DATA SCIENCE PBL Ⅱ

Data-Assisted Exploration for Organic Electronics Materials

データ駆動による有機エレクトロニクス材料の探索

Riku KANO

1. PURPOSE

有機エレクトロニクスの研究において、分子のエネルギー準位をコントロールすることは有機材料の物性を向上させる上で極めて重要である。有機薄膜太陽電池においては、電子供与性の分子種（**ドナー**）の最高被占有軌道(Highest Occupied Molecular Orbital; **HOMO**)と電子受容性の分子種（**アクセプター**）の最低空軌道（Lowest Unoccupied Molecular Orbital; **LUMO**）のエネルギー差が太陽電池で発生する電圧と深い関係にある。発電効率を高めるには有機分子のバンドエンジニアリングが必要不可欠である。

また、有機 EL の開発においてもエネルギー準位をコントロールすることは重要である。青色や赤色などといった特定の色をもった光を放出する有機 EL 照明を開発するには、その色の波長に対応するエネルギーギャップを持った有機分子を作成する必要がある。近年、熱活性化遅延蛍光（Thermally Activated Delayed Fluorescence; **TADF**）を利用した有機 EL 素子の研究が活発に行われている。TADF とは分子の周囲の熱エネルギーをもとに、三重項状態（Triplet; T）から一重項状態（Singlet; S）へと**逆系間交差**（Reverse Intersystem-Crossing; **RISC**）させることによって、一重項状態から蛍光を取り出す過程である（図 1）。この現象を応用することにより、今までは実現することが困難であった高い量子効率を持つ素子を開発することができるようになり、高い注目が集められている。その量子効率を向上させるには RISC が発生する確率を高めれば良い。エネルギー準位間の遷移確率や遷移の速度k_{RISC}はフェルミの黄金律をもとにして導くと以下のようにして表せる。

$$k_{\mathrm{RISC}} \propto \left|\langle \mathrm{S}|\, \hat{H}_{SOC}\, |\mathrm{T}\rangle\right|^2 \exp\left(-\frac{\Delta E_{\mathrm{ST}}}{k_B T}\right) \tag{1}$$

ここで、$\hat{H}_{SOC}$はスピン-軌道相互作用を考慮したハミルトニアン、$\langle \mathrm{S}|, |\mathrm{T}\rangle$はそれぞれ一重項状態、三重項状態の波動関数、$\Delta E_{\mathrm{ST}}$は一重項状態と三重項状態のエネルギー差、$k_B$はボルツマン定数、$T$は絶対温度である。(1)式の指数部分に着目すると、RISC の反応速度k_{RISC}を高めるにはΔE_{ST}のエネルギー差を 0 に近づけることが鍵となる。TADF を応用した有機発光デバイスを開発する上で、一重項状態と三重項状態のエネルギー差を狭めることは重要になる。

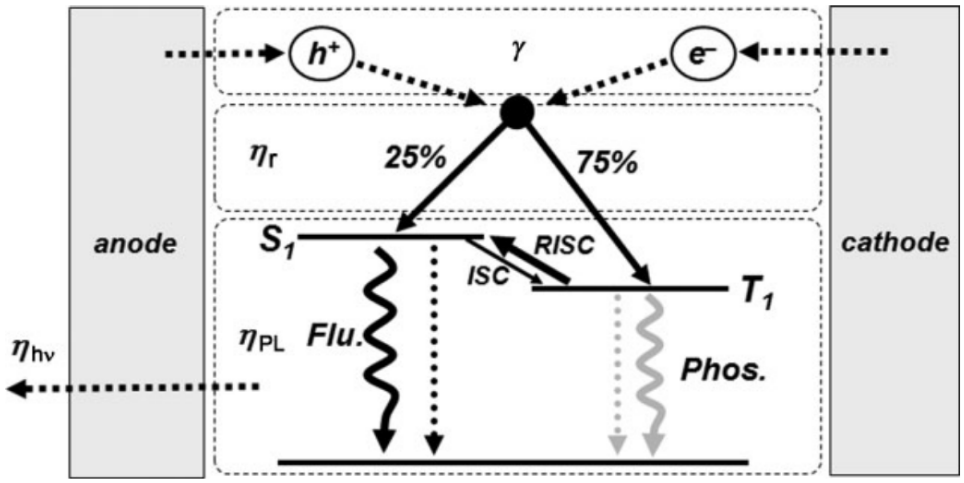

Figure 1. Schematic view of the electroluminescence mechanism: carrier injection, transport, recombination; and radiative decay processes. The thermally activated delayed fluorescence (TADF) process is highlighted. (γ: ratio of holes and electrons in carrier injection, transport, and recombination processes; η_r: singlet and triplet exciton formation ratio; η_{PL}: photoluminescence efficiency; $\eta_{h\nu}$: light out-coupling efficiency; ISC: intersystem crossing; and RISC: reverse intersystem crossing.)

図 1 TADF の原理の概要。Ref[1]からの引用

このエネルギー差を計算で予測する方法として、時間依存密度汎関数理論(Time Dependent Density Functional Theory; TD-DFT)などといった手法を駆使した第一原理計算がある。TD-DFT は、通常の密度汎関数理論とは異なり、時間にも依存するコーン・シャム方程式を用いて、分子の電子状態を逐次的に導く理論である。しかし、この計算手法の計算量は、計算する系に含まれる波動関数の基底の個数をNとすると少なく見積もっても$O(N^3)$になる。それゆえ、分子を１つ１つ虱潰しのように探索して行った場合、膨大な時間を要することになる。

そのような課題を克服するために、このリサーチでは一重項状態と三重項状態のエネルギー差を、深層学習モデルを駆使することによって高速にスクリーニングを行うモデルを開発することに取り組むこととした。

2. METHOD

データセット

データセットはRef[3]を利用してモデル開発に取り組んだ。このデータセットはGaussianと呼ばれる有償ソフトウェアによって計算された48182件のTD-DFT計算の結果を収蔵している。このデータセットに含まれるデータの概要は以下のテーブルのようになる。今回はこの中から$\mathrm{E(S1), E(T1), E(T2), E(T3)}$を用いて一重項状態と三重項状態のエネルギー差$\Delta E_{S1-T1}, \Delta E_{S1-T2}, \Delta E_{S1-T3}$を算出し目的変数と定めた。説明変数としてはSMILESを用いることを考えていたが、一重項状態や三重項状態などといった電子励起状態は3次元の構造に大きく依存するため、SMILESといった2次元情報から記述子を作成するだけでは不十分であると考えられた。図2にSMILESが全く同一の化合物の3次元構造をVESTAという可視化ソフトウェアで2つ可視化した。両者を比較すると3員環が持つ官能基の立体的な構造が異なっている。このような立体構造の差異が生じると分子の電子状態にも差異が生じるため、エネルギー状態も変化することになる。このようなことが想定されるため、今回は元論文のGaussianの計算ログから構造最適化が済んだ状態の分子の3次元座標を抽出し、その3次元座標と原子種を説明変数としてモデル構築をすることとした。

Table 1 データの概要

カラム名	単位	詳細
ID	-	ID
Doi	-	分子の構造をX線で決定した論文のDOI
Formula	-	化学式
NAts	-	重原子の個数
SMILES	-	分子のSMILES表記
HOMO	eV	HOMOのエネルギーの計算値
LUMO	eV	LUMOのエネルギーの計算値
E(S1)	eV	S1状態のエネルギーの計算値
f(S1)	-	S1状態への振動子強度
E(S2)	eV	S2状態のエネルギーの計算値
f(S2)	-	S2状態への振動子強度
E(S3)	eV	S3状態のエネルギーの計算値
f(S3)	-	S3状態への振動子強度
E(T1)	eV	T1状態のエネルギーの計算値
E(T2)	eV	T2状態のエネルギーの計算値
E(T3)	eV	T3状態のエネルギーの計算値

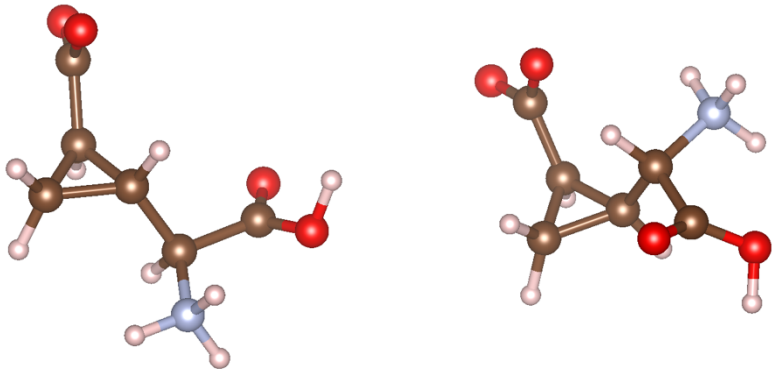

図2 ID:ZOGTIY と ZOGTOE の3 次元構造の比較。両者の SMILES は同一である。

モデル

3 次元座標から目的の物性を予測するモデルはこれまでにさまざま提案されてきているが、今回は PaiNN[2]をベースとしたモデルを用いることにした。アーキテクチャの概要は図 3 に示してある。このモデルの特徴としては、物理の世界の現象を記述する方程式に課される制約条件や対称性をモデルに組みこんでいる。いわゆる帰納バイアスを考慮した message passing モデルの一つである。他の特徴としてはスカラー値だけでなく、分極率などといったテンソル値も予測を目的とするモデルとなっている。このモデルは、原子種Zとその原子の座標rをインプットとして、潜在空間に埋め込み、message ゲートや update ゲートを通して、周囲の原子と message をやりとりすることによって潜在表現を獲得していくスキームを取る(図 3 の橙色と黄色の箇所)。

今回は原子の潜在基底を 128 次元のベクトルで表現し(図 3 の緑色の embedding 操作に相当する)、message ゲートと update ゲートをそれぞれ 3 個ずつをモデルに組み込んだ。message ゲートと update ゲートで潜在表現を獲得した後、得られたスカラー値の潜在表現をもとにそれぞれの目的変数を予測する多層パーセプトロンに表現ベクトルを受け渡し、目的変数を予測させる(図 3a の黄色から青色の箇所)。また、今回はΔE_{S1-T1}, ΔE_{S1-T2}, ΔE_{S1-T3}を目的変数として予測するシングルタスク学習だけでなく、E(S1), f(S1), E(T1),

E(T2), E(T3)も予測値に加えたマルチタスク学習モデルも同時に学習させ、性能の比較を試みた。データセットは 8:1:1 の割合でそれぞれ訓練データ、検証データ、テストデータへとランダムに分割した。学習の最大エポック数は 300、バッチサイズは 128、初期の学習率は 0.0005 とし、予測値と実測値の二乗平均誤差を損失関数と設定し訓練を行なった。ここでシングルタスク学習とマルチタスク学習では以下のような損失関数L_{single}, L_{multi}を用いることとした。右辺のLは全て二乗平均誤差である。また、今回は簡単のため、重みλは全て 1.0 とした。

$$L_{single} = \lambda_{\Delta E_{S1-T1}} L_{\Delta E_{S1-T1}} + \lambda_{\Delta E_{S1-T3}} L_{\Delta E_{S1-T2}} + \lambda_{\Delta E_{S1-T3}} L_{\Delta E_{S1-T3}}$$

$$L_{multi} = \lambda_{\Delta E_{S1-T1}} L_{\Delta E_{S1-T1}} + \lambda_{\Delta E_{S1-T3}} L_{\Delta E_{S1-T2}} + \lambda_{\Delta E_{S1-T3}} L_{\Delta E_{S1-T3}} + \lambda_{\mathrm{E(S1)}} L_{\mathrm{E(S1)}} + \lambda_{\mathrm{f(S1)}} L_{\mathrm{f(S1)}} + \lambda_{\mathrm{E(T1)}} L_{\mathrm{E(T1)}} + \lambda_{\mathrm{E(T2)}} L_{\mathrm{E(T2)}} + \lambda_{\mathrm{E(T3)}} L_{\mathrm{E(T3)}}$$

テストデータの評価指標としては平均絶対誤差を用いてモデルの性能評価を行うこととした。実装には PyTorch、PyTorch Lightning を使用し、実験管理を進めていった。

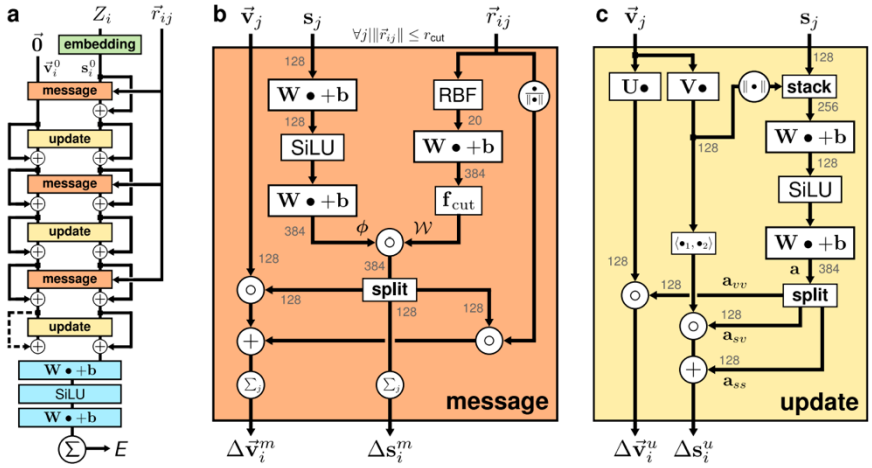

FIG. 2: The architecture of PaiNN with the full architecture (**a**) as well as the message (**b**) and update blocks (**c**) of the equivariant message passing. In all experiments, we use 128 features for $\mathbf{s}_i$ and $\vec{\mathbf{v}}_i$ throughout the architecture. Other layer sizes are annotated in grey.

図 3 PaiNN のアーキテクチャの概要図。Ref[2]からの引用。

3. RESULTS

学習させた機械学習モデルの評価指標は Table2 のような結果となった。シングルタスク学習とマルチタスク学習では、シングルタスク学習がより高い性能を示すことになった。マルチタスク学習では他の予測値を加えて性能向上が見込めるかと想定したが上手くいかなかった。また、真値と予測値の y-y プロットは図 5, 6 のようになった。マルチタスク学習では上手く学習ができておらず、予測値に偏りが見られる。一方でシングルタスク学習ではΔE_{S1-T1}の予測は概ね赤線状に存在しており、上手く予想ができているように見受けられるが、$\Delta E_{S1-T2}, \Delta E_{S1-T3}$では予測値と真値の間にシフトが見られる結果となった。

Table 2 モデルの予測値と真値の平均絶対誤差(eV)

目的変数	シングルタスク学習	マルチタスク学習
ΔE_{S1-T1}	0.1541	0.2264
ΔE_{S1-T2}	0.1770	0.2511
ΔE_{S1-T3}	0.1987	0.2734

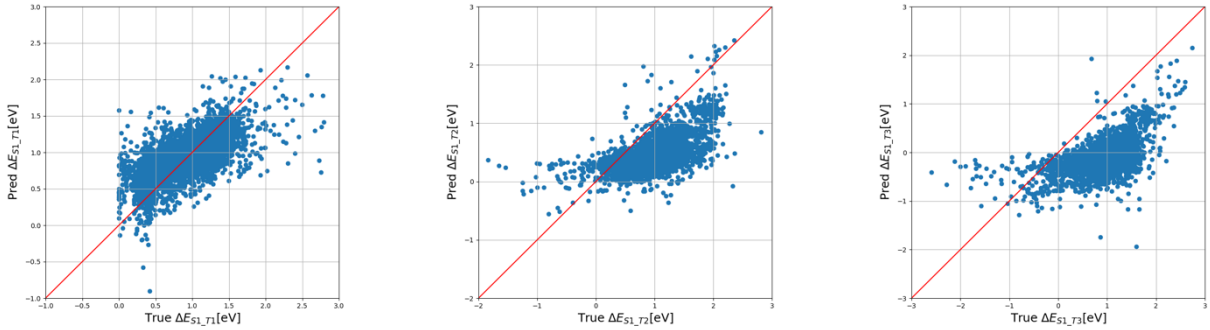

図 5 マルチタスク学習によって得られたモデルの y-y プロット

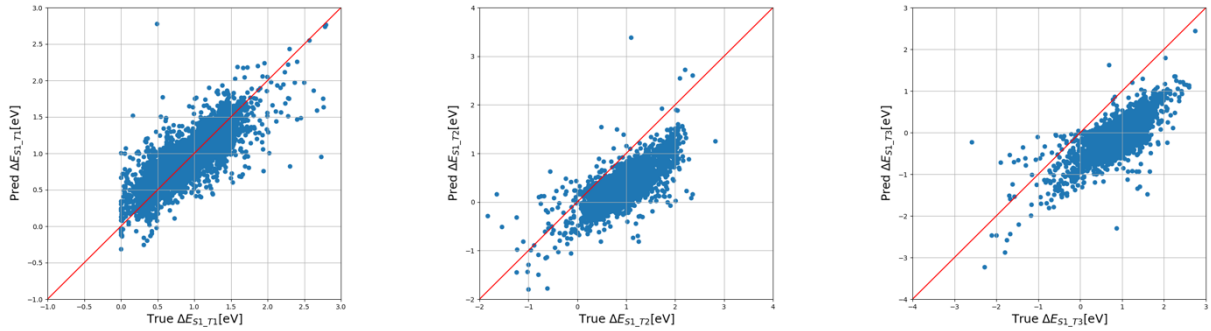

図 6 シングルタスク学習によって得られたモデルの y-y プロット

4. DISCUSSION

今回はシングルタスク学習による予測値が、マルチタスク学習による予測値より秀でる結果となったが、この結果は深く追究できる余地が残されている。時間の関係上、このリサーチに深く取り組むことが困難であったため、試行錯誤を繰り返し損失関数の重みを調整することによってマルチタスク学習によるモデルがより良いモデルとなることも想定される。

シングルタスク学習の結果を参照すると、ΔE_{S1-T1}の予測値の絶対平均誤差は 0.1541eV となった。正確な予測にはまだまだ程遠い結果となったが、TADF の性能を示す有機材料を選別する一助になることは間違いないと考えられる。ただ、真値が 0 eV のデータの予測値を見てみると上手く予想できていないようにも伺える。これに関しては、元データの分布が図 7 のようになっているためだと考えられる。図 7 の一番上のグラフの 0 eV 付近を見てみると小さなスパイクが見られる。このようなデータの山の裾部分の予測は外挿領域に差しかかるので予測が多少困難になりやすい為、0eV 付近の予測に少し乱れがあると考えられる。0eV 付近の予測を成し遂げるために、更なるデータ収集が必要になる。

また、今回は説明変数として原子種Zと原子の座標rのみを用いたが他の説明変数も必要になるケースが考えられる。例えば、部分電荷を持つ化合物やカチオン、アニオンといった分子を想定すると原子が持つ電荷の情報も必要になってくる。なぜならば、電荷の有無によって波動関数の形状が変化し、得られる分子のエネルギーが変化することが起こりうるからだ。実際、今回扱かった分子の中にはカチオンやアニオンが含まれていたが電荷の情報を考慮せずモデル構築に進んだ。過去の研究でカチオンやアニオン分子に対して電荷移動などを考慮したモデルを開発し、分子のエネルギーを予測した報告例[4]もあるので、類似した手法を適用することにより今回使用したモデルを更に改善することができると期待される。

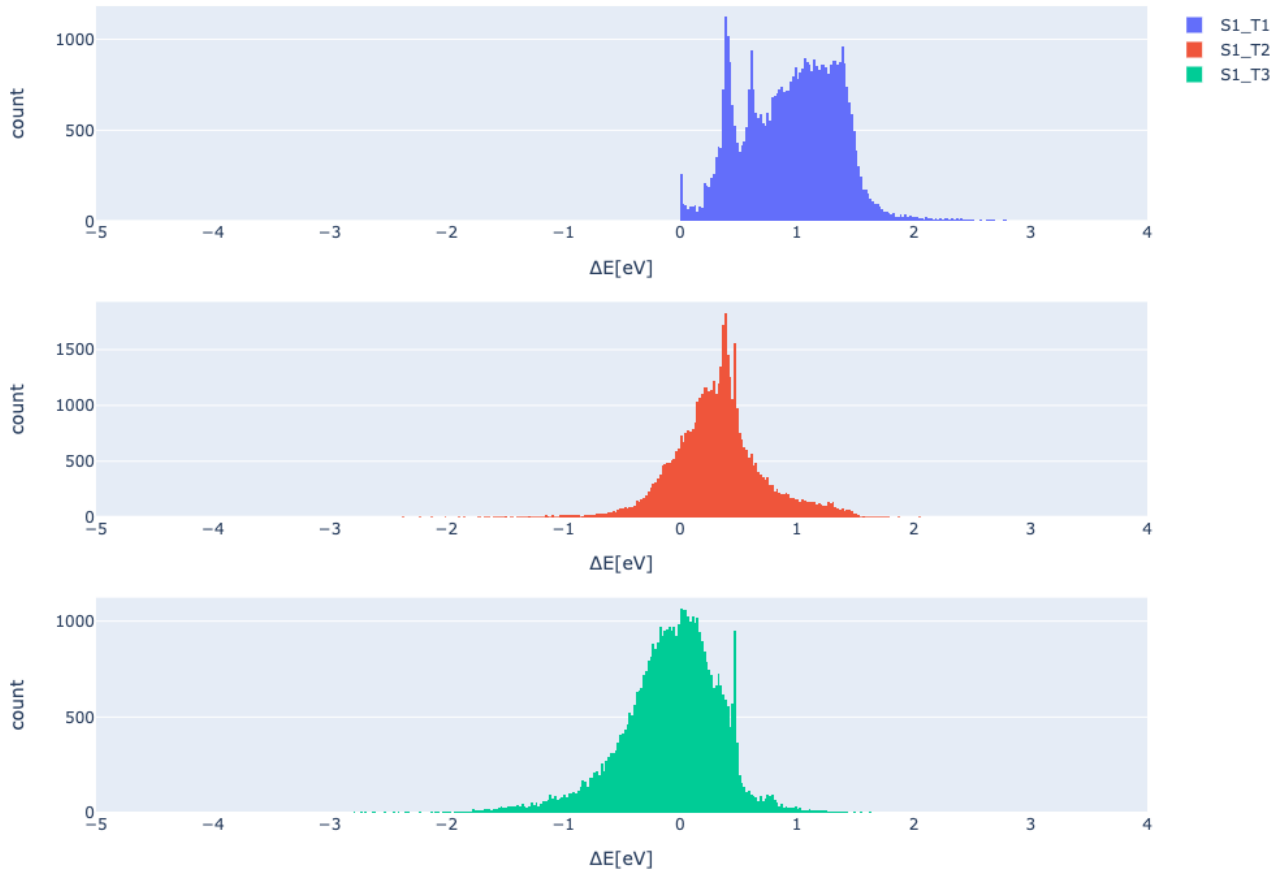

図 7 $\Delta E_{S1-T1}, \Delta E_{S1-T2}, \Delta E_{S1-T3}$ *のヒストグラム*

5. CONCLUSION

このリサーチではPaiNNベースの深層学習モデルを構築し、一重項状態と三重項状態のエネルギー差を予測するモデルを、シングルタスク学習、マルチタスク学習で訓練させた。その結果、シングルタスク学習で訓練させたモデルがより良い性能を示すことが判明した。シングルタスクのΔE_{S1-T1}では絶対平均誤差が0.1541eVを示し、TADFを利用した有機エレクトロニクス材料探索の試金石として利用可能であることが想定される。

今後の課題としては、このリサーチに深く取り組めなかったため、更なるハイパーパラメータ調整や機械学習モデルの改善などが挙げられる。また、今回はスクリーニングのためのモデル開発を行なったが、Diffusion modelやGenerative Adversarial Networkなどの生成モデルを活用した新規材料探索も注目が集められているのでこのような分野に注力することを展望している。

6. REFERENCE

[1] Ayataka Endo *et al*. Thermally Activated Delayed Fluorescence from Sn^{4+}–Porphyrin Complexes and Their Application to Organic Light Emitting Diodes — A Novel Mechanism for Electroluminescence. Advanced Materials **21**, 47(2009). https://doi.org/10.1002/adma.200900983

[2]https://arxiv.org/abs/2102.03150

[3] Omar, Ö.H., Nematiaram, T., Troisi, A. *et al.* Organic materials repurposing, a data set for theoretical predictions of new applications for existing compounds. *Sci Data* **9**, 54 (2022). https://doi.org/10.1038/s41597-022-01142-7

[4] Ko, T.W., Finkler, J.A., Goedecker, S. *et al.* A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat Commun* **12**, 398 (2021). https://doi.org/10.1038/s41467-020-20427-2