The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

固体系材料インフォマティクスのための畳み込みニューラルネットを活用する3Dボクセルデータ記述子

A universal 3D voxel descriptor for solid-state materials informatics with convolutional neural networks

梶田晴司　大庭伸子　旭良司
Seiji Kajita　Nobuko Ohba　Ryoji Asahi

株式会社 豊田中央研究所
Toyota Central R&D Labs., INC.

Materials informatics (MI) is a promising approach to liberate us from the time-consuming trial and error process for material discoveries. Contrary to molecular systems, however, practical successes of the solid-state MI are very scarce because existing descriptors insufficiently describe 3D features of field quantities (e.g., electron distributions and local potentials). We develop a simple, generic 3D voxel descriptor that compacts any field quantities, in such a suitable way to implement convolutional neural networks. We examine the reciprocal-lattice 3D voxel space descriptor encoded from the electron distribution by a regression task with 680 oxides data. The present scheme outperforms other descriptors in the prediction of Hartree energies that are significantly relevant to the long-wavelength distribution of the valence electrons.

1. はじめに

材料探索は研究者のカンと経験で実施されてきたが，材料インフォマティクス (materials informatics; 以下 MI) のアプローチではデータを活用した超高効率な材料探索が期待されている．機械学習を適用するには物質の特徴をなんらかのプロトコルで数値ベクトルに変換する必要がある．この数値ベクトルは記述子と呼ばれ，機械学習アプローチの成否を決める重要なものである．

分子系の MI は創薬を中心に発展が進んでいるが，固体の MI はいまだ緒に就いた段階である．固体の MI が難しい理由は，電子密度分布やポテンシャルといった場の量を表現する記述子が限られるためである．物質は点で表される原子核と，連続的な広がりを持つ電子で構成されている．原子核に着目した点描像の記述子として coulomb matrix(CM)[1, 2, 3] や smooth overlap of atomic positions(SOAP)[4, 5, 6] が提案されている一方，場の量を表現する記述子は少ない．電子が局在的に分布する分子系と異なり，電子の広がりが固体特性の本質的な起源になる場合が多いため，場の量を取り扱う記述子は固体系でより重要となる．

そこで我々は，結晶内の場の量を表現する逆格子 3 次元ボクセル空間 (reciprocal-lattice 3D voxel space: 以下 R3DVS) 記述子を開発した [7]．そして 680 種類の酸化物の電子分布から R3DVS 記述子を生成し，それを畳み込みニューラルネット (convolutional neural networks: 以下 CNNs) の学習データに使用してエネルギーに関連する物理量を回帰する．SOAP と CM 記述子を用いた機械学習とのベンチマークをとり，R3DVS 記述子の特徴や回帰精度を包括的に評価する．

2. 関連研究

記述子には大きく分けて 2 つのカテゴリがある．1 つは原子番号や電気陰性度，バンドギャップ，密度，擬ポテンシャルのコア半径などの物性量をもとに，その分野に精通した研究者が良い記述子をデザインする方策である [8, 9, 10, 11, 12, 13, 14]．このような発見的な記述子はしばしば「ハンドクラフト記述子」と呼ばれ，古くは 1960 年代から実施されており現在でも多くの成功を収めている．

2 つ目のカテゴリは物質の特徴を理論的に数値ベクトルへ射影する方策で，便宜上，本研究では「理論的記述子」と呼ぶ．ハンドクラフト記述子と違い，理論的記述子は計算式にそって一意的に作られるため，使用者の資質に依存せずに様々な問題へ適用できる柔軟性がある．このカテゴリに分類される手法には，原子隣接密度の類似度を記述する SOAP，構成原子核間のクーロンポテンシャルを行列表示した CM，動径対称関数による構成原子の局所構造の表現 [15, 16, 17]，原子の分布関数のフーリエ変換やウェーブレット変換による表現 [18, 19]，などの記述子が提案がされている．

さて記述子にはいくつかの条件が要求される．物質に並進と回転，原子ラベルの交換操作を施してもその性質は変わらないことに起因し，記述子も同様の操作で不変であることが求められる．例えば角度方向の積分をした原子の動径分布関数を使って回転不変性を取り入れる工夫があるが [18]，当然この平均化により固体構造の角度情報は犠牲になる．固体系の記述子はさらに厄介であり，周期性に起因するユニットセル選択の不変性が追加される [7]．固体系かつ単一元素に限定されない理論的記述子として SOAP や CM が挙げられるが，種類は非常に少ない．

3. 逆格子 3D ボクセル空間 (R3DVS) 記述子

ある結晶のユニットセルの基本並進ベクトルを $\mathbf{a}_i, i = 1, 2, 3$ とおき，その逆格子の基本並進ベクトルを $\mathbf{b}_i$ とおく．これらは $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$ の関係をもつ．結晶内に分布する場の関数 $s(\mathbf{r})$ のフーリエ係数の絶対値は，

$$|s(\mathbf{g})| = \frac{1}{v_c}\left|\int_{v_c} \exp(-\mathrm{i}\mathbf{r}\cdot\mathbf{g})s(\mathbf{r})d\mathbf{r}\right| \quad (1)$$

である．ここで v_c はユニットセルの体積，$\mathbf{g}$ は逆格子ベクトルを表す．さて通常の計算，例えば第一原理計算で得られる場のデータは $\mathbf{r}$ に関して連続ではなく，$\mathbf{r} = (m_1/M_1)\mathbf{a}_1 + (m_2/M_2)\mathbf{a}_2 + (m_3/M_3)\mathbf{a}_3$ で区切られたボクセルデータである．それに対応

連絡先: 梶田晴司，〒 480-1192 愛知県長久手市横道 41-1 豊田中央研究所、fine-controller@mosk.tytlabs.co.jp

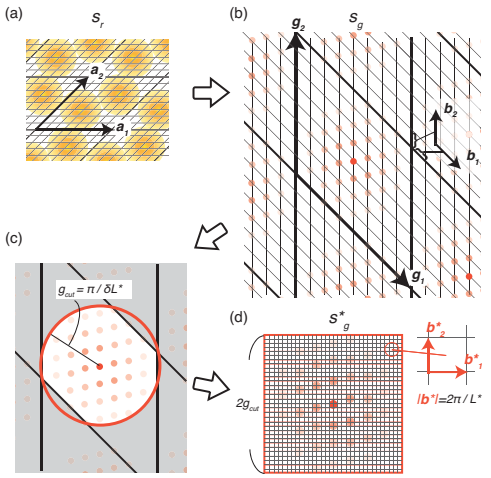

図 1: R3DVS 記述子作成の模式図

し逆格子ベクトルも同様に，$\mathbf{g} = m'_1\mathbf{b}_1 + m'_2\mathbf{b}_2 + m'_3\mathbf{b}_3$ となる．ここで整数 m_i, m'_i はボクセルのインデックス番号，M_i はインデックスの最大数を表し，$0 \leq m_i, m'_i < M_i$ である．離散化された場の関数を $s(\mathbf{r}) \sim s_r(m_1, m_2, m_3) = s_r(\mathbf{m})$，および $s(\mathbf{g}) \sim s_g(m'_1, m'_2, m'_3) = s_g(\mathbf{m}')$ と記す．式 (1) は以下のように離散化される．

$$|s_g(\mathbf{m}')| = \frac{\Delta}{v_c}\left|\sum_{\mathbf{m}} \exp[-2\pi \mathrm{i} \sum_{i=1}^{3} m_i m'_i / M_i] s_r(\mathbf{m})\right| \qquad (2)$$

ここで Δ はボクセルの体積である．

R3DVS 記述子のアルゴリズムを以下に記す．

1. 式 (2) に従いボクセルデータ s_r から s_g を生成する (図 1(a), (b))．
2. 逆格子空間の原点 $\mathbf{g} = 0$ を中心に半径 g_{cut} の球でくり抜く (図 1(c))．さらにその球に外接する立方体で s_g を囲み，その立方体に合わせて逆格子ベクトル $\mathbf{b}^*_i$ を取り直す．
3. $\mathbf{b}^*_i$ で規定されるボクセルメッシュに s_g の値を再配置させて s^*_g を作る (図 1(d))．

以上のプロトコルから得られる $|s^*_g|$ が R3DVS 記述子である．この記述子は構造の異なる結晶であっても場の関数を一定の立方格子ボクセル数で規格化できるため，CNNs といった機械学習手法の入力データとして使用できる．

本記述子の生成で使用するパラメータは，$g_{cut} = \pi/\delta L^*$ を定める δL^* と，$|\mathbf{b}^*_i| = 2\pi/L^*$ の L^* である．δL^* を小さく，L^* は大きくとれば記述子の精度は向上するが，ボクセル数が多くなるため記述子のデータ量と CNNs の学習時間が増加する．本計算では $\delta L^* = 0.4$ Å, $L^* = 12.8$Å に設定した．このパラメータによる R3DVS のボクセル数は 32^3 である．

4. 計算条件

4.1 データ

無機結晶構造データベース (Inorganic crystal structure database: ICSD) に登録されている酸化物の中で，構造が既知

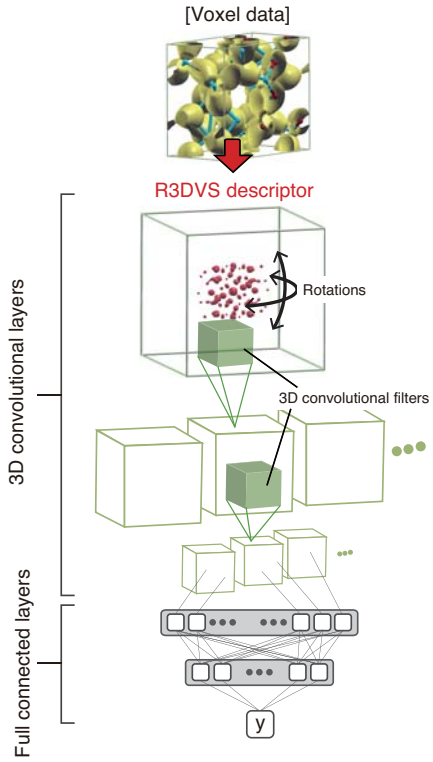

図 2: R3DVS 記述子と 3D CNNs 学習モデルの模式図

かつユニットセルに含まれる原子の数が 50 以下のものを 680 種選択した．この酸化物に対し密度汎関数理論にもとづく第一原理計算を VASP[20] で実施した．固体の全エネルギー E を構成する以下のエネルギー項を目的変数として使用する．

$$E = \sum_i \epsilon_i - E_H + \Delta E_{xc} + E_I$$

$$\Delta E_{xc} = E_{xc} - \int v_{xc}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \qquad (3)$$

ここで ϵ_i，E_H，E_{xc}，v_{xc} および E_I はそれぞれ電子の i 番目の 1 電子軌道エネルギー，ハートリーエネルギー (古典的な電子-電子静電エネルギー)，交換・相関エネルギー，交換・相関ポテンシャル，原子核同士の静電エネルギーを表す [21]．$\rho(\mathbf{r})$ は電子密度分布である．これらエネルギー項に加え，固体の凝集エネルギーとバンドギャップも目的変数として採用する．

4.2 学習モデル

酸化物の価電子密度分布から R3DVS 記述子を生成し，図 2 に示す CNNs の入力データに用いた．回転不変性を CNNs に獲得させるため，ランダムに 3 次元回転させた R3DVS データを 30 枚生成して入力データを増やした．実装は python ライブラリ keras を用い [22]，バックエンドは tensorflow を使用した [23]．畳み込み層のフィルタ数は 16 枚，ボクセルフィルタのサイズは 3^3 に設定した．ネットワーク構造の詳細は文献 [7] 参照のこと。

SOAP 記述子はそれ自体をカーネルに用いた SOAP カーネルリッジ，CM 記述子に使用する回帰モデルはガウスカーネルリッジを採用し，それぞれの正則化パラメータは 3.0 および 0.01 と設定した．これらのカーネル回帰モデルは scikit-learn を使用して実装した [24]．

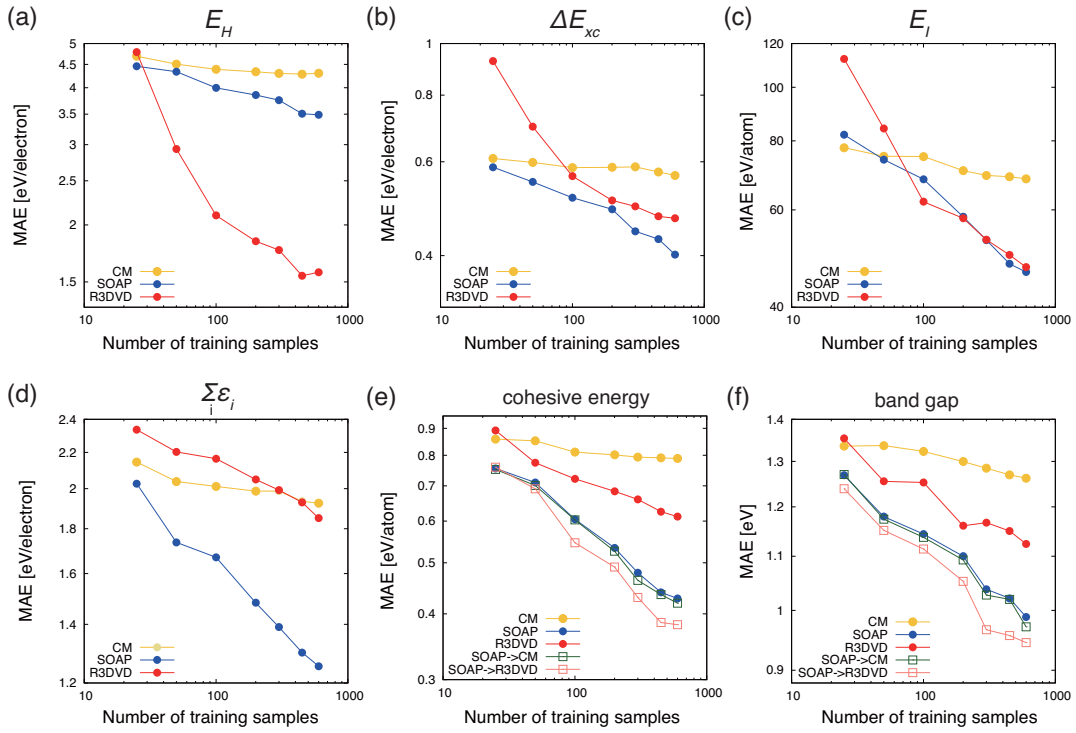

図 3: R3DVS と SOAP，CM 記述子による (a) ハートリーエネルギー (古典的な電子-電子の静電相互作用)，(b) 交換-相関相互作用項，(c) 原子核の静電エネルギー，(d) 一電子軌道エネルギーの和，(e) 凝集エネルギー (f) バンドギャップの回帰の平均絶対誤差 (Mean absolute errors; MAE).

5. 結果

680 種類の酸化物の中から 80 種をランダムサンプルしたものをテストデータとし，残りをトレーニング標本として用いた．CNNs の初期値依存性の影響を抑えるため学習は 5 回実施し，それぞれの予測値の平均を取る．この回帰テストを 20 回実施し，その回帰誤差の平均を図 3 に示す．

トレーニング標本数が増えるほど R3DVS, SOAP, CM 記述子による回帰性能は向上する．R3DVS 記述子は特にハートリーエネルギー (図 3(a)) の回帰性能が他よりも優れている．この理由は，R3DVS 記述子が電子密度分布から生成されており，電子間の静電エネルギーの予測に適するためである．交換・相関エネルギー項 ΔE_{xc}(図 3(b)) や原子核の静電エネルギー E_I(図 3(c)) は SOAP 記述子による回帰性能とほぼ同等だが，1 電子軌道エネルギーの和 $\sum_i \epsilon_i$(図 3(d)) の回帰性能は悪い．精度が悪化した原因は，1 電子軌道エネルギーには電子の運動エネルギーおよび電子と原子核との相互作用エネルギーが含まれているため，R3DVS 記述子が有する電子密度分布の情報だけでは表現できないためと考えられる．凝集エネルギー (図 3(e)) とエネルギーギャップ (図 3(f)) が SOAP 回帰よりも精度が悪いのも同様の理由と考えられる．

次に，複数の記述子を用いることで回帰精度が向上するか検討する．たとえ電子局在性が強いイオン結晶であっても実際の電子はある程度の広がりをもつため，点描像の SOAP はこの電子の広がりを記述するには表現力不足である．逆に R3DVS は電子密度分布を良く記述するが，局在性が強い原子核の存在は無視されている．この洞察に立脚し，点描像と場の描像を両方とり込むように記述子を同時使用する．

まずトレーニング標本を用いて SOAP 記述子による回帰モデル $\tilde{y}_{SOAP}$ を作り，トレーニング標本の目的変数 $y_{correct}$ と $\tilde{y}_{SOAP}$ との誤差 $\Delta y = y_{correct} - \tilde{y}_{SOAP}$ を記録しておく．次に Δy を目的変数として R3DVS 記述子を使った CNNs に学習させ，回帰モデル $\Delta\tilde{y}_{R3DVS}$ を作る．そしてテストで使用する予測値を $\tilde{y} = \tilde{y}_{SOAP} + \Delta\tilde{y}_{R3DVS}$ により算出する．これは SOAP 回帰を目的変数の主要項とみなし，R3DVS による CNNs 回帰で摂動項を表現する方策である．結果を図 3(e) と (f) の SOAP→R3DVS の凡例で記されるプロットで示す．期待通り凝集エネルギーとバンドギャップの回帰性能は，SOAP 回帰単独のものより向上する．比較のため SOAP 回帰と CM 回帰の組み合わせも検討したところ（図 3(e) と (f) の SOAP→CM の凡例）回帰性能にほとんど改善が見られなかった．以上より，点描像の SOAP と場の描像の R3DVS を併用すると相補的に作用し予測精度が向上することが確かめられた．

6. 結論

電子密度分布から生成した R3DVS 記述子は、特にハートリー項 (電子密度分布による静電エネルギー) に対し回帰性能が高かった．一方，電子の運動エネルギーや原子核ポテンシャルとの相互作用エネルギーが含まれる 1 電子軌道エネルギーの回帰性能は悪い．密度汎関数法によれば電子密度分布から電子の運動エネルギーは決定されるとはいえ，その運動エネルギー汎関数は未だ解析解が見つかっていない．また，原子核のポテンシャルもまた電子密度分布から導くことが難しい量である．つまり CNNs にとって電子密度分布の情報だけをもとに，1 電子軌道エネルギーを精度よく予測するのは難易度が高すぎると考えられる．原子を点と捉える SOAP と場の描像の R3DVS 記述子を併用すると予測精度は向上した．今回の結果

を踏まえると，点描像と場の描像の記述子どちらが優れているかではなく，互いに相補的な情報であるため，両記述子を共存させた予測モデルを作成することが肝要と結論される．

本検討では R3DVS の検証という立場のため，第一原理計算のエネルギー項を目的変数とした．しかし MI の実務上では計算が困難な物性値，例えば輸送係数などを予測するといった使い方が本来のものであろう．R3DVS 記述子により固体系の場の量が表現できるようになったため，様々な物性値の予測が可能となるだろう．

謝辞

SOAP のプログラムを提供いただいた、豊田中央研究所の陣内亮介博士に感謝する．

参考文献

[1] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108** 058301 (2012).

[2] K. Hansen, et al, J. Chem. Theory Comput. **9** 3404 (2013).

[3] F. Faber, A. Lindmaa, O. A. von lilienfeld, and R. Armiento, Int. J. Quant. Chem. **115** 1094 (2015).

[4] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87** 184115 (2013).

[5] W. J. Szlachta, A. P. Bartók, and G. Csányi, Phys. Rev. B **90** 104108 (2014).

[6] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, Phys. Chem. Chem. Phys. **18** 13754 (2016).

[7] S. Kajita, N. Ohba, R. Jinnouchi, R. Asahi, Sci. Rep. **7** 16991 (2017).

[8] J. A. Van Vechten, Phys. Rev. **182** 891 (1969).

[9] A. Zunger, Phys. Rev. B **22** 5839 (1980).

[10] P. Villars, et al., J. Alloys Comp. **317-318** 26-38 (2001).

[11] P. Villars, et al., J. Alloys Comp. **367** 167-175 (2004).

[12] A. Seko, et al., Phys. Rev. Lett. **115** 205901 (2015).

[13] A. D. Sendek, et al., Energy Environ Sci. **10**(1), 306-320 (2017).

[14] F. A. Faber, A Lindmaa, O. A von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117** 135502 (2016).

[15] J. Behler and M. Parrinello, Phys. Rev. Lett. **98** 146401 (2007).

[16] J. Behler, Int. J Quant. Chem. **115** 1032 (2015).

[17] N. Artrith, and A. Urban, Comp. Mat. Sci. **114** 135 (2016).

[18] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, Int. J Quant. Chem. **115** 1084 (2015).

[19] M. Hirn, S. Mallat, N. Poilver, Multiscale Model Simul. **15**(2) 827 (2017).

[20] G. Kresse and J. Furthmüller, Phys. Rev. B **54** 11169 (1996).

[21] R. G. Parr and W. Yang, "Density-Functional Theory of Atoms and Molecules", Oxford University Press (1989).

[22] https://keras.io

[23] https://www.tensorflow.org

[24] http://scikit-learn.org/stable