調査レポート

拡散モデルを用いた分子構造生成

デジタルエンジニアリングチーム
主任コンサルタント
石田純一

発展著しい機械学習分野で，近年特に注目を集めている技術の１つが拡散モデルである．拡散モデルは DALLE 2（OpenAI 社），Imagen（Google 社）といった高精度な画像生成エンジンのコア技術として活用されたことで，世界的な研究開発に火が付いた．既に画像だけではなく音楽生成，分子生成，動画生成といった幅広い用途への適用が進められている．本稿では，特に分子構造生成における拡散モデルの適用事例を調査し，その概要と将来的な展望を述べる．

1 はじめに

人工知能（Artificial Intelligence，AI）技術は過去十年の間に著しい進展を遂げた．特にコンピュータビジョンと自然言語処理は AI 技術開発の中核分野となっており，人間が作成したものと見分けられないレベルでの画像・文章生成技術が既に実現している．これら 2 つの分野にまたがる技術として，文章の指示に従い画像を生成する text-to-image が注目を集めている．2022 年に OpenAI 社，Google 社から立て続けに発表された画像生成エンジン DALL･E 2 と Imagen は，それまでの技術とは一線を画す卓越した画像生成能力を有しており，プロンプトと呼ばれる指示文章に忠実な画像を生成することができる．さらに Stability AI 社により高精度な学習済みモデル Stable Diffusion がオープンソースで公開され，ローカル端末や Google Colaboratory といった環境で気軽に画像生成を楽しむことができる．

これらの技術のコアにあるのが，拡散モデルと呼ばれる訓練データの分布を学習し新規データを生成する機械学習モデルである．従来，生成モデルとして敵対的生成ネットワーク（Generative Adversarial Network，GAN）の研究が最も盛んに進められてきたが，パラメータ調整が非常に難しく，生成データに多様性がない（モード崩壊）といった課題があった．拡散モデルを用いることで高精度かつ多様なデータを生成できることから，近年は画像生成のみならずテキスト生成，音楽生成，分子生成，動画生成など適用範囲が広がっている．

当部では科学技術分野における多様なソリューションを提供しており，特に材料、化学ならびに創薬のための AI を重要な開発項目と位置付けている．そこで本稿では，拡散モデルの概要について説明したのち，特に分子構造生成モデルについて最新の適用事例を示すとともに，今後の展開について述べる．

2 拡散モデル

拡散モデルの直観的なイメージを**図 1** に示す．学習対象データx_0（ここでは分子構造）に対し，段階的にノイズを付加することで，完全なノイズデータx_Tを生成する．続けて，ノイズを除去する逆プロセスを経て真のデータを復元する．ノイズ除去の方法として，ノイズそのものを学習対象とするノイズ除去拡散確率モデル（Denoising Diffusion Probabilistic Model，DDPM)やスコアと呼ばれるデータ分布の勾配を学習対象とするノイズ除去スコアベースモデル（Denoising Score-based Model，DSM）といった異なるアプローチが知られており，それらがまとめて拡散モデルと呼ばれる [1,2]．本稿では代表的な手法である DDPM と DSM の概要について説明する．なお，これらのアプローチの学習対象は係数部分を除き等価であることが証明されており，確率微分方程式で統一的に表現する試みも行われている．

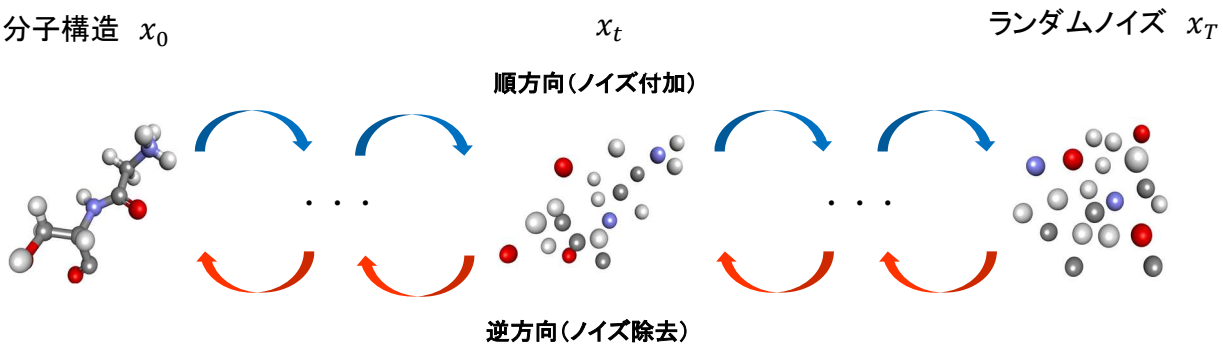

図 1　3 次元原子座標を学習対象とした場合の拡散モデルによる分子生成イメージ

2.1 DDPM

DDPM では，学習データが完全なランダムノイズになるまでガウスノイズを段階的に付加する．その後，これを逆にたどりノイズを除去するニューラルネットワークを訓練することでデータを復元する．分子構造等のデータx_0に対するTステップのノイズ付加過程は以下の式で表される．

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{2}$$

ここで$q(x_t)$は潜在変数x_tの分布，$\mathcal{N}$は多変量正規分布，β_tはステップtにおける分散を制御するハイパーパラメータであり，ノイズ付加されたデータはガウス分布から生成される．一方，ノイズ除去を行う逆方向プロセスは学習可能な分布p_θを用いて以下のように表され，ノイズ除去データも学習可能なパラメータθを含むガウス分布から生成される．

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)\right) \tag{4}$$

ここで$p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$であり，完全なランダムノイズに対応する．また$p_\theta(x_t)$は逆プロセスにおける潜在変数$x_t$の分布である．学習データ$x_0$に関する対数尤度$\log p_\theta(x_0)$を最大化するため，変分下限を計算する．文献[1]に従う再パラメータ化および単純化を行うことで，目的関数L_t^{simple}は以下のようなノイズに関する二乗誤差として表される．

$$L_t^{\text{simple}} = \mathbb{E}_{t\sim[1,T],x_0,\varepsilon_t}[||\varepsilon_t - \varepsilon_\theta(x_t,t)||^2] \tag{5}$$

ここでε_θはノイズを表すニューラルネットワークであり，ε_tはステップtにおけるノイズである．直観的には，ノイズ付加データのうち，どの部分がノイズであるかをニューラルネットワークが学習する．得られたデータ分布の勾配を用いて，新しいデータがサンプリングされる．

2.2 DSM

DSM では，真のデータ分布$q(x)$のデータxに関する勾配を，ニューラルネットワークp_θの勾配を用いて推定する．

$$s_\theta(x) = \nabla_x p_\theta(x) \tag{6}$$

ここでs_θはスコア関数と呼ばれている．真のデータ分布の勾配$\nabla_x q(x)$をスコア関数で近似することで，以下のサンプリング手法によりデータを生成することが可能となる．

$$x_{i+1} \leftarrow x_i + \varepsilon\nabla_x q(x) + \sqrt{2\varepsilon}z_i, i = 0,1,\cdots,K \tag{7}$$

ここでz_iはガウス分布から生成されるノイズであり，$\varepsilon \to 0$, $K \to \infty$とすることで解が収束する．

s_θは一般にスコアマッチングと呼ばれる手法により推定されるが，データ分布が疎な領域では推定精度が下がるという欠点がある．これに対処するため，データ全域に段階的にノイズを付加し完全にランダム化する．その状態から段階的にノイズを除去し，異なるノイズレベルにおけるスコアを段階的に学習することで，推定精度を向上させることができる．

さらに文献[2]では，順方向・逆方向プロセスを確率

微分方程式（Stochastic Differential Equation, SDE）の解として滑らかに表現することで，モデルの高精度化に成功している．ノイズ付加順方向 SDE は以下のように書き下すことができる．

$$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w \qquad (8)$$

ここでxは連続的な時間変数tに関するデータ分布である．wは標準 Wiener 過程であり，$\mathrm{d}w$は無限小のホワイトノイズに対応する．fはドリフト係数，gは拡散係数と呼ばれる．一方，逆方向プロセスは以下のようになる．

$$\mathrm{d}x = [f(x,t) - g(t)^2\nabla_x \log q_t(x)]\mathrm{d}t + g(t)\mathrm{d}w \qquad (9)$$

右辺に登場する$\nabla_x \log p_t(x)$をスコア関数で近似するため，以下の目的関数を最小化する。

$$\mathbb{E}_{t,x_t}[\lambda(t)||s_\theta(x_t,t) - \nabla_{x_t}\log q_t(x_t)||^2] \qquad (10)$$

$\lambda(t)$は重みパラメータであり，スコアマッチングにより式(10)を計算しニューラルネットワークs_θを学習する．微小変位$\mathrm{d}x$を用いてxを順次更新することにより，データが生成される．

3 分子構造生成モデル

拡散モデルを分子生成に活用した研究事例は，本稿の執筆時点でも増加の一途を遂げている．**表 1** には，2022 年 9 月現在報告されている拡散モデルによる分子生成の報告事例を整理した．適用先のターゲットは低分子，タンパク質，結晶材料の 3 種類に大別され，それらに対し様々な目的で拡散モデルが適用されている．以下，各ターゲットに対する研究開発の状況を説明する．

3.1 低分子

低分子は比較的小規模な分子系であり，材料科学・創薬の多くの場面で活用されるため，以前から生成モデルの適用対象として活発な研究が進められてきた．拡散モデルの適用例として，既に新規構造生成，コンフォーマー生成，分子動力学計算におけるトラジェクトリ生成といった様々な用途が報告されている [3-7]．

新規構造生成は，学習データセットの分子構造分布を学習し，学習データに含まれない未知の分子構造を生成することを目的とする．実用上は所望の物性を有する新規構造を生成することがより重要である．文献 [3,4]では，低分子が従う並進・回転対称性を保つための特殊なタイプのニューラルネットワークを活用し，拡散モデルを構築している．低分子は構成原子の 3 次元座標や周辺情報を組み込んだ行列形式で表現されており，いずれのケースでも従来手法に比べてより多様かつ現実的な分子生成に成功している．

コンフォーマー生成は特に創薬分野において重要視されている技術である．生体中での多様な構造変化に対応した構造群がコンフォーマーとして呼ばれ，それらの集団的な振る舞いによって分子の特性が決まる．コンフォーマー生成では特にねじれ角の自由度が重要であるという発想に基づき，文献 [6]ではねじれ角のみを学習対象としている．これにより，3 次元座標を生成する場合より数桁高速なコンフォーマー生成を可能としている．また，得られた構造は OMEGA や RDKit といった一般的なソフトウェアを上回る多様性や，真の構造アンサンブルに近い化学的特性を示している．

生成モデルは分子動力学（Molecular Dynamics，MD）計算のような動的構造の生成にも活用されている．文献 [7]では拡散モデルを MD 計算におけるトラジェクトリ生成に活用し，原子の座標を直接生成することによって，トラジェクトリの安定性を評価する指標において他の機械学習手法より優れた結果を示した．

3.2 タンパク質

タンパク質は低分子に比べ構成原子数が膨大であり，アミノ酸と呼ばれる 20 種類の構成要素の組み合わせから成る．生体内ではタンパク質と他の分子系との相互作用が重要であり，低分子系に比べより多様な目的で拡散モデルの事例が報告されている [8-13]．

構成原子数が膨大であることから，生成データ対象をタンパク質の主要骨格である主鎖の特徴に限定した事例が目立った．生成データからタンパク質の構造を復元するため，アミノ酸残基間の角度や距離，主鎖に結合した側鎖の結合角を拡散モデルによって生成するなど，様々な工夫が施されている [9-12]．生成可能なタンパク質の大きさに制約が課される場合があるが，ネイティブ構造に近い構造の生成，条件付き構造生成，モチーフ周辺のタンパク質の構造補完（イ

表 1　拡散モデルを活用した分子生成に関する主要文献（2022 年 9 月現在）

適用系	目的	生成対象	手法	概要
低分子	新規構造生成	3 次元座標，原子種類	DDPM	等変量グラフネットワークを利用し３次元分子構造を生成 [3]
		隣接行列，特徴行列	DSM	学習データ分布外かつ所望の物性を有する分子を生成 [4]
	コンフォーマー生成	3 次元座標	DDPM	座標を生成する拡散モデルによりコンフォーマーを生成 [5]
		ねじれ角	DSM	フレキシブルな自由度であるねじれ角を生成しコンフォーマーを作成 [6]
	トラジェクトリ生成	各時刻の 3 次元座標，原子種類，速度	DSM	分子動力学トラジェクトリを生成 [7]
タンパク質	新規構造生成	3 次元座標	DDPM	任意のモチーフ構造にインペインティングすることでタンパク質の主鎖を生成 [8]
		原子間角度	DDPM	最大 128 残基の多様かつ新規な主鎖を生成 [9]
		Cα 原子座標，クォータニオン，側鎖結合角，アミノ酸種類	DDPM	タンパク質の構造・配列を生成，マスクされた領域のインペインティングも可能 [10]
		残基間角度，距離	DSM	ネイティブに近いタンパク質構造を生成，また条件付き生成も可能 [11]
	抗体生成	アミノ酸種類，Cα 原子座標，配向	DDPM	抗原と作用する抗体の相補性決定領域（CDR）の配列と構造を生成 [12]
	複合体生成	3 次元座標	DSM	タンパク質骨格テンプレート，低分子グラフ構造を入力として複合体構造の座標を生成 [13]
結晶	新規構造生成	3 次元座標，原子種類	DSM	結晶材料構造を生成 [14]

ンペインティング）などいくつかの成功例が報告されている．また，薬剤設計において重要となる抗原と相互作用する抗体の相補性決定領域（Complementary Determining Region，CDR）の配列・構造設計に活用された事例や，タンパク質とリガンドを入力とした結合構造の生成，妥当なリガンドポーズの生成など多くの活用方法が模索されている．

3.3　結晶材料

拡散モデルは低分子，タンパク質系への適用がメインストリームとなっているが，結晶材料生成における適用事例も報告されている．結晶系では並進・回転対称性に加えて，周期的対称性を考慮したモデルを構築する必要がある．文献 [14]では，これらの対称性を保持したニューラルネットワークを用いて結晶材料を扱っており，構造の再構成，新規構造の生成，特性最適化において既存手法を上回る精度を報告している．本モデルによって生成された仮想 2 次元材料データベースも他の研究グループによって公開されている [15]．

4　今後の展望

以上，拡散モデルを用いた分子生成技術に関して概説した．猛烈なスピードで世界的に研究が進められており，**表 1** に記載した研究事例は文献 [14]を除き 2022 年に報告されたものであることは特筆に値する．特に創薬分野においては多様なアプリケーションが模索されており，コンフォーマー生成やドッキング構造の生成では既に従来手法を上回る精度が報告されている．実際の薬剤設計の現場へ導入可能なレベルでの，優れた実力が証明されつつあると言える．

材料科学への応用方法として，生成モデルを用いた新規化合物のデータベース構築や，それを活用した量子化学計算による有望素材のスクリーニングといった方向性が想定される．実用的な材料開発を指

向した場合，特定の環境下における化学的安定性や複数の物性に最適化した材料が要求されるが，現在の生成モデルではそれらすべての要求に応えることは難しい．そのため，大量に生成された候補材料に対してスクリーニングを行うことで，未知の機能性材料を効率的に探索できる可能性がある．

一方で，拡散モデル一般の課題として挙げられるのがサンプリング速度の遅さである．データをノイズ化するために大量のステップを要するため，データ生成に長大な時間を要する．高速化に向けた検討は盛んに進められており、サンプリング間隔を間引いたり圧縮された潜在空間におけるデータに対して拡散モデルを適用するといった試みが行われている。高精度化と高速化の両立は拡散モデルにおける重要な研究テーマとして今後も活発に開発が進められるであろう．

5 結び

本稿では，近年注目を集める拡散モデルについて，DDPM，DSM の 2 つのアプローチと材料・創薬分野における主な研究開発動向を概説した．創薬を中心として多様な適用例が報告されており，拡散モデルを用いることで従来の機械学習生成モデルに対して優位または競争力のある結果を達成できる．特にコンフォーマー生成やドッキング構造生成では，従来手法を大きく上回る精度が報告されており，基礎研究に留まらず実際の開発現場へと導入可能な糸口が見えつつある．一方で，生成分子の特定環境下での化学安定性や学習データに含まれていない物性の特性は保証されていないため，量子化学計算等の従来の手法との適切な組み合わせが必要となるだろう．また，化合物は実際に合成されることで初めて実用的な価値が生まれる．コンピュータ上での予想から一歩踏み込んだ，実材料に関する検証が今後期待される．

引 用 文 献

1) Ho, Jonathan, Ajay Jain, and Pieter Abbeel: Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851
2) Song, Yang, *et al.*: Score-based generative modeling through stochastic differential equations, *arXiv:2011.13456* (2020)
3) Hoogeboom, Emiel, *et al.*: Equivariant diffusion for molecule generation in 3d, *arXiv:2203.17003* (2022)
4) Lee, Seul, Jaehyeong Jo, and Sung Ju Hwang: Exploring Chemical Space with Score-based Out-of-distribution Generation, *arXiv:2206.07632* (2022)
5) Xu, Minkai, *et al.*: Geodiff: A geometric diffusion model for molecular conformation generation, *arXiv:2203.02923* (2022)
6) Jing, Bowen, *et al.*: Torsional Diffusion for Molecular Conformer Generation, *arXiv:2203.02923* (2022)
7) Wu, Fang, *et al.*: A score-based geometric model for molecular dynamics simulations, *arXiv:2204.08672* (2022)
8) Trippe, Brian L., *et al.*: Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem, *arXiv:2206.04119* (2022)
9) Wu, Kevin E., *et al.*: Protein structure generation via folding diffusion, *arXiv:2209.15611* (2022)
10) Anand, Namrata, and Tudor Achim: Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models, *arXiv:2205.15019* (2022)
11) Lee, Jin Sub, and Philip M. Kim: ProteinSGM: Score-based generative modeling for de novo protein design, *bioRxiv* (2022)
12) Luo, Shitong, *et al.*: Antigen-specific antibody design and optimization with diffusion-based generative models, *bioRxiv* (2022)
13) Qiao, Zhuoran, *et al.*: Dynamic-Backbone Protein-Ligand Structure Prediction with Multiscale Generative Diffusion Models, *arXiv:2209.15171* (2022)
14) Xie, Tian, *et al.*: Crystal diffusion variational autoencoder for periodic material generation, *arXiv:2110.06197* (2021)
15) Lyngby, Peder, and Kristian Sommer Thygesen: Data-driven discovery of novel 2D materials by deep generative models, *arXiv:2206.12159* (2022)