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

AIによる新物質探索手法

Material discovery by AI

武田 征士[*1] Seiji Takeda
Hsiang Han Hsu[*1] Hsiang Han Hsu
濱 利行[*1] Toshiyuki Hama
山根 敏志[*1] Toshiyuki Yamane
益田 幸治[*1] Koji Masuda
中野 大樹[*1] Daiju Nakano

[*1]IBM 東京基礎研究所
IBM Research - Tokyo

Discovering new materials that possess on-demand properties is the central demand in every industrial domain. We constructed the first material discovery system with end-to-end pipeline consisting of several technical pieces; feature encoding, regression, solution search, and structure generation. Those pieces are coordinated to coherently work together by newly defining two kinds of feature vectors; data-driven feature and pre-defined feature, and developing an algorithm to generate molecular structures by using those feature vectors. The capability of the system to discover new small organic molecules is demonstrated by a public dataset of commercial drugs.

1. 序論

あらゆる産業は，新物質の発見によってこそもたらされる非連続な飛躍を繰り返すことで，今日にみる巨大な発展を遂げてきた．航空宇宙産業，自動産業など重工業から，医薬・バイオ産業，ナノテクノロジー産業，情報産業に至るまで，任意の製品性能を実現するためには，その性能の要求する材料物性（たとえば機械的・熱的・電気的特性など）を備えた物質を新規かつ高速に発見・開発し，製品に取り込むことが極めて重要である．

今日の一般的な材料（物質）開発のプロセスは，技術者の知識や直観に基づき，実験やスーパーコンピュータによる物理化学計算を繰り返す伝統的な試行錯誤に依る部分が大きい．しかしながら物質の探索空間は極めて膨大であり，化学構造のトポロジーや用いる原子の種類に強い制約を設けても，10^{60} 種類を優に上回る未知の物質が存在し得るとされている[Bohacek 96]．このため，新材料開発には10年から20年ほどの期間を要するというのが一般的な通念である．加えて，得られる新材料も，各材料分野固有の知見に基づいて開発・デザインされるため，既存材料から大きく離れた構造や特性を備えた物質が得られることは困難であった．

このような人間の知見や発想による制約を離れ，物質デザインの自由度を飛躍的に高め開発速度を向上させるために，近年 AI（機械学習・統計）を活用した物質探索の研究開発が加速している．AI を物質探索に取り込むアプローチは，一般にケモインフォマティクス，あるいはマテリアルズ・インフォマティクスとも呼ばれる．AI による物質探索の目標は，所望の物性値を持ち得る化学構造をデザインすることである．ここで対象とする物質を有機化合物（薬品，ポリマー，発光材料など）とすると，解くべき課題は，化合物の分子構造，特徴ベクトル，物性値の3要素を双方向的に繋ぐシステムを設計することである（図1）．すなわち，化合物を特徴ベクトルにエンコードし，それを用いて物性値を予測する「順問題」と，所望の物性値が得られる特徴ベクトル候補を出力し，その特徴ベクトルから分子構造を列挙する「逆問題」の両方を達成すべく，各要素間を繋ぐアルゴリズムを適切に設計する必要がある．特に，特徴ベクトルへのエンコード手法と分子構造へのデコード（構造生成）手法の設計が重要である．

これまで報告されてきた多くのアプローチは，順問題のみにフォーカスしたものや，化合物の形状に強い制約を設けた上での逆問題の解法などであった．本研究では，分子構造の部分構造に着目することで，順問題を十分な精度を以って解き，かつ逆問題にも対応可能なアルゴリズムを設計した．

連絡先： 武田征士，IBM 東京基礎研究所，
seijitkd@jp.ibm.com

2. 関連研究

2.1 バーチャル・スクリーニング

順問題のみを対象とする研究は，多くの場合バーチャル・スクリーニングと呼ばれる物質探索手法を想定している．これは，高精度の物性値予測モデルに，物性値が未知の多様な分子構造を数万～数千万個単位で投入し，所望の値に近い物性値を持つ物質を選別（スクリーニング）する手法である．高い予測精度が要求されるため，特徴ベクトルには，部分構造の出現を one-hot エンコードした Morgan Fingerprint [Pyzer-knapp 15a] や，原子間クーロン相互作用をエンコードした Coulomb Matrix [Rupp 12]が用いられる．また多くの場合，大量の化合物をディープラーニングにより学習する[Pyzer-knapp 15b]．学習後のモデルに投入する新規化合物は，部分構造の網羅的な組み上げや，遺伝的アルゴリズムにより生成され[Vership 13]，これら仮想的な化合物の集合をバーチャル・ライブラリーと呼ぶ．

本手法はシンプルながらも強力で，高効率な有機 LED 材料の発見が報告されている[G.-Bombarelli 16]．一方で，試行錯誤の AI 版ともいえる網羅性ゆえに，バーチャル・ライブラリー生成やディープラーニングにかかる計算コストが大きい．

2.2 生成モデル

変分オートエンコーダによる分子構造の生成モデルが近年提案された[G.-Bombarelli 18]．あらゆる分子のトポロジカルな構造は，SMILES（Simplified Molecular Input Linear Entry System）という文法に基づき，文字列による表記が可能である． G.-BomBarelliらは，SMILESを入力とし，SMILESおよびそれに帰

属する物性値を出力とする，数10万個の化合物により学習させた変分オートエンコーダを構成した．ガウス過程により，所望の物性値に相当し得る潜在空間上の点を選択し，それをSMILESにデコードすることで，候補化合物を得るという手法である．本手法はディープラーニングによる生成モデルの，分子構造への最初の適用事例であり大きな注目を集めたが，SMILES 文法が破綻した文字列を多く出力するという問題があるため，文法生成ルールを適用するなど改良の途上にある[Kusner 17].

2.3 部分構造の連結生成

順問題と逆問題の両方を扱う直感的な手法として，部分構造を部品と見なし連結することで化学構造を組み上げる(生成する)アプローチがある．[Arakawa 05] および[Huan 15] では，事前に定義した構造ユニットの個数を特徴ベクトルとして予測モデルを構築し，逆問題求解時にはそれらユニットを連結することで新構造を生成する．この手法は構造ユニットが事前に定義されているため，それを含まない物質に対応できず，また組み上げのパターンすなわち部分構造同士の連結ルールが決まっているため，限定的な構造しか生成できないという課題がある．

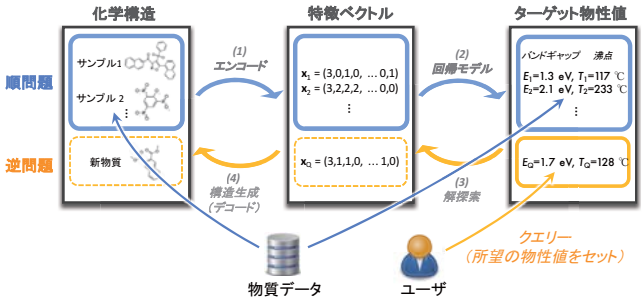

図1： 本提案手法の概要図．化学構造からターゲット物性値を予測する「順問題」と，所望の物性値を満たす新規な化学構造をデザインする逆問題を解く必要がある．

3. 提案手法

提案手法の概要を図1に示す．本手法の主な特徴は以下の2点である:

- 部分構造の個数カウントからなるデータ駆動特徴量と，構造生成に必要な事前定義特徴量の2種類を提案し，それらを同時に用いることで，順問題と逆問題を解くことを可能とした．
- 事前定義特徴量を基に分子構造を生成し，かつ部分構造を変形させながらデータ駆動特徴量によりフィルタする生成手法を提案することで，効率的な構造生成を可能とした．

3.1 特徴エンコード

分子構造の物質探索に用いる特徴量は，以下の性質を兼ね備えている必要がある:(1)分子構造の特徴を十分に捉え，高精度の予測モデルが作成可能であること，(2) 与えられた物質データ集合により柔軟に定義されること(データ駆動であること)，(3)それを用いて化学構造を広く逆生成できること，(4) 専門家が中身を理解できること．これらを満たす特徴量を，以下に定義する．

3.1.1 データ駆動特徴量

分子構造をノード(原子)とエッジ(結合)からなるグラフとみなしたとき，部分構造とは部分グラフと同義である．既存物質として与えられた分子構造 m_n の集合を $\mathcal{M} = \{m_1, m_2, \dots m_N\}$ とし，分子構造 m_n を構成する部分構造の集合を $S^{(n)}=\{s_1, s_2, \cdots\}$ とする．ただし，$S^{(n)}$は m_n を構成するあらゆる部分構造(最小は原子，最大は m_n 自体)からなる．集合$\mathcal{M}$の全ての要素に対して，網羅的に抽出された部分構造の集合は，$\mathcal{S}^{Full} = \bigcup_{i=1}^{N} \mathcal{S}^{(i)}$ で得られる．ここで，m に含まれる部分構造 s の個数を $N_D(m, s)$とし，分子構造 m_nを表現する特徴 $\boldsymbol{x}_D$を以下で定義する．

$$\boldsymbol{x}_{\mathrm{D}}^{(n)} \coloneqq \left(N_D\left(m_n, \mathcal{S}_1^{Full}\right), N_D\left(m_n, \mathcal{S}_2^{Full}\right), \dots\right) \qquad (1)$$

$\boldsymbol{x}_D$ は出現頻度の低い部分構造カウントを多く含み冗長であるため，特徴選択を行う．本手法では目標物性値 y への LASSO 回帰: $\mathcal{L}: \boldsymbol{x}_D \mapsto y$ により特徴選択し，$\boldsymbol{x}_D^{Select}$を得る．これをデータ駆動特徴量と呼ぶ．

3.1.2 事前定義特徴量

上記により得られたデータ駆動特徴量は，既に順問題に対応出来ているが，これらの情報のみでは具体的な分子構造を生成できない．なぜなら部分構造同士の連結や包含関係に関する情報を持たないからである．そこで，構造生成可能な特徴量を事前に定義し，データ駆動特徴量は，生成された構造から尤もらしいものを選別するフィルタリングに利用する．

構造の骨格生成に必要な情報として，x_{HA}, x_5, x_6(重原子，5員環，6員環の個数)，原子情報として，x_A (原子 A の個数)，結合情報として，$x_{||}$, $x_{|||}$, x_{arom} (二重結合，三重結合，芳香環の個数)を要素にもつ特徴量 $\boldsymbol{x}_P$を定義し，これを事前定義特徴量と呼ぶ．

データ駆動特徴量と事前定義特徴量を連結した以下のベクトルを，分子構造を表現する特徴ベクトルと定義する．

$$\boldsymbol{x} = (\boldsymbol{x}_D^{Sel}, \boldsymbol{x}_P) \qquad (2)$$

3.2 回帰モデル

上記の特徴ベクトル $\boldsymbol{x}$ を用いて物性値 y を予測する回帰モデル $\mathcal{L}: \boldsymbol{x}_D \mapsto y$ を構築する．カーネルリッジ回帰，サポートベクター回帰，ランダムフォレスト回帰などの一般的なモデルにより十分な精度が得られている．

物理シミュレーションによりバーチャル・ライブラリーの物質の物性値を計算した，人工データが大量にある特別な例[G.-Bombarelli 16]を除き，材料開発の分野において多くの場合利用可能なデータ数は数十から数百程度である．したがって，大量のデータを必要とするディープニューラルネットワークは必ずしも最良の選択ではない．

3.3 解探索

上記のモデル$\mathcal{L}$ の逆問題を解くことで，所望の物性値 y を満たす特徴ベクトルの候補 $\boldsymbol{x}_{cand}$を得る．$\mathcal{L}^{-1}$ を直接求めることは不可能なので，最適化アルゴリズムによる解探索を行う．ここでは一般的な PSO (Particles Swarm Optimization:粒子群最適化)アルゴリズムを用い，$\arg\min \left|y_{target} - \mathcal{L}(\boldsymbol{x})\right|^2 + \beta_r(\boldsymbol{x})$ を求める．$\beta_r(\boldsymbol{x})$はペナルティで，事前定義特徴量の重原子の個数，各原子の個数，環状構造の個数間の関係に化学構造として矛盾が生じた際に，任意の高い値$10 < \beta_r$ を与える．

3.4 構造生成

上記で得られた特徴ベクトル候補 $\boldsymbol{x}_{cand}$ を満たす分子構造を生成する．生成の流れは，事前特徴量 $\boldsymbol{x}_P$ の各情報をもとに，分子構造の骨格(グラフトポロジー)生成，原子情報(各原子個数)の導入，結合情報の導入を順次実施することで，徐々に構造を具体化してゆく．生成は SMILES の文字列操作により行う．得られた構造を，データ駆動特徴量 $\boldsymbol{x}_D^{Select}$ に含まれる部分構

造個数と照合することで，構造の選択すなわちフィルタリングをする．

ここで問題は，各ステップが進行するたびに生成される構造の個数が指数関数的に増大・発散するため，計算負荷が増え生成が困難になることである．そこで発散を抑えるために，各ステップごとに$\boldsymbol{x}_D^{Select}$の情報によるフィルタリングを行うことを考える．ただし$\boldsymbol{x}_D^{Select}$の示す部分構造の多くは，酸素原子や窒素原子，芳香環などを含むため，生成の初期ステップでは利用することができない．そこで各ステップに応じて，これらの部分構造をフィルタとして利用可能な形に変形する．そのワークフローを図2に示す．すなわち，最初は$\boldsymbol{x}_D^{Select}$に出現するすべての部分構造を，トポロジカルな骨格構造のみに還元し，各原子情報や結合情報の導入ステップに応じて，もし元の部分構造が対応する原子や結合を持っている場合は，それらを復元することにする．こうして変形された各部分構造が生成構造に含まれる個数をカウントし，それを$\boldsymbol{x}_D^{Select}$と照合することで，不適切な構造を除外するというフィルタリングを行う．

以上のプロセスを逐次的に実行することで，事前定義特徴量とデータ駆動特徴量の両者をともに満たす構造が生成され，かつ途中経過における個数の発散を抑えることが可能となる．

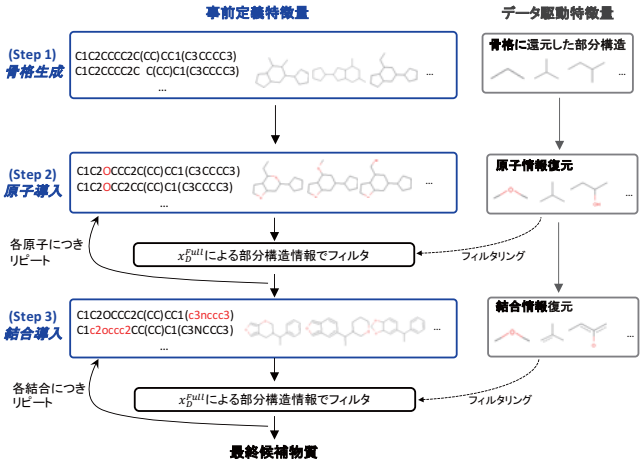

図2：構造生成のワークフロー．事前定義特徴量の情報から構造を生成し，各ステップに応じて変形させた部分構造の個数カウントとデータ駆動特徴量を比較し，構造を選別する．

4. 評価実験

本手法の実装に際し，分子からの部分構造抽出，分子に含まれる部分構造の個数カウント，および SMILES から分子構造への変換・描画，の3点については化学構造解析用のオープンツールである RDKit [GitHub 18]を用いた．

4.1 データセット

薬物の公開データベース ZINC15 中の，FDA-approved drugs データセットを用いて評価実験を実施した．本データセットに含まれる1,562 個の分子構造につき，目標物性値として，脂溶性および細胞膜透過性の指標となる logP（オクタノール/水 分配係数）および TPSA (Topological Polar Surface Area : 極性表面積) を原子団寄与法により計算した．図3にその分布を示す．A から D までの各象限に含まれる分子構造のうちランダムに取り出した5つずつの分子構造を描画しているが，logP や TPSA が高くなるにつれ，構造が大きく（原子数が多く）かつ複雑になることが確認される．本実験では，最も構造が複雑な象限 B における物質，具体的には目標物性値として$\{(y_{logP}, y_{TPSA}) \mid 3.0 \leq y_{logP} \leq 3.2 \cap 230 \leq y_{TPSA} \leq 240\}$を満たす分子構造を新規にデザインする．

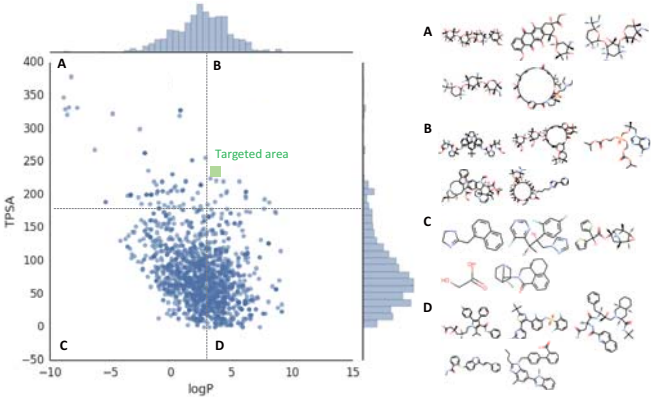

図3：物質データセットの logP および TPSA の分布．右に各象限内の化合物の例を示している．

4.2 特徴エンコード

1,562 個の分子構造につき，最大半径が2の Morgan Fingerprint を適用することで，3,760 個の網羅的な部分構造 S^{Full} を得た．3.1.1 節に従い，logP および TPSA について独立の Lasso 回帰モデル$\mathcal{L}_{logP}$および$\mathcal{L}_{TPSA}$を構築した．各々5-fold cross validation による L1 正則化項の最適化，特徴選択を行い，重複する部分構造を取り除いたところ，104 種類の部分構造が得られた．各分子構造に含まれるこれら部分構造の個数をカウントすることで得られる$\boldsymbol{x}_D^{Sel}$および，別途計算した$\boldsymbol{x}_P$を連結して，114 次元の特徴ベクトル $\boldsymbol{x}$ を得た．

4.3 回帰モデル

得られた特徴ベクトル $\boldsymbol{x}$ を用いて，logP および TPSA に対する回帰モデル$f_{logP}: \boldsymbol{x} \mapsto y_{logP}$ および $f_{TPSA}: \boldsymbol{x} \mapsto y_{TPSA}$ を構築した．モデルはリッジ回帰，ランダムフォレスト，サポートベクター回帰を用い，それぞれ 5-fold cross validation によりハイパーパラメータを最適化した．図4に，最良の結果が得られたリッジ回帰の結果を示す．本モデルにおいては，テストセットに対するスコア(R^2, RMSE) が，logP と TPSA に対して (0.98 ,0.36), (0.99, 3.4)がそれぞれ得られた．以降の節では本モデルを予測モデルとして用いる．

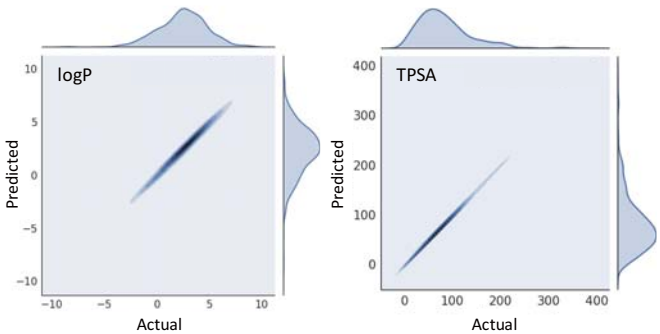

図4：回帰結果のカーネル密度分布．

4.4 解探索

f_{logP}およびf_{TPSA}を用いて，PSOにより4.1で設定した目標物性値を満たす特徴ベクトルを求める．探索範囲は，データセットの各特徴量ごとの最大・最小値の範囲内とした．1,000個の粒子による探索を実施し，特徴ベクトル候補を見つけるたびに粒子の位置をランダムに初期化することで，100個の候補を発見した時点で終了とした．

4.5 構造生成

得られた特徴ベクトル候補$\boldsymbol{x}_{cand}$を用い，3.4節のワークフローに従って分子構造を生成した．最終的に得られた分子構造の一部を図5に示している．いずれも原子数が多く複雑な形状を有しており，また構造の多様性にも富んでいる．人間の技術者がこのような構造を高速に列挙するのは極めて困難である．ここで提案する手法による生成時間（図2のワークフローに要する時間）は，構造の大きさにもよるが1分子あたり0.1から1秒程度である．したがって，従来的な試行錯誤による分子構造のデザインプロセスと比べ，速度と構造の多様性において高い優位性を実現している．また，所望の物性値から逆問題を解くため，バーチャル・スクリーニングを用いた手法と比べ，無用な構造を生成する必要がないため，探索効率が高いといえる．

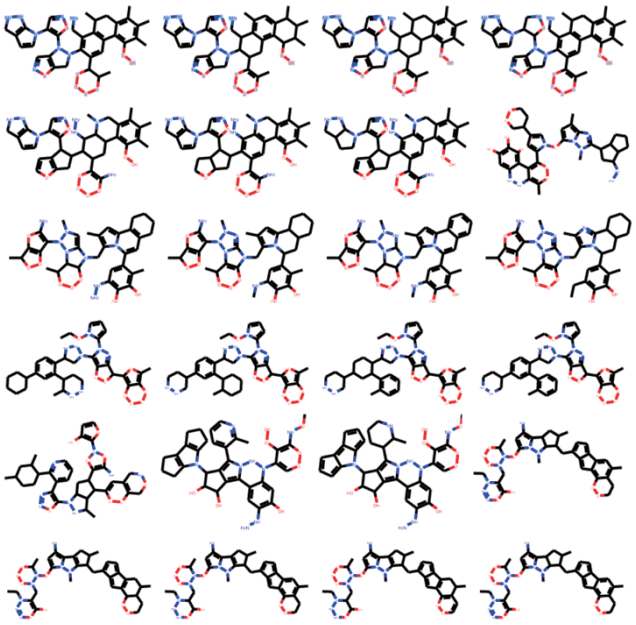

図5：生成された分子構造の一部．

5. 結論

本研究では，特徴エンコードから構造生成までend-to-endで動作するAIを活用した物質探索手法を提案した．与えられた物質データに含まれる部分構造カウントからなるデータ駆動特徴量と，構造生成に適用の容易な事前定義特徴量を組み合わせることで，順問題と逆問題の両方に対応可能なワークフローを構築した．ZINC15の薬物データを用いることで，複数の目標物性値を満足するような複雑な分子構造を高速に生成可能であることを実証した．

謝辞

本研究の一部成果に関しましては，IBM Research Frontiers Instituteの参画企業である，JSR株式会社，株式会社本田技術研究所，日立金属株式会社，長瀬産業株式会社，キヤノン株式会社の皆様から，積極的な意見を頂きました．また本研究のアルゴリズム構築にあたり，IBMアルマデン研究所のDr. Jed PiteraとDr. Victoria Piunovaから有意義なアドヴァイスを頂きました．

参考文献

[Bohacek 96] Bohacek, R.S., McMartin, C., and Guida, W.C.: The art and practice of structure-based drug design: A molecular modeling perspective, *Medicinal Research Reviews*, 16 (1), 3-50 (1996)

[Pyzer-knapp 15] Pyzre-knapp, E.O., et al., What is high-thoughput virtual screening? A perspective from organic materials discovery, *Annual Review of Materials Research*, 45, 195-216 (2015).

[Rupp 12] Rupp, M., et al., Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.*, 2018, 053801 (2012).

[Pyzer-knapp 15] Pyzer-Knapp, E.O., Li, K., and A-Guzik, A., Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery, *Materials Review*, 25, 6495-6502 (2015).

[Vership 13] Virship, A.M., et al., Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds, *Journal of the American Chemical Society*, 135, 7296-7303 (2013).

[G.-Bombarelli 16] Gomez-Bombarelli, R., et al., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, 15, 1120-1128 (2016).

[G.-Bombarelli 18] Gomez-Bombarelli, R., et al., Automatic chemical design using a data-driven continuous representation of molecules, *ACS Central Science*, 4(2), pp.26 8-276 (2018).

[Kusner 17] Kusner, M. J., Paige, B., Miguel, J., Hernández-Lobato, Grammar Variational Autoencoder, *Proceedings of the 34th International Conference on Machine Learning, PMLR*, 70, 1945-1954 (2017)

[Arakawa 05] Arakawa, M., Yamada, Y., Funatsu, K., Development of the computer software for automatic chemical structure generation using group contribution method, *Journal of Computer Aided Chemistry*, 6, 90-96 (2005).

[Huan 15] Huan, T.D., M.-Kanakkithodi, A., and Ramprasad, R., Accelerated materials property predictions and design using motif-based fingerprints, *Physical Review B*, 92, 014106 (2015).

[GitHub 18] GitHub and SourceForge, RDKit: Open-Source Cheminformatics Software (http://www.rdkit.org/)