することで一つあるいは複数種類の三次元構造を得ることが出来，その構造はX線結晶解析と比較すると生理的条件に近い構造であることが利点である．加えてタンパク質が試料中で複数の状態をとる場合にそれらの構造を同時に得ることが可能であるが，急速凍結により固定しているので構造間のダイナミクスの情報は失われている．本稿ではそのダイナミクスの理解に向けて，機械学習の可能性について言及し，2020年における単粒子解析の深層学習適用の最新動向の一つを紹介してその将来を展望する．

2.　単粒子解析における構造分類

本題の前に単粒子解析法の代表的なソフトウェアの一つであるRELION[1~3]を例に，クライオ電子顕微鏡像取得後の標準的なフローを俯瞰する（図1）．まず，動画補正およびCTF補正が取得したクライオ電子顕微鏡画像のSN復元のために行われる．動画補正は高性能直接電子検出器の登場により可能となったもので，単粒子解析が現在の高分解能に達した主要因の一つ[4]である．次に，粒子抽出により目的のタンパク質の座標を画像中から収集し切り出す．このフェーズは特に深層学習の適用が目覚ましく自動抽出の手法が多く発表されている[5,6]．続く二次元クラス分類は，粒子画像を教師なし学習によってあらかじめ指定した数のクラスに分類することで，目的粒子のクラス群とノイズ画像のクラス群に分割する．ノイズ画像のクラス群に属する画像を除去することで粒子データのスクリーニングを行う．これらの粒子像を使った初期モデル生成により，目的タンパク質の大まかな三次元構造を計算するが，モデル生成の困難さによっては既知の類

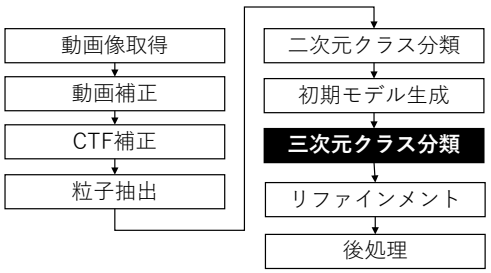

図1　RELIONによる単粒子解析の一般的なワークフロー．本稿のテーマとなる構造分類は三次元クラス分類フェーズに該当する．

*〒820–8502 福岡県飯塚市川津680–4
TEL: 0948–29–7938;　FAX: 0948–29–7938
E-mail: nobuyamamizu@sifi.co.jp
2020年9月1日受付，2020年10月12日受理

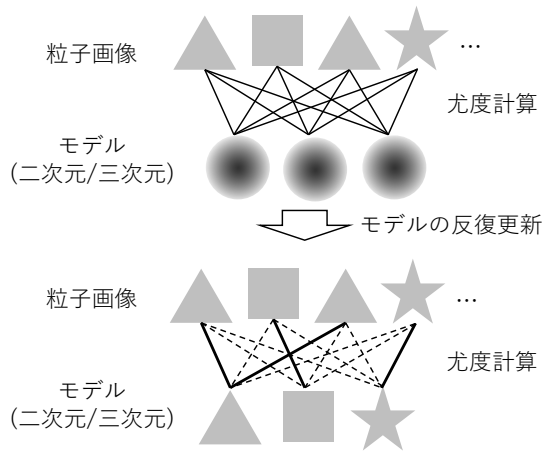

図2　RELION によるクラス分類の模式図．ランダム分割されたデータの重ね合わせからスタートし，最終的にそれぞれの構造へと分岐する．

似構造で代替する．三次元クラス分類は本稿の主題であるタンパク質の構造分類と投影パラメーターの推定を同時に行う．二次元クラス分類にて除き切れなかったノイズ画像も同時にここで排除される．最後に，リファインメント，後処理では三次元クラス分類で分けられたデータについて精密な投影パラメーター（投影角および平行移動）推定，粒子単位の動画補正および CTF 補正のやり直し，溶媒領域のマスクによるノイズの影響の排除を通して，最終構造が決定される．

前述のクラス分類について二次元，三次元ともに考え方は同じである（図 2）．最初に画像をランダム分割してガウス球に近いモデルを生成する．この分割数が，構造を何種類とするかのパラメーターとなる．画像ごとに様々な投影パラメーターでの各モデルに対する尤度を計算しその重みに応じて逆投影することでモデルを更新する．ランダム分割された初期の構造モデルはほぼ違いがないが，わずかな揺らぎが，更新に従って拡大されていき最終的に構造に応じたモデルへとそれぞれ発展していく．実際に，構造分類に成功した事例が数多く報告されている．

一方で，この構造分類法の問題点を挙げる．第 1 に，構造を離散的に分ける過程で構造間の関係は失われており，特にタンパク質のダイナミクスを直接観察することは出来ない点である．第 2 に，データを何種類に分割するかについて，ユーザーにゆだねられているため多くは経験あるいは予測に基づいて分割数が設定される点である．以降これらの問題を解決する糸口として深層学習を適用した例とその可能性を紹介する．

3.　構造分類と多様体仮説

機械学習分野において多様体仮説[7]という概念がある．第 1 に「高次元空間上で表現される実世界のデータは，ずっと低次元の多様体の近傍に集中している」（図 3A），第 2 に「分類問題において，各クラスのデータは低密度領域によって区

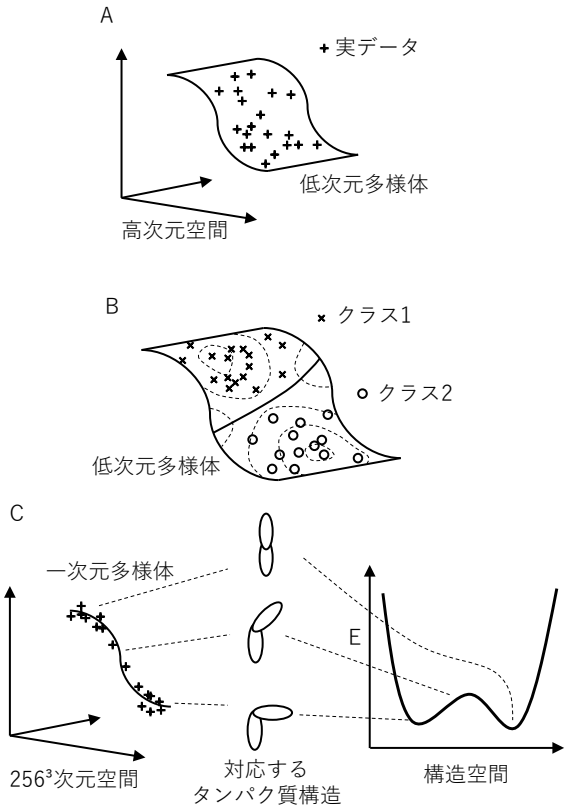

図3　多様体仮説．
（A）実世界の高次元空間中に埋め込まれた，実データが近傍に集中する低次元多様体．（B）低次元多様体中の低密度領域を境界とするクラス分類．（C）256 立方のボクセルデータで表現されたヒンジ構造を有するタンパク質構造が集中する一次元多様体（左）とそれに対応する構造空間（右）．

切られた部分多様体の近傍に集中している」（図 3B）という二つの仮説である．

多様体についての直感的な説明をすると，例えば細胞膜は，実世界という三次元空間上に埋め込まれた二次元多様体である．細胞膜は自由に変形が可能だが，表面の自由度は二次元に限定されている．

我々の関心に多様体仮説を当てはめるなら，例えば一つのヒンジ構造をもち，ヒンジ角度が 0°（始点）または 90°（終点）で安定であるようなタンパク質分子を仮定する（図 3C）．このタンパク質の構造を一辺 256 pixel の三次元密度マップで表現すると，全ての状態は 256^3 次元空間にプロットすることが出来る．しかし，実際の状態はヒンジの角度による一次元の自由度で拘束されており，さらに構造空間の中でヒンジの始点と終点の角度をとる二状態に分布が集中するはずである．これを多様体仮説で解釈するならば，このタンパク質の状態は，ヒンジの角度に対応した一次元の多様体の近傍に集中しており，多様体の二つの端点の中間は低密度領域になっている．このとき多様体上の高密度領域ではタンパク質構造がエネルギー的に安定で，低密度領域では不安定であるとして，図 3C 左図の多様体はちょうど図 3C 右図のような構造

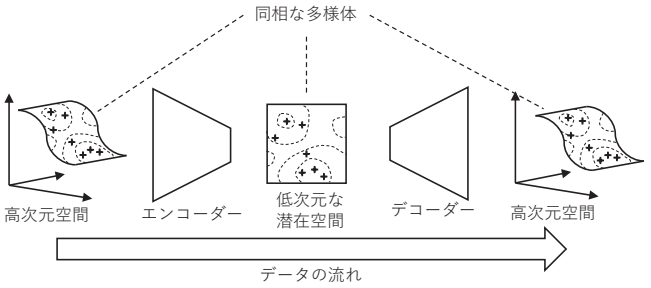

図4　オートエンコーダーによる次元削減．ネットワークは学習データが近傍に集中する多様体を復元する必要があるため，潜在空間および出力先でも入力データのもつ同相な低次元多様体を保存されることになる．

空間に対応したものと考えることも出来る．

RELION で行われる離散集合への分割による構造分類は図 3B のクラス 1，2 の 2 種の構造分類に対応する．多様体仮説で解釈すれば理想的には低密度領域によって区切られた部分多様体ごとの集合になっているはずだが，そのデータ群全体がもともと属している多様体の振る舞いは分からないため，図 3C で示されるような構造間の関係を類推する必要がある．逆にその多様体が何らかの方法で取得出来れば，クラス間の位置関係や多様体上のデータの粗密からより深い情報が引き出せるのではないだろうか．

4.　多様体推定のための次元削減：オートエンコーダー

入力データそのままの高次元空間から関心領域が属する低次元多様体を直接推定することは難しいため，低次元多様体が埋め込める範囲で次元削減する必要がある．次元削減の代表的な手法として主成分分析による線形変換があるが，ここでは深層学習の一種であるオートエンコーダー（autoencoder：自己符号化器）[8] による非線形次元削減を紹介する．

オートエンコーダーとは，入力と出力を同じデータとするようなニューラルネットワークのことである（図 4）．学習は入力と出力を比較することで行われ，その差を小さくするように各層のパラメーターが学習される．オートエンコーダーは，前段にエンコーダーと呼ばれる次元削減を行うネットワーク，後段に削減された次元上のデータから元の入力を復元するデコーダーと呼ばれるネットワークから構成される．エンコーダーの出力かつデコーダーの入力となる空間のことを「**潜在空間**」という．潜在空間はデコーダーが学習データを復元するだけの情報量を保持するようにエンコーダーによって構築されるため，入力の高次元空間上においてデータが集中する低次元多様体と同相な多様体が埋め込まれることになる（図 4）．潜在空間上の各点は，それぞれ「**潜在変数**」として，対象の異なる特徴を表現することになる．

オートエンコーダーの拡張として変分オートエンコーダー[9]（VAE: variational autoencoder）がある．オートエンコーダーの潜在空間を確率分布で構築することにより，潜在空間の連続性を向上させたものである．この特性は，深層学習でよく使われる敵対的生成ネットワーク（GAN: generative adversarial network）[10] に潜在空間上からサンプリングした特徴を入力させる VAEGAN[11] などでも応用されている．前述したように，この VAE が示した，次元削減された潜在空間の連続性の向上は，タンパク質の複数構造やそれらのダイナミクスを表現する構造空間を表現出来る可能性がある．

5.　cryoDRGN

本稿では，単粒子解析における構造分類と三次元再構成を同時に行う深層学習アプローチである cryoDRGN[12,13] を紹介する．cryoDRGN では，対象物の回転や並進に対して不変になるようにエンコードを行うように VAE を拡張した spatial-VAE[14] を応用している．2020 年 8 月本稿執筆現在では，International Conference on Learning Representations（2020）での発表以外では，プレプリント誌に投稿されたのみであるが，ソフトウェア実装が公開されている．

cryoDRGN で利用された spatial-VAE と VAE の違いは，デコーダーの入出力にある．すなわち，入力として，潜在変数 z のほかに，座標 x を渡し，出力は座標 x 上の値を推定するとする（図 5）．その結果として，潜在空間は入力画像の座

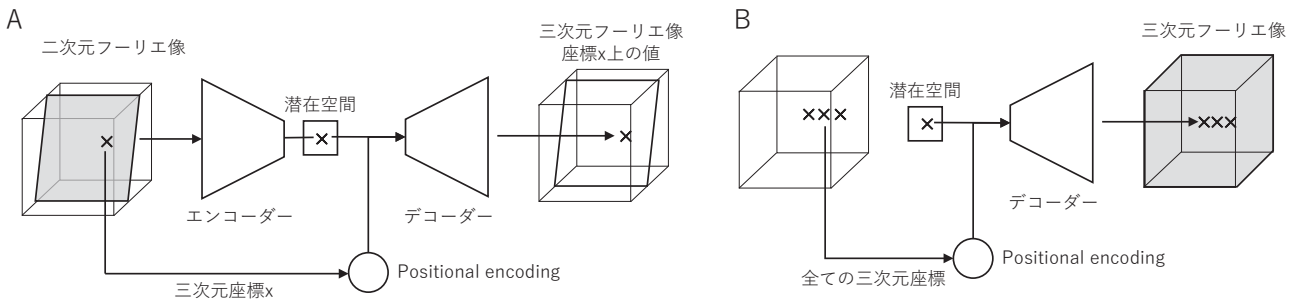

図5　cryoDRGN のネットワークアーキテクチャ．
（A）学習時．エンコーダーは，入力として粒子の二次元フーリエ像が受けとられ，次元圧縮された潜在空間を出力する．その後，デコーダーは，入力として，粒子の投影パラメーターから計算した三次元座標 x とエンコードされた潜在変数を受け取り，座標 x 上の値を出力する．この出力値と元の二次元フーリエ像の値が同じになるように学習する．（B）生成時．デコーダーのみを用いて一つの潜在変数に対する，全ての三次元座標上の値を計算する．

標変換の情報を除去され，対象物の構造の特徴のみを表現出来るように構築されうる．

ここで，cryoDRGN に実装されたタンパク質構造解析のアーキテクチャを図 5 に示す．spatial-VAE によって投影パラメーターによる変換の情報は潜在空間から除去されるため，潜在空間がタンパク質の構造空間としてエンコードされていることが期待される．粒子画像の二次元フーリエ像を入力とし，潜在変数 z を出力するエンコーダーと，潜在変数 z と三次元フーリエボクセルの三次元座標 x を入力とし，タンパク質の三次元再構成の三次元フーリエ変換の入力座標上の値を返すデコーダー部分からなる．ここでは詳細は省くが，三次元座標 x は Positional encoding というニューラルネットワークに適した変換が行われている．三次元再構成する際はある潜在変数に対して三次元座標全ての出力を計算する．これにより，全ての三次元フーリエ空間の値が得られ，それをフーリエ合成することにより，三次元ポテンシャルマップを得ることが出来る．

6. GroEL/ES 実データを用いた潜在空間の観察

今回，cryoDRGN が生成する潜在空間について，当研究室で収集した GroEL/ES[15,16] 複合体のクライオ電子顕微鏡データで考察してみた（図 6）．GroEL/ES は分子量 57 kDa の EL サブユニットの 7 量体リング（以下，SR: single-ring），分子量 10 kDa の ES サブユニットの 7 量体リング（以下，ES）を基本構造とし，それらの組み合わせによる 6 種類の複合体をとる（図 6A）．さらに SR が背中合わせに重なったバレル型と，そこへ ES が蓋のように結合したバレット型は，リングがつくるカゴの開閉を伴う連続的な構造変化を起こす．離散的および連続的構造変化が同一試料内に含まれるため，潜在空間の観察対象として適している．

cryoDRGN の学習データの準備として，撮影した電子顕微鏡画像（4,903 枚，図 6B）から二次元クラス分類による粒子選別（図 6C）を経て 5,908 枚の粒子画像を得た．これらの粒子画像に対して，クラス分類せずに単一の三次元構造を再構成し（4.1 Å 分解能），cryoDRGN の学習に必要な各粒子の投影パラメーターを得た．これらの計算は cryoSPARC[17] を用いて行った．なお，構造のうち ES 単体については粒子抽出および二次元クラス分類の時点で上手く拾えなかったため，以降の解析では考慮しないことにした．

cryoDRGN の潜在空間次元数は任意に指定可能である．今回は多様体学習による変形を受けないそのままの構造を目視で確認出来るように三次元で計算を行った．

学習後の潜在空間の様子を図 6D に示す．各粒子画像に対応する潜在変数が点で表されている．ES を除く 5 種類の複合体の存在が期待されることから，K-means 法により 5 種類のクラスターに分類した結果が図 6E である．各クラスター

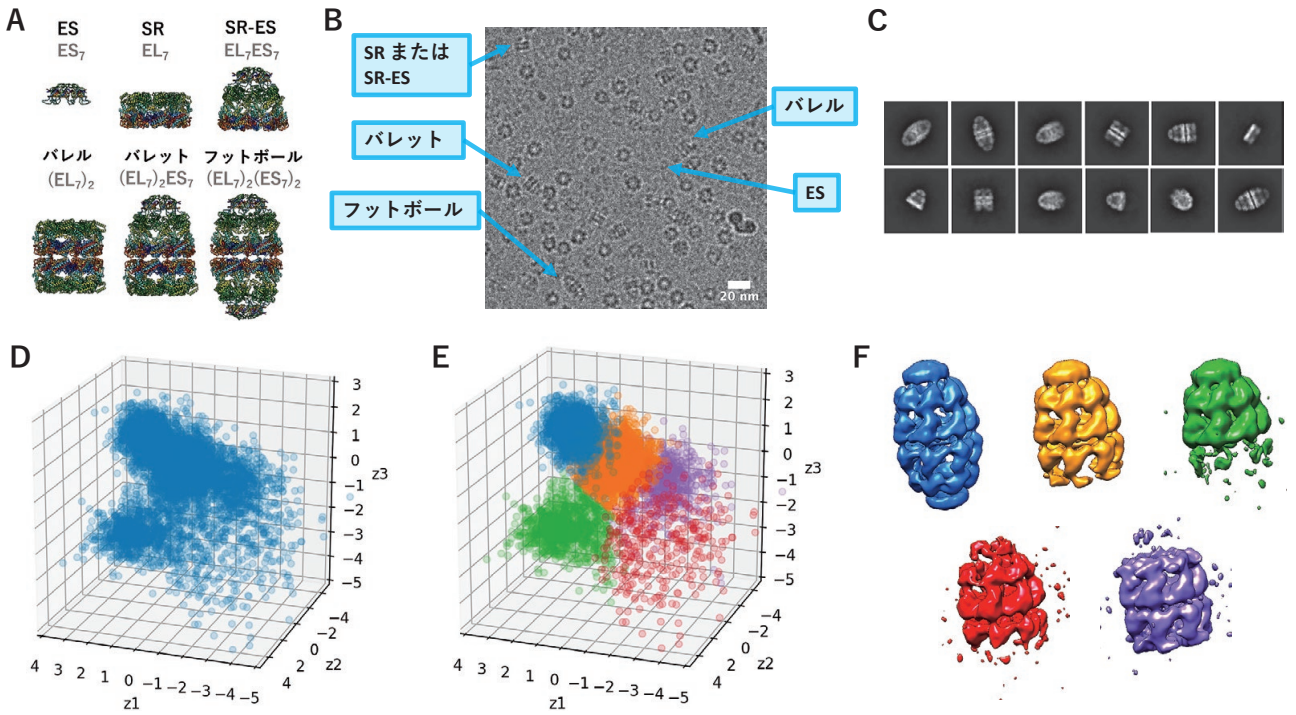

図 6　GroEL/ES 実データを用いた潜在空間の観察

（A）GroEL/ES の 6 種類の複合体構成の構造モデル（PDB バレル：5W0S，バレット：1PCQ，フットボール：1PKO．ES, SR, SR-ES はそれぞれ 1PCQ, 5W0S, 4PKO の当該部分を抜き出し）．（B）クライオ電子顕微鏡像．（C）選択した二次元クラス平均像．（D）cryoDRGN により生成された三次元の潜在空間上のプロット．（E）（D）に対し K = 5 で K-means を適用しクラス別で色分けした結果．（F）（E）による各クラスターの粒子像について推定済みの投影パラメーターで逆投影を行った三次元再構成像．対応クラスターに色を合わせてある．

の構造を確かめるため，それぞれ独立に逆投影して三次元構造を再構成した（図 6F）．青，黄，紫のクラスターはそれぞれフットボール型，バレット型，バレル型と確認出来，緑は密度下部にノイズが目立つものの SR-ES 型の特徴を示している．この結果から，SR 型以外については期待通り低密度領域で区切られた部分多様体が獲得されたと言え，多数の複合体種を含む試料の構造解析に有用であると確認出来た．さらにクラスターの位置関係に注目すると，バレット型（黄）を中心として，そこから ES が外れたバレル型（紫），ES がついたフットボール型（青），SR が外れた SR-ES 型（緑）が取り囲んでおり，類似した構造が直感的に分かりやすい形で近接していて解釈性も高い．一方で赤のクラスターの分布は比較的散漫で，バレット型，バレル型，SR-ES 型のクラスターに隣接しており，逆投影による三次元構造もそれらの混合のようで，構造の特徴づけは難しい．残る複合体種は SR 型であるから SR 型の粒子のクラスターと期待されるが，最も分子量が小さいこともあり，SN の悪い画像や判別の難しい画像などと合わさって分かりにくいクラスターになっている可能性があり，精査が必要である（投稿準備中）．

7．今後の展望

単粒子解析における深層学習の利用は，今まで粒子抽出については盛んに行われていたが，ついに三次元再構成の領域にまでその適用範囲を広げたのは驚くばかりである．今回 cryoDRGN とそれが生成する潜在空間に焦点を当てて紹介した．データに含まれる構造の数，種類を確かめるための強力なツールとして期待される．

深層学習は特徴量を非明示的にエンコードしてくれる点が優れているがそこから何を引き出すかといった意味付けが今後の課題といえる．例えば cryoDRGN による構造空間の構築によって構造を分類し，クラスターに分けることは可能となった．さらに，ダイナミクスの情報を抽出するために，クラスター間の遷移について cryoDRGN の提案者らにより，k 近傍グラフをたどる方法が試みられている[11]．一方で，クライオ電子顕微鏡法が，全ての水和したタンパク質の構造を捉えることが出来ていることを最大限活かし，ダイナミクスの情報を抽出したい．そのためには，潜在空間を構造空間として捉えるために，構造のもつエネルギー状態と対応させる．そうすれば単なる構造分類ではなく，タンパク質のもつダイナミクスに繋げることが出来る．加えて，特定の安定した 3 次構造をもたないタンパク質や糖の構造解析への道も開けよう．ダイナミクスの探索はまだまだ挑戦的で，我々を魅了する課題である．

謝　辞

GroEL/ES の電子顕微鏡画像収集は理化学研究所の重松秀樹博士のご協力のもと行われた．本研究は，JST，CREST，JPMJCR1865 の支援を受けたものである．

文　献

1）Scheres, S.H.W.: *Journal of Molecular Biology*, 415(2), 406–418 (2012)
2）Scheres, S.H.W.: *Journal of Structural Biology*, 180(3), 519–530 (2012)
3）Zivanov, J., Nakane, T., Forsberg, B.O., *et al.*: *eLife*, doi: 10.7554/eLife.42166 (2018)
4）Li, X., Mooney, P., Zheng, S., *et al.*: *Nature Methods*, 10, 584–590 (2013)
5）Wagner, T., Merino, F. and Stabrin, M.: *Communications Biology*, 2, 218 (2019)
6）Wang, F., Gong, H., Liu, G., *et al.*: *Journal of Structural Biology*, 195(3), 325–336 (2016)
7）Rifai, S., Dauphin, Y., Vincent, P., *et al.*: *Advances in Neural Information Processing Systems*, 24, 2294–2302 (2011)
8）Hinton, G.E. and Salakhutdinov, R.R.: *Science*, 313(5786), 504–507 (2006)
9）Kingma, D.P. and Welling, M.: *arXiv*:1312.6114v10 (2014)
10）Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al.*: *Advances in Neural Information Processing Systems*, 27, 2672–2680 (2014)
11）Larsen, A.B.L., Sønderby, S.K. and Larochelle, H.: *Proceedings of Machine Learning Research*, 48, 1558–1566 (2016)
12）Zhong, E.D., Bepler, T., Berger, B. and Davis, J.H.: *bioRxiv*, doi: 10.1101/2020.03.27.003871 (2020)
13）Zhong, E.D., Bepler, T., Davis, J.H. and Berger, B.: *International Conference on Learning Representations 2020*, arXiv:1909.05215 (2020)
14）Bepler, T., Zhong, E.D., Kelley, K., *et al.*: *arXiv*: 1909.11663 (2019)
15）Hayer-Hartil, M., Bracher, A. and Hartl, F.U.: *Trends in Biochemical Science*, 41(1), 62–76 (2016)
16）Yan, X., Shi, Q., Bracher, A., *et al.*: *Cell*, 172(3), 605–617 (2018)
17）Punjani, A., Rubinstein, J.L., Fleet, D.J. and Brubaker, M.A.: *Nature Methods*, 14, 290–296 (2017)