「マルチメディア，分散，協調とモバイル
(DICOMO2021)シンポジウム」 令和3年6月-7月

MatVAE：少量の実験データでも学習可能な 実験候補化合物を提示するための入れ子型変分オートエンコーダ

刑部好弘[†1] 淺原彰規[1]

概要：国際的に高性能かつ低環境負荷な材料の開発が急務であり、人工知能技術やデータ分析技術により材料開発効率化を図るマテリアルズ・インフォマティクス（MI）に注目が集まっている。特に、試作実験前に予め有望な材料を実験候補として提示し、要求性能を満たす新材料を短期間で見つけるため技術として、性能改善確度の高い化合物を生成する深層生成モデルがいくつか提案されている。しかし、一般的にこうしたモデルは十分な精度を得るために膨大な実験データを必要とする。現実の材料開発現場では多くても千件程度しか蓄積できないことが多く、少量の実験データでも適用可能な手法が必要だ。そこで報告者は、大規模オープンデータを用いて化合物の構造的特徴を学習する外側の VAE と、前記 VAE を介して得られる潜在変数と物性値の関係性を小規模実験データから学習する内側の VAE をネストして構成する深層生成モデル MatVAE を提案した。本研究では、学習データの範囲を超える高性能な材料を生成させるため、内側 VAE の学習時に潜在変数の所定の成分$z[k]$と物性値が強い相関をもつような損失関数$\mathcal{L}_{\mathrm{cor}}$を導入した。オープンデータによる検証では、候補生成時に$z[k]$を操作することで、生成化合物の物性値も制御可能であることが確認できた。さらに、過去の実験データを用いた模擬的新材料探索実験では、高性能な新材料の発見に要する実験回数を従来比 1/4 に削減した。

MatVAE: Nested Variational Autoencoders for Generating Compounds with Small Experimental Dataset

YOSHIHIRO OSAKABE[†1] AKINORI ASAHARA[1]

1. 序論

素材産業では，マテリアルズ・インフォマティクス（Materials Informatics，以下 MI）という，IT を用いて研究開発効率向上を目指す取り組みが積極的に推進されている．素材産業における技術革新は関連する広範な製造業にとっても重要であり，AI やビッグデータ分析技術の飛躍的進展からもそれを活用する MI に注目が集まっている．特に，過去の実験データや数値シミュレーションデータを活用することで，要求性能を満たす新材料を従来よりも短期間で見つけることができるようになると期待されている．

たとえば有機材料開発では，実験候補となる化合物を効率よく洗い出すために，バーチャルスクリーニング法が用いられる．これは，あらかじめ実験データ等で学習した機械学習モデルを用いて，大量の化学式に対し物性値を予測させてスクリーニングする方法である．予測精度が高いと無駄な実験をせずに済むので，実験回数を削減することができる。しかし、この方式には２つの問題点がある。

第一の問題は予測モデルの外挿性である。バーチャルスクリーニング法で一般に用いられる機械学習モデルは内挿型の予測モデルであり、あくまで学習データの範囲でしか有効な予測が立てられない。ところが、既知材料を凌駕する性能をもった新材料を探索するというタスクは、いわゆる外挿領域を予測する問題設定になる。したがって、本来発見したかった有望な化合物をむしろふるい落としてしまう可能性がある。

第二の問題は、スクリーニング対象の質である。そもそも良い化合物がリストアップされていなければ、スクリーニングの精度が高くても新材料発見に要する期間は短縮しにくい。これまで、このスクリーニング対象となる化合物を準備する「構造発生プロセス」は、単純な部分構造の組合せや材料開発者の熟練度などに依存しやすいボトルネックとなっていた。たとえば、既知化合物の部分構造を機械的につなぎ合わせて構造発生する BRICS[1]という手法では、自然界に存在しえない化学構造も大量に作られてしまい非効率である。

そこで近年，性能改善確度の高い化合物をより直接的に生成する、深層生成モデルが提案されている。たとえば、Gómez-Bombarelli らは変分オートエンコーダ（VAE）を用いて、文字列形式で与えられた化学式表現 SMILES[2]から、連続的な潜在表現を得る ChemicalVAE を提案した[3]。VAE の潜在表現は一般に実数ベクトルで表され、これを潜在ベクトルと呼ぶ。ChemicalVAE は、この潜在ベクトルから材料の物性値を予測するニューラルネットワークを VAE と

1 株式会社日立製作所 研究開発グループ
†1 yoshihiro.osakabe.fj@hitachi.com

は別に持ち、この予測器の出力も VAE の学習に利用する。この特徴により、ChemicalVAE の潜在空間は化学構造の類似度と物性値の類似度の両方が反映される。ChemicalVAE は、任意の潜在ベクトルを指定することで新しい化学式を生成できることに加え、いくつかの例については潜在空間を探索することで最適な物性値を持つ化合物を発見できることを示した。ところが，一般的に VAE のような深層生成モデルは、十分な性能を得るために数千～数十万規模の膨大な学習データを必要とする．しかし、化学式と性能指標値である物性値がペアになった実験データを大量に収集することは難しい。現実の材料開発現場では多くても百～千件程度しか蓄積できていないことが多く，少量のデータでも機能する手法が必要である．

そこで報告者は、独立に学習した 2 つの変分オートエンコーダ(VAE)をネストした深層生成モデル「MatVAE」を提案した[MLPS]。SMILES により表現された化学式を実数ベクトルに変換する外側の VAE と、そのベクトルと実験条件や物性値との関係性を学習し潜在表現に変換する内側の VAE を入れ子構造にすることで、望ましい物性値を持ちうる確度の高い候補化合物を生成することができる。オープンデータを用いた検証では、1000 件以下の実験データで学習をした場合でも、既存性能を超す物性値をもった化合物を ChamicalVAE と比較して 5 倍も多く生成できることを示した[MLPS]。ネットワーク構造と学習方式を工夫したことで少量の実験データでも性能を発揮できるようになったものの、生成時は既知化合物のうち最も性能の良い化合物の潜在空間における近傍を探索しているに過ぎない。潜在空間上の近傍であるからといって、必ずしも物性値の良い外挿領域の候補が生成できるわけではなく、さらなる効率化のためにも改善が求められる。

そこで本研究では、近傍探索よりも直接的に高性能な化合物を生成できる手法を提案する。具体的には、潜在ベクトルの所定の成分と物性値が強い相関を持つような損失関数を MatVAE に導入することで、物性値に対して線形に変化する次元を潜在空間につくり出す。もし理想的な潜在空間が学習によって獲得できた場合には、その所定の成分にバイアスを加えることによって、生成される化合物の物性値をバイアス量に応じて変化させられるようになることが期待できる。

2. VAE を用いた構造発生と提案手法

本章では、VAE を用いた有機化合物の構造発生について、既存手法である ChemicalVAE と、本稿で提案する MatVAE の手法について説明する。

2.1 既存手法（ChemicalVAE）の概要

生成モデルは、観測されたデータxが潜在変数zから生成されるという仮定に基づき、その変換規則$p(x|z)$を学習することで、学習したデータと類似するデータを生成する統計モデルである。特に深層ニューラルネットワークを用いて構成した生成モデルを深層生成モデルと呼び、変分オートエンコーダ（VAE）はその一つである。VAE は連続的な潜在変数zを多次元ガウス分布と仮定し、入力データxを潜在変数zに変換するエンコーダと、潜在変数zを入力データxに逆変換するデコーダを同時に学習する。十分に学習すると、実数ベクトルで与えられる任意の潜在変数zをデコードすることで、学習したxに類似する新たなサンプルx'を得ることできるようになる。

この VAE を化学式の構造発生に応用したものが ChemicalVAE である[3]。図 1 の模式図に示した通り、ChemicalVAE は SMILES 表記の化学式を one-hot エンコードした配列Hを学習し、潜在変数zを獲得する。さらに、潜在変数zから物性値を回帰する予測器も用意され、同時に訓練される。ChemicalVAE の学習に利用される損失関数$\mathcal{L}$には、VAE が One-hot ベクトルHを復元できているかを評価

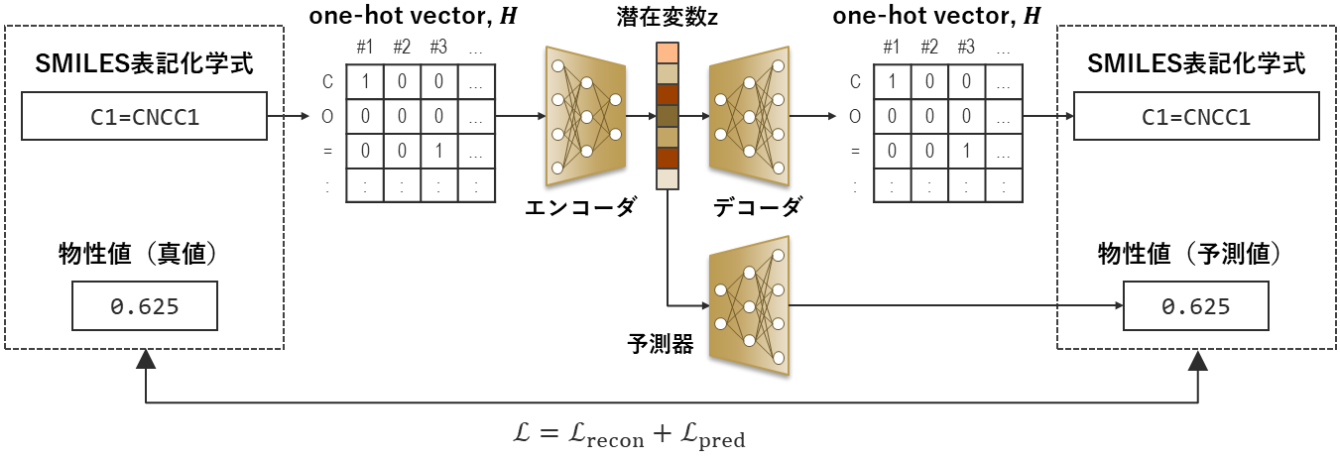

図 1　ChemicalVAE の概略図。

(Figure 1　Schematic design of ChemicalVAE. The pair of the encoder and decoder network is trained to generate chemical formula. The predictor network is trained to predict the property value of the compounds with the latent variable z.)

する再構成誤差関数$\mathcal{L}_{recon}$と、予測器の物性値予測誤差関数$\mathcal{L}_{\mathrm{pred}}$の和が用いられる。このようにして、One-hot ベクトルHに基づく化学構造の類似性を反映した潜在空間が得られるだけでなく、同時に訓練された予測器によって潜在変数zから精度の高い予測物性値を得ることができるようになる。ある種、バーチャルスクリーニングのための予測器を内包した深層生成モデルと理解することできる。しかし、予測器が十分な性能を発揮するためには、大量の実験データが必要になってしまう点が課題である。

2.2 提案手法（MatVAE）

実験候補材料となる化学式を提案する生成モデルを実現するために、VAE は 2 つの規則を学習しなければならない。一つ目は、実際に存在しうる化学構造の特徴である。SMILES に代表されるような文字列形式の化合物表現を学習データとする場合、元素記号列のパターンを学習することが主なタスクとなる。二つ目は、化合物の構造的特徴と物性値との関係性である。学習に大量の実験データを必要とするのは、この二つの学習対象を同一のモデルに獲得させようとしているからだと考えた。

そこで報告者は、これら二つの学習対象ごとに VAE を分離することを考え、MatVAE を提案した。このモデルは、オープンデータから取得した大量の化学式をもとに化合物の構造的特徴を潜在表現として獲得する VAE と，その潜在表現と物性値との関係性を少量の実験データから獲得する VAE の 2 種類を組み合わせるアプローチをとっている。前者の VAE は次元圧縮の機能も果たす。ゆえに、その潜在ベクトルの次元を十分に小さくすれば、後者の VAE は少量の実験データでも学習が可能である。本研究では、この MatVAE をさらに発展させ、損失関数に改良を施した。特に言及のない限り、以下では改良した後のモデルに関して

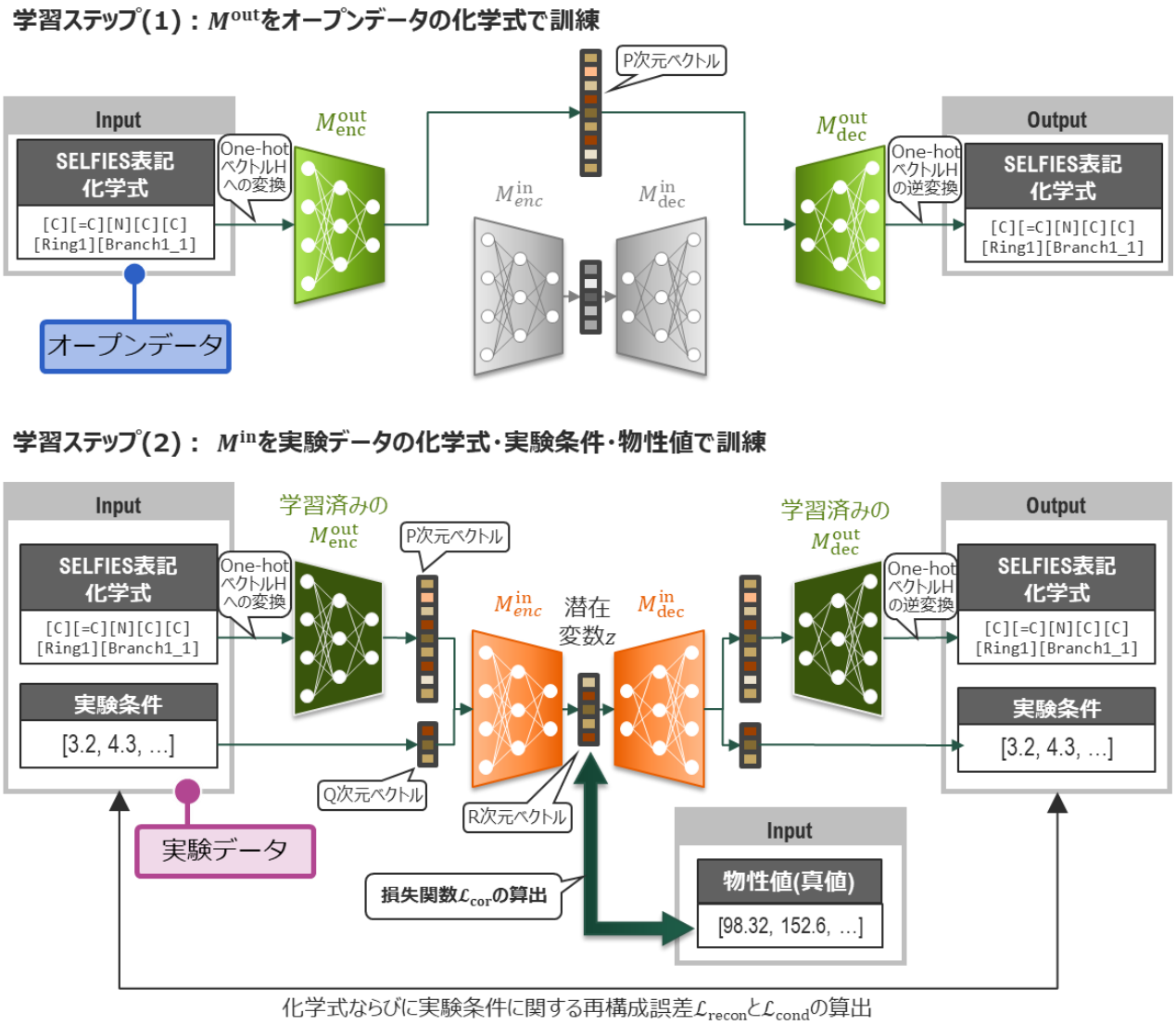

図 2　提案手法 MatVAE のモジュール構成模式図と学習ステップの概要。2 つの変分オートエンコーダ（VAE）は別々の学習データセットにより独立に訓練される。

(Figure 2　Schematic design of the proposed model, MatVAE, and the abstract of its learning steps. The two VAEs are trained independently over different datasets.)

の説明にとどめる。改良前の MatVAE については[4]を参照していただきたい。

図 2 に提案モデル MatVAE の模式図を示す。このモデルは、先に述べた通り二種類の VAE を入れ子型に組み合わせて構成される。MatVAE を構成している外側の VAE（Outer VAE，以下M^{out}）は，オープンデータから取得した大量の化学式をもとに化合物の構造的特徴を潜在表現として獲得する VAE である。そのエンコーダ$M^{\mathrm{out}}_{\mathrm{enc}}$は、数層の 1 次元畳み込みレイヤーと数層の全結合レイヤーで構成されている。$M^{\mathrm{out}}_{\mathrm{enc}}$は、文字列形式の化学式を one-hot エンコードと呼ばれる方法で変換した多次元配列Hを入力として受け取り、P 次元の潜在ベクトルを出力する。元素種（記号種）を M、化学式の文字列長を N としたとき、1 つの SELFIES 文字列は$M \times N$次元の配列になる。複数の化合物が与えられた場合、Hは 3 次元配列となる。一方、そのデコーダ$M^{\mathrm{out}}_{\mathrm{dec}}$は、複数の全結合レイヤーと Gated Recurrent Unit（GRU）レイヤーで構成されており，$M^{\mathrm{out}}_{\mathrm{enc}}$が出力した P 次元ベクトルを入力として受け取って、one-hot ベクトルHを出力する。

M^{out}のこうしたネットワーク構造は ChemicalVAE と概ね同一であるが、化学式の文字列表現には SMILES ではなく SELFIES[5]を用いた。SELFIES は SMILES と互換性がある表記法で，VAE や GAN などの生成モデルに使うことを想定して提案されたものである。SMILES では省略されることの多い副鎖や官能基などの分岐構造や，シス・トランス構造などの幾何異性体についても明示的に書き下すことを特徴とする。その分，文字列としては冗長になるが，非常にロバストな表現学習を可能にする点が強みである。たとえば，次に示す 2 つの表記はどちらも同じ化合物を示しており，一意に相互変換が可能である。

・SMILES：C1=CNCC1

・SELFIES：[C][=C][N][C][C][Ring1][Branch1_1]

他方、MatVAE を構成する内側の VAE（Inner VAE，以下M^{in}）は，M^{out}の潜在ベクトルである P 次元ベクトルと，Q 個の実験条件などの説明変数の配列からなる Q 次元ベクトルを結合した，(P+Q)次元ベクトルを入力として受け取る。ここでの実験条件とは，化学構造以外のファクター，たとえば，合成時の温度や配合量，触媒などに関するデータを指す。M^{in}はそのエンコーダ$M^{\mathrm{in}}_{\mathrm{enc}}$、デコーダ$M^{\mathrm{in}}_{\mathrm{dec}}$ともに複数の全結合レイヤーで構成され，その中間出力たる潜在変数zは，P+Q>R なる R 次元ベクトルである。

図 2 に示す通り、学習過程は 2 ステップに別れており、まず初めにM^{out}が訓練される。用いられる学習データは、オープンデータから抽出した数万から数十万件の化学式データセットである。$M^{\mathrm{out}}_{\mathrm{enc}}$に入力される one-hot ベクトル$H$と同じ多次元配列をデコーダが出力するように損失関数を与え，ニューラルネットワークの重み等パラメータを最適化する[6]。この仕組みは一般的な VAE と同じである。

次に、百から千程度の実験データによりM^{in}が訓練される。M^{out}の学習と同様に、$M^{\mathrm{in}}_{\mathrm{enc}}$が受け取る入力ベクトルと$M^{\mathrm{in}}_{\mathrm{dec}}$の出力ベクトルが等しくなるよう、再構成誤差が損失関数として与えられる[6]。ただし、M^{out}由来の P 次元ベクトル部分の損失関数$\mathcal{L}_{\mathrm{recon}}$と、実験条件由来の Q 次元ベクトル部分の損失関数$\mathcal{L}_{\mathrm{cond}}$は個別に算出される。さらに本研究では、潜在変数$z$の所定の成分が物性値と強く相関するような損失関数$\mathcal{L}_{\mathrm{cor}}$を提案する。

多くの新材料開発では，ある物性値について既存性能を凌駕するような新しい化合物の発見を目標としている。つまり，対象物性値に関し既知データの外挿領域に存在するような化合物を提案したい。そこで，M^{in}の R 次元潜在ベクトルの任意の成分$z[k]$（ただしkは$0 \leq k < R$なる整数）が物性値と強い相関をもつよう，次の通り損失関数$\mathcal{L}_{\mathrm{cor}}$を設計した。

$$\mathcal{L}_{\mathrm{cor}} = \sum_{b} \frac{z[k] \cdot \boldsymbol{y}^{b}}{\|z\|\|\boldsymbol{y}\|} \tag{1}$$

ここで$\boldsymbol{y}$は物性値を示す。たとえば、考慮している物性値が B 種類あるとき，b種目の物性値$\boldsymbol{y}^{b}$と潜在ベクトルの第 k 成分（$z[k]$）とのコサイン類似度によって定義される。この際，どの物性値とどの成分が相関をもつか，つまりbとkの対応はあらかじめ決定しておく。共分散が大きくなるほど相関は強くなるので、最終的なM^{in}の損失関数$\mathcal{L}$はαを負の値を取るハイパーパラメータとして次の式で与える。

$$\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{cond}} + \alpha\mathcal{L}_{\mathrm{cor}} \tag{2}$$

こうして学習したM^{out}とM^{in}を組み合わせ，生成モデルとして用いる。M^{out}との組合せによってM^{in}の潜在空間zは，化学構造の類似度と物性値の類似度が同時に距離として反映された空間になっている。さらに，$\mathcal{L}_{\mathrm{cor}}$の効果によって，所定の成分$z[k]$が物性値と相関をもつ。したがって，理想

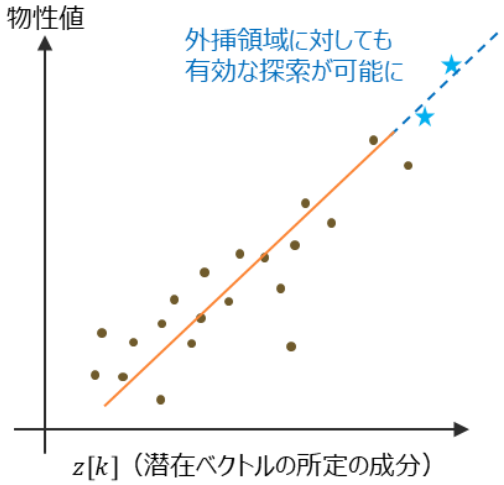

図 3　損失関数$\mathcal{L}_{\mathrm{cor}}$によって理想的な潜在空間を獲得した場合の外挿領域探索方法のイメージ。

(Figure 3　The image of the extrapolated domain search method when the ideal latent space is obtained by the proposed loss function $\mathcal{L}_{\mathrm{cor}}$.)

的には図 3 に示すような散布図を得ることができる。ここで、横軸が$z[k]$の値、縦軸が物性値である。丸印で示された点は学習データに含まれている点を表している。このような理想的な学習ができた場合には，$z[k]$を候補生成時のパラメータとして用いることができる。たとえば図 3 を例にすれば、既知データの傾向から、十分に大きな$z[k]$を与えれば星印で示されているような物性値をもつ化合物を生成することができるだろうと期待できる。したがって、既知化合物のなかでも特に高性能な化合物データをエンコードし、その$z[k]$にバイアスを加えてからデコードすることにより、性能改善確度の高い候補化合物を生成することができる。そしてこの手法によればは、従来の潜在空間における近傍サンプリングよりも、さらに直接的に高性能な化合物を与える潜在変数zを得ることができる。

3. 損失関数$\mathcal{L}_{\mathrm{cor}}$の効果検証実験

新たに提案した損失関数$\mathcal{L}_{\mathrm{cor}}$の効果を検証するための実験を行った。本来であれば、生成モデルが提案した化合物の物性値は実測を通じて評価すべきである。しかし、本研究では提案手法の情報科学的側面の検証を優先し、合成容易性スコア（Synthetic Accessibility Score、以下 SA Score）[7]を物性値の代替とみなして実験を行う。SA Score は合成の難易度を$\mathcal{S}=1$（易）から$\mathcal{S}=1$（難）の間の実数で定量化した指標値で、オープンソースソフトウェアの RDkit[8]を用い化学式のみから算出できる。

MatVAE は入れ子構造になっており、当然のことながら$\mathcal{L}_{\mathrm{cor}}$は$M^{\mathrm{out}}$の学習とは無関係である。したがって、$M^{\mathrm{out}}$を通じて最終的に出力される化学式の SA Score に$\mathcal{L}_{\mathrm{cor}}$がどの程度影響するかを直接調べることは難しい。そこで、本章の実験では、シンプルな 1 段の VAE を用いた。入れ子型の MatVAE に$\mathcal{L}_{\mathrm{cor}}$を適用した結果については次の章で述べる。

実験に利用したモデルのネットワーク構成を述べる。エンコーダM_{enc}は、3 層の 1 次元畳み込みレイヤーと、それに続く 2 層の全結合レイヤーで構成される。デコーダM_{dec}は、2 層の全結合レイヤーと、それに続く 3 層の GRU レイヤーで構成される。なお、GRU の隠れ層は 488 次元である。

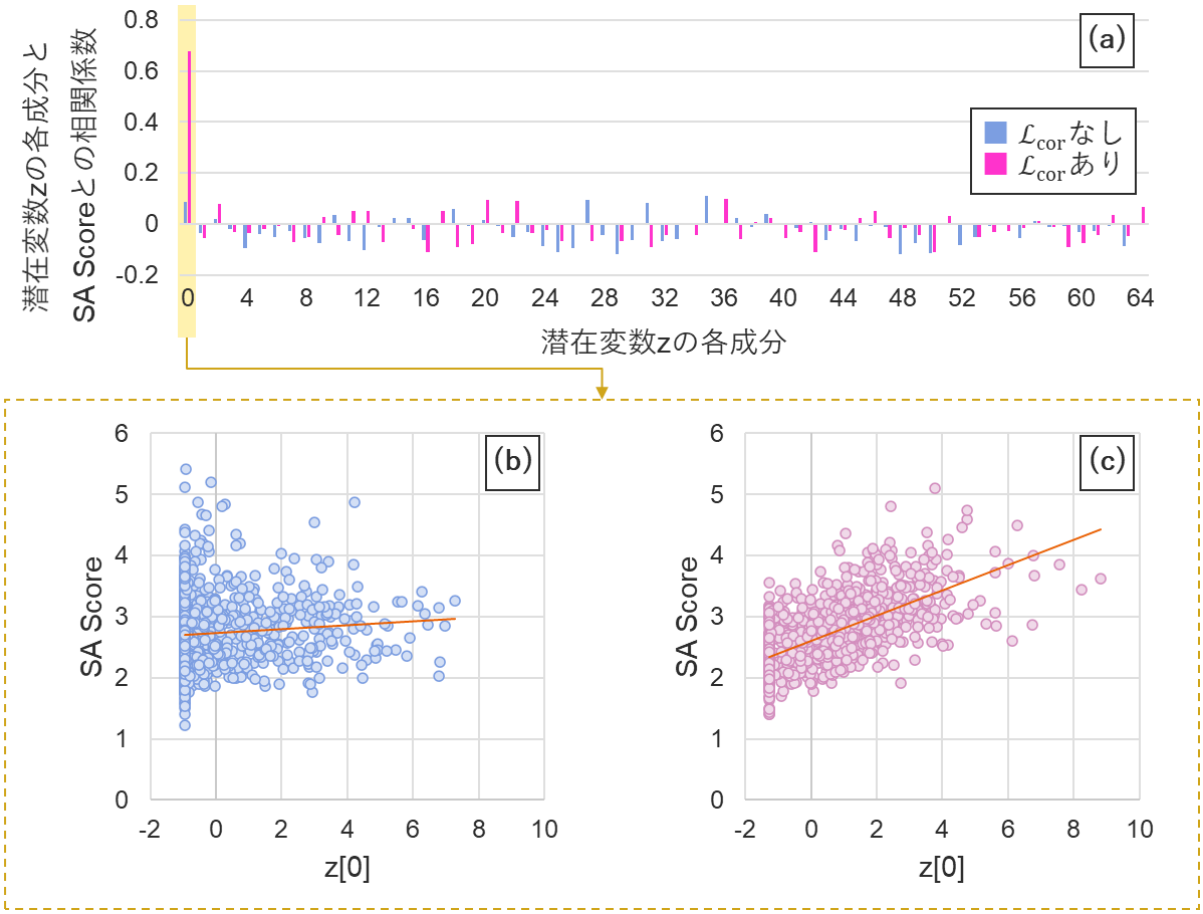

図 4　損失関数$\mathcal{L}_{\mathrm{cor}}$の効果を検証した実験結果。(a) 潜在変数の各成分についての SA Score に対する相関係数。(b) $\mathcal{L}_{\mathrm{cor}}$無しならびに (c) $\mathcal{L}_{\mathrm{cor}}$有りの場合における、潜在変数の第 0 成分の値と SA Score の散布図。

(Figure 4　The results for verifying the effectiveness of the loss function $\mathcal{L}_{\mathrm{cor}}$. (a) The correlation coefficient for SA Score for each component of a latent variable z. Scatter plot of the 0th component of the latent variable and SA Score in the case of (b) with and (c) without $\mathcal{L}_{\mathrm{cor}}$.)

SELFIES の記号種は 107 種、化学式の文字列長は最大で 96 だったため、One-hot エンコードして得られた多次元配列Hは$107 \times 96 \times (\text{batch size})$である。この多次元配列$H$から 65 次元の潜在変数$z$を得るよう構成し、その第 0 成分である$z[0]$と SA Score の相関が強くなるよう$\mathcal{L}_{\text{cor}}$を作用させた。

実験に用いたデータは全て ZINC[9]と呼ばれる市販化合物のオープンデータベースから無作為に抽出したデータセットを用いた。学習データは、オープンデータとみなすものと、実験データとみなすものの 2 種類を用意し、オープンデータで学習させたのち、さらに実験データで学習を行った。次の表 1 にデータセットの詳細を記す。

表 1　学習データセットの詳細

(Table 1　The details of the training data sets.)

役割	内容	データ数	条件
オープンデータ	SELFIES	500,000	なし
実験データ	SELFEIS SA Score	1,000	$2.5 < \mathcal{S} < 4.5$

損失関数$\mathcal{L}_{\text{cor}}$の効果を確かめるため、学習後のモデルについて潜在変数zの各成分と SA Score との相関を調べた。実験データセットに含まれている 1000 件分について算出した相関係数を図 4(a)に示す。$\alpha = 0$として学習した$\mathcal{L}_{\text{cor}}$無しの場合（青）と比べ、$\mathcal{L}_{\text{cor}}$を作用させた場合（マゼンタ）では、$z[0]$が SA Score とかなり強い相関を持っていることが分かる。残りの成分については、$\mathcal{L}_{\text{cor}}$の有無で相関係数の大きさがさほど変化していないことも確認できる。したがって、$\mathcal{L}_{\text{cor}}$は確かに狙い通り機能しているということが分かった。

次に、$\mathcal{L}_{\text{cor}}$によりもたらされた相関関係が候補化合物生成時に有用な関係として得られているのかを調べた。図 3 で示したような散布図を、$z[0]$と SA Score について描画した結果が図 4(b)ならびに(c)である。(b)$\mathcal{L}_{\text{cor}}$無しの場合と比べ、(c)$\mathcal{L}_{\text{cor}}$有りの場合の方が$z[0]$と SA Score との間に線形な関係があることが見て取れる。

実際、この$z[0]$を候補化合物生成時のパラメータとして用いることができるかを調べた。既知の化学式をM_{enc}に通して得た潜在変数の$z[0]$に正及び負のバイアスを一定値加え、バイアス印加後の潜在変数をM_{dec}に通して得た化学式の SA Score を算出した。正のバイアスを加えたときに元の SA Score よりも大きくなった場合、そして負のバイアスを加えたときに元の SA Score よりも小さくなった場合を成功ととらえ、1000 回試行したときの成功割合をプロットしたものが図 5(a)である。印加するバイアスの正負によって、生成される候補化合物の SA Score が選択的に変調できることが確認された。また、バイアスの絶対値に応じて成功割合も大きくなっていることも分かる。また、以上の結果から、物性値と相関をもたせた$z[0]$以外の成分は既存の化合物を変換して得た潜在変数の値をそのまま用いても問題ないことが示唆された。

最後に、新材料開発の最終的な目的である、既存性能を超える化合物を提案方式によって生成できるかを調べた。用意した実験データは、その SA Score が$2.5 < \mathcal{S} < 4.5$の範囲になるようフィルタリングしている。したがって、バイアスを加えたzから生成された化合物の SA Score がその範囲を超えていれば、外挿領域に存在する化合物を生成できたといえる。$z[0]$に正のバイアスを加え 1000 件の化合物を生成したとき、$\mathcal{S} > 4.5$なる SA Score をもつ化合物の生成割合を調べたのが図 5(b)である。このケースは非常に高い確

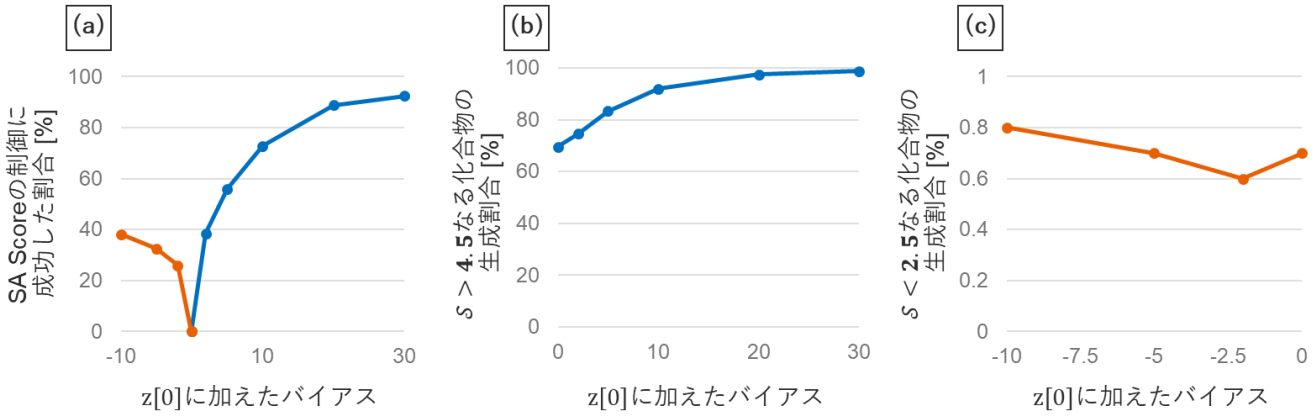

図 5　$z[0]$にバイアスを加えて生成した候補化合物の SA Score の評価結果。(a) 印加したバイアスによって生成化合物の SA Score の制御に成功した割合。(b) 正のバイアスを加えたとき、生成化合物の SA Score が外挿領域である 4.5 以上となった割合。(c) 負のバイアスを加えたとき、生成化合物の SA Score が外挿領域である 2.5 未満となった割合。

(Figure 5　The SA Score improvement rate of the candidate compound produced by biasing z [0]. (a) The rate of the successful control of the SA Score of the generated compounds by the bias to z [0]. (b) The rate at which the SA Score of the generated compound was higher than 4.5, which is the extrapolation region, when a positive bias was applied. (c) The rate at which the SA Score of the generated compound was lower than 2.5, which is the extrapolation region, when a negative bias was applied.)

率で、外挿領域の化合物の生成に成功している。他方、$z[0]$に負のバイアスを加え 1000 件の化合物を生成したとき、$\mathcal{S} < 2.5$なる SA Score をもつ化合物の生成割合を調べたのが図 5(c)である。こちらは非常に難易度が高く、1%を超えることがなかった。SA Score が大きくなるということは、より複雑な構造をもつということである。したがって、SELFIES の文法規則を満たす範囲で、より長く多様な文字列を出力すれば大きい SA Score を得ることはできる。一方、SA Score を小さくするためには、単純な炭素鎖のようにシンプルな化学式を出力しなければならない。今回の学習データは、市販されている工業薬品などからキュレーションしたため、工業的に価値の低い、単純な構造をもった低分子化合物はデータセットにそれほど多く含まれてはいなかった。こうしたことが影響して、図 5(b)と(c)のような結果の差異が生じたと考察できる。

以上の結果から、$\mathcal{L}_{\mathrm{cor}}$は期待通り指定した潜在変数の所定成分と物性値とが強く相関するように機能すること、そしてその所定成分にバイアスを加えることで生成される化合物の物性値も制御可能である見込みが得られた。こうして提案化合物の質が改善されれば、新材料発見に要する実験回数を削減でき、材料開発の効率化が図れる。実際、この損失関数$\mathcal{L}_{\mathrm{cor}}$を入れ子型の MatVAE に導入し、過去の実験データでの有効性を調べた。対象材料データが機密情報のため詳細結果の公開は控えるが、従来の XGBoost（XGB）モデルを用いたバーチャルスクリーニング法と比較し、実験回数を 1/4 に削減できるとの結果を得た。

4. 結論

本研究では、性能改善確度の高い化合物を実験候補として提示することで、材料開発期間を短縮するための深層生成モデルを検討した。報告者が以前提案した、学習データと学習対象の異なる 2 種類の VAE を入れ子にした構造を持つ深層生成モデル「MatVAE」を発展させ、その性能を検証した。特に、目的変数に対して強い相関をもつような潜在空間を構成する損失関数を考案したことで，潜在空間での近傍探索よりもさらに直接的に望ましい物性値をもつ候補化合物を生成できるようになった。オープンデータを用いた検証では、性能改善対象となる物性値と強い相関を持たせた潜在変数の所定の成分$z[k]$を候補生成時のパラメータとして用いることの妥当性が確認された。実際、過去の実験データを用いた検証でも、高性能な化合物を提示しうることを確認した。XGB モデルに基づくバーチャルスクリーニング法と比較し，高性能材料の発見に要する実験回数を 1/4 に削減する効果があることが確認できた。

参考文献

[1] Degen, J. et al.: On the art of compiling and using “drug-like” chemical fragment spaces. ChemMedChem, Vol.3, No.10, pp.1503–1507 (2008).

[2] Weininger, D.: SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. Journal of Chem. Inf. and Comput. Sci, Vol.28, No.1, pp.31–36 (1988).

[3] Gómez-Bombarelli, R. et al.: Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science, Vol.4, No.2, pp.268–276 (2018).

[4] Osakabe, Y., & Asahara, A.: MatVAE: Independently Trained Nested Variational Autoencoder for Generating Chemical Structural Formula. AAAI Spring Symposium 2021: MLPS (to be published) (2021).

[5] Krenn, M. et al.: Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. Machine Learning: Science and Technology (2020).

[6] Kingma, D. P., Welling, M.: Auto-Encoding Variational Bayes. (2013). https://arxiv.org/abs/1312.6114

[7] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics, Vol.1, No.1, pp.8 (2009).

[8] Landrum, G.: Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. (2013). https://rdkit.org

[9] Sterling, T., & Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. Journal of Chemical Information and Modeling, Vol.55, No.11, pp.2324–2337 (2015). https://zinc15.docking.org