

材料科学における
多次元出力変数の教師あり学習
に関する研究

岩山 めぐみ

総合研究大学院大学
複合科学研究科
統計科学専攻

指導教員：吉田 亮 教授

令和4（2022）年度

諸言

材料研究の諸分野において、マテリアルズインフォマティクスと呼ばれる新しい学際研究に注目が高まっている。マテリアルズインフォマティクスの教師あり学習では、モデルの入出力変数は多様な形式をとる。入力変数の典型的な例は、化学組成や分子の化学構造、材料の配合比、プロセス条件等である。これらの入力変数の多くは非数値型変数として与えられることが多い。そのような場合、モデルは一旦入力変数を記述子という数値ベクトルに変換した上で出力変数である物性や機能への写像を定める。材料研究で対象となる出力変数の典型的な例は、密度や融点のようなスカラー量で表される特性値である。一方、マテリアルズインフォマティクスでは、電子顕微鏡の画像データや物性の温度依存曲線、紫外線可視近赤外分光法や核磁気共鳴分析から得られるスペクトル関数のように、行列・テンソル・関数型変数で与えられるケースも数多くある。しかしながら、このような多次元出力変数を対象とした教師あり学習の研究はほとんど進んでいない。

本研究では、出力変数が関数で与えられた場合のモデリング・推定・予測の方法論について考える。特に材料研究における関数出力回帰の未開拓問題の発掘と幅広い問題に適用可能な汎用的方法論の構築を目的とする。適用対象としては、分子の光吸収スペクトルの予測、材料微細構造（電子顕微鏡画像）の予測、物性の温度・周波数依存性の予測などを想定する。カーネル回帰と深層学習を組み合わせた関数出力回帰モデルを提案し、有機分子の光吸収スペクトルの予測、複合薄膜材微細構造の画像予測への適用例を通じて、提案手法の予測性能を解析する。特にデータが少ない状況において、提案手法は著しく高い汎化性能を示すことを明らかにする。関数出力回帰では、関数全体を同時に学習・予測することでマルチタスク学習と同様のメカニズムが働き、従来のスカラー変数の予測よりも圧倒的に高い汎化性能を引き出せることが実験的に示される。この作業仮説を検証するために、通常の教師あり学習で関数の値を独立に予測するケースとスカラーの関数特徴量を予測するケースとの比較を行い、関数全体の予測が圧倒的に高い汎化性能を示すことが実験的に確認された。

材料研究において、関数型出力変数を予測する機械学習の用途は、他にも数多く考えられる。ほとんどの材料特性は、温度または周波数の関数として与えられる。材料の誘電特性は温度と周波数の関数として定義される。線膨張係数、体積弾性率、比熱、熱伝導率、 χ パラメータ等の物性も温度依存的に決まる。これらの温度依存性曲線から、ガラス転移温度、結晶融点、結晶化温度等の物性が計算される。物性の加工条件への依存性も関数出力回帰の典型的な応用例である。さらに、材料構造の2次元あるいは3次元イメージングでは、出力変数は画像座標空間における多次元関数として与えられる。このように、材料研究では、関数型出力回帰を適用できる未開拓問題が数多く残されている。本研究の提案手法は、原理的には任意の高次元関数データを扱うことができるように設計されている。本研究は、これらの潜在的問題を統一的視点から捉えた上で、少数データにも適用可能な関数出力変数の方法論を構築した。

目次

第 1 章	はじめに	5
1.1	材料科学における機械学習	5
1.2	材料科学における多次元出力変数	7
1.3	研究の目的と概要	8
1.4	本論文の構成	9
第 2 章	材料科学における関数出力回帰	10
2.1	スペクトル関数の予測	10
2.2	材料の微細構造の予測	13
2.3	その他の問題設定	13
2.4	提案手法の特徴	14
第 3 章	関数出力回帰の方法	16
3.1	提案手法：Kernel Regression with Functional Outputs	16
3.2	関連手法 1：関数データ解析における回帰	19
3.3	関連手法 2：Conditional GAN	22
3.4	関連手法 3：Long Short-Term Memory	24
3.5	関連手法 4：注意機構を用いた Seq2Seq	25
3.6	関連手法 5：マルチタスク学習	25
第 4 章	モデル設計と性能指標	27
4.1	Kernel Regression with Functional Outputs のモデル設計	27
4.2	Conditional GAN のモデル設計	28
4.3	性能評価指標	33
第 5 章	結果と考察	42

5.1	スペクトル関数の予測	42
5.1.1	UV-Vis スペクトルの予測	42
5.1.2	近赤外スペクトルの予測	48
5.2	電子顕微鏡像の予測	49
5.3	少数データに対する KRFO の予測メカニズム	51
第 6 章	まとめ	56
付録 A	実験結果の補足	58
A.1	スペクトル予測の実験結果補足	58
A.2	微細構造予測の実験結果補足	63
	謝辞	70
	参考文献	71

第 1 章

はじめに

1.1 材料科学における機械学習

近年、さまざまな分野で機械学習を活用した技術革新が注目を浴びている。材料科学においては、マテリアルズインフォマティクス (MI) というデータ科学と材料科学の学際領域に大きな期待が寄せられている [1]。2011 年米国にて、材料開発の開発期間やコストの大幅な削減をミッションとする国家プロジェクト Materials Genome Initiative (MGI) が始動したことを発端に、日本でも 2015 年に JST イノベーションハブ構築支援事業「情報統合型物質・材料開発イニシアティブ」が始動し、MI への注目度が増すこととなった。

材料開発では、新規材料の発見から量産・製品化までにおよそ 10 年以上の年月を要するといわれている。開発期間が長期化する要因の一つとして、研究者の経験と勘に依存した試行錯誤型の材料設計や材の選定、生産プロセス条件の探索が挙げられる。一般に材料設計やプロセス条件決定におけるパラメータ空間は超巨大である。例えば、低分子化合物の化学空間は、 10^{60} 種を超える候補物質が存在するといわれている [2]、プラントのプロセス条件の探索空間は、連続あるいは非連続的な多数の製造工程から構成され、検討すべき条件の組み合わせは膨大な数になる。このような広大なパラメータ空間から人間の経験と勘のみで最適解を発見するのは非現実的である。MI に課されたミッションは、材料開発プロセスのボトルネックとなっている作業工程に機械学習の技術を導入し、従来の属人的な材料開発が対象としてきた探索空間を大幅に拡張することで、研究開発コストの削減と革新的特性を有する新材料の創製を実現することである。

MI の基本的なワークフローは順問題と逆問題から成る [3, 4, 5]。順問題では、化学構造や合成条件、プロセス条件を入力として、材料の機能、物性、構造的特徴等の出力変数を予測する。逆問題では入力と出力が反対になる。逆問題は、順問題のモデルの逆写像を

求めて、所望の機能、物性、構造を有する入力変数の予測を目的とする。近年の計算科学的アプローチによる材料シミュレーションの技術進歩により、第一原理計算や分子動力学計算、有限要素法等を用いることで順方向の予測精度は大幅に改善した。しかしながら、これらの手法で複雑な現実系の時間・空間スケールを再現することは、現在の計算機の性能では難しい。また計算可能な系だとしても、探索対象のパラメータ空間が広大であるため、シミュレーションモデルを用いて直接的に逆問題を解くと計算コストは無視できないほど大きくなる [6]。そこで、統計モデルとシミュレーションを相補的に用いて、計算コストを抑制しつつ高精度な順方向の予測モデルを創出しようという研究が進められている。

MI のデータ解析の特徴は、入出力変数の特殊性と高次元性にある。典型的な入力変数である化学組成や化学構造、結晶構造等は、一般に固定長ベクトルに基づく特徴表現の方法が非自明な変数である。このような非数値変数を順問題・逆問題のワークフローに組み込むには、変数の形式に応じて記述子と呼ばれる固定長の数値ベクトルに変換する必要がある [3, 4]。MI では、化学組成、分子、結晶構造対象に様々な数値化の方法が研究されてきた [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]。

物質の構成要素の一つである化学構造は、原子と結合状態により表される。化学構造を機械学習で扱いやすくするための最も標準的な化学構造の表記方法として SMILES 記法 (Simplified Molecular Input Line Entry System: Weininger, 1988) がある。SMILES 記法は化学構造を文字列で記述する。化学構造は、ラベル付きの無向グラフと考えることができる。無向グラフ $G = (V, E, L_V, L_E)$ は、頂点集合 V と、エッジ集合 E から成る。頂点は原子、エッジは結合を表し、頂点には元素種 ($L_V \in \{C, O, \dots, N\}$)、エッジには結合次数 ($L_E \in \{1, 2, \dots\}$) を表すラベルが付与される [3]。SMILES 記法では、ラベル付き無向グラフを一定のルールに従い化学構造を文字列に変換する [19]。原子は元素記号によって表される、単純な水素結合は省略する、隣接原子は隣に記載する、二重結合と三重結合はそれぞれ「=」と「#」で記載する、分岐構造は括弧で表記する、環構造は接続箇所の二つの原子に数値を割り当て記述する、といった規則があり、この規則に従い記述することにより、2次元の化学構造のグラフを1次元的な文字列に変換する。

MI では、SMILES 記法により文字列に変換された化学構造をフィンガープリント (Fingerprints) と呼ばれる記述子に変換して、入力変数として使用する。フィンガープリントは、特定の部分構造の有無や個数により、化学構造を整数値ベクトルに変換する。通常、計算対象の部分構造は数百から数千種類ほどであり、部分構造の取り方や数え上げのアルゴリズムについて様々な手法が研究されてきた [12, 13, 14, 15, 16]。また、分子構造のトポロジーや物理化学的な特徴量を記述子とすることもある [20, 21, 22, 23]。化学組成は、可変数の元素種とその含有量からなる集合変数であり、元素特徴量と組成情報を固定

長のベクトルに変換する様々な方法が提案されてきた [8, 9, 10, 11, 7]. 結晶構造の記述子は、単位セル内の構成原子の局所的隣接関係をベクトル化する [24, 9, 10, 11]. また近年では、物質の構造をグラフとして扱い、グラフニューラルネットワークを用いてデータから自律的に特徴量を抽出する方法も広く普及している [25, 26, 27]. 例えば、結晶系における原子の周期的な構造を、結晶グラフと呼ばれる隣接関係を表すグラフに変換して無限に配置された単位セルの構成原子の配位関係を表現し、グラフニューラルネットワークで予測モデルを構築する方法が提案されている [25].

1.2 材料科学における多次元出力変数

出力変数である機能、物性、構造の特徴量はスカラー変数で表されることが多い。しかしながら、MI で予測の対象となる変数は、スカラー量に限らない。例えば、材料の表面や界面構造を観察するために走査型電子顕微鏡 (SEM) や透過型電子顕微鏡 (TEM) が広く用いられ、また、不純物の同定や物質の光吸収特性を解析するために、ガスクロマトグラフ装置 (GC-MS) や紫外可視分光法によるスペクトル計測が行われる。その場合、データは関数の形で与えられる。材料研究では、電子顕微鏡の画像や赤外分光法・核磁気共鳴分析から得られるスペクトルのように行列・テンソル・関数型の多次元変数として取り扱うべきデータが数多く存在している。本論文では、このような多次元出力変数を統一的に関数型出力変数と呼ぶことにする。

これまでの MI の研究は、主にスカラー変数の予測を対象としてきた。一方、関数型出力変数を対象とした MI の体系的な研究はほとんど進んでおらず、ごく限られた問題において、関数や画像データの回帰・予測を試みた研究が散発的に発表されているというのが現状である。例えば、分子の紫外可視光 (UV-Vis) 吸収スペクトルを予測する問題設定では、入力変数は化学構造を表す記述子ベクトル、出力変数は波長領域に定義された光吸収スペクトル関数となる。実際、有機分子の UV-Vis スペクトルを予測するために注意機構 (attention mechanism) を導入した Seq2Seq を適用した先行研究がある [28]。複合材料の微細構造の研究では、温度、圧力、組成等の加工条件が微細構造に及ぼす影響を定性的・定量的に理解したい。通常、微細構造を調べるために SEM や TEM による観測が行われる。これを教師あり学習の問題形式に帰着させると、入力変数は処理加工条件を表す実数ベクトル、出力変数はグレースケースの電子顕微鏡像を示す輝度行列となる。すなわち、行列が出力変数の回帰分析に帰着する。また別の視点からこの問題設定を眺めると、画像生成のタスクと捉えることもできる。実際に条件付き GAN (conditional generative adversarial networks: cGANs) [29] やエンコーダー・デコーダーモデル [30] 等の深層生

成モデルを適用して材料微細構造を予測する研究も行われている [31, 32, 33]. また, 関数データ解析 (functional data analysis) [34] という分野で研究されてきた関数型出力変数の回帰手法を用いて問題解決を図ることも考えられる.

1.3 研究の目的と概要

MI において関数出力回帰という視点に基づいた体系的な研究はほとんど進んでいない. また, 本来は関数出力回帰に分類される MI の問題が数多く存在するにもかかわらず, これらの多くは未発掘な状態となっている. そこで本研究では, 材料科学における高次元出力回帰の問題発掘と汎用的な方法論の提案を目的とする. 特に, カーネル回帰と深層学習を組み合わせた関数出力回帰の方法論を構築し, 三つのケーススタディに適用してその有用性を実証する.

関連研究として, Banko et al.[31] によって提案された cGAN による微細構造画像の予測の研究がある. ただし, cGAN には, 敵対的な学習プロセスに起因するトレーニングの不安定性と限られたデータに対する学習性能の低下という問題点がある. また, Urbina et al.[28] は, 注意機構 [35] を導入した Seq2Seq を用いて化学構造から光吸収スペクトルへの写像を学習している. しかしながら, これらのモデルは, 多次元出力変数の回帰ではパラメータ数が非常に大きくなるため, 十分な量のデータを利用できないケースにおいて学習性能が大幅に低下する. そこで本研究では, 関数データ解析における関数出力回帰のアイデアを拡張する形で Kernel Regression with Functional Outputs (KRFO) という方法論を考えた. KRFO のモデルは, 深層学習による表現学習とカーネル回帰を組み合わせて記述子ベクトルから関数出力への写像を表現しており, 特に多次元出力変数の場合, cGAN や Seq2Seq の深層生成モデルに比べて, パラメータ数を大幅に節約できる. 後述するように, この特性から KRFO はデータが少ない状況においても優れた予測性能を発揮する.

ケーススタディの内二つは, UV-Vis (170~780nm) と近赤外波長 (NIR:780~2,500nm) の二つの異なる領域における有機分子からの光吸収スペクトルの予測を行う. 出力変数は波長に対するスペクトル関数である. UV-Vis の場合, トレーニングのデータ数は数千のオーダーであるが, 近赤外波長では, トレーニング分子の数は 60 に満たずデータ数が非常に少ない. 三つ目のケーススタディでは, 任意の組成および加工条件に対する薄膜複合材料の微細構造の電子顕微鏡画像を予測する. これらのケーススタディを通じて, 限られた訓練データに対する潜在的な予測能力や学習メカニズムを検証する. 特に, スペクトル関数のケーススタディでは, 事前に定量化されたスペクトルのスカラー特徴量を予測す

る通常の回帰や波長毎にスペクトル関数の強度を独立に回帰した結果との比較実験を実施し、KRFO の優位性とマルチタスク学習 [36, 37] と関数出力回帰の関連性について考察する。なお本論文の内容の多くは、Iwayama et al.[38] の研究成果で発表したものに基づく。

1.4 本論文の構成

本論文の構成は以下の通りである。第 1 章では、材料科学における機械学習の背景等を紹介し、本研究の目的と概要を述べた。第 2 章では、分子の光吸収スペクトルの予測、材料微細構造（電子顕微鏡画像）の予測、物性の温度・周波数依存性の予測等、材料研究における関数出力回帰の適用範囲の広さと潜在的可能性を論じ、先行研究のレビューを行う。第 3 章では、KRFO の方法論を説明する。また、条件付き敵対的生成ネットワークなどの深層学習に基づく手法や関数データ解析の分野で研究されてきた既存手法を取り上げ、提案手法の特色やデータが少ない状況における利点等を論じる。第 4 章では、KRFO のモデル構造の設計等を詳細に説明する。第 5 章では、有機分子の光吸収スペクトルの予測、複合薄膜材微細構造の画像予測への適用例を通じて、提案手法と既存手法の予測性能を比較する。特にデータが少ない状況において、提案手法は著しく高い汎化性能を示すことを報告する。この観測結果に対し、関数全体を予測することでマルチタスク学習と同様のメカニズムが働く可能性があることが示唆される。この仮説を検証するために、通常の教師あり学習で関数の値を独立に予測するケースとスカラーの関数特徴量を予測するケースとの比較を行い、関数全体の予測が圧倒的に高い汎化性能を示すことを実験的に確認する。第 6 章は、まとめと今後の課題の章である。

第2章

材料科学における関数出力回帰

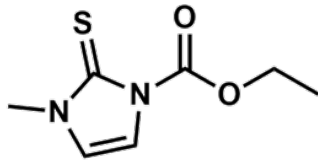
材料科学の研究開発や生産現場には多くの関数型データが存在しているものの、機械学習による有効な活用方法は未だ示されていないものが多い。本研究では、材料科学において関数データを出力変数とする回帰分析の問題を発掘・整理し、そのような問題形式を統一的に取り扱うための方法論を開発する。これ以降、入力変数を $X \in \mathbb{R}^p$, 出力変数を $Y(X, t) \in \mathbb{R}$ と表す。 Y の引数 $t \in \mathbb{R}^q$ は、例えば、画像空間の座標やスペクトル関数の定義域の波長に相当する。以下の節では、材料科学におけるいくつかの問題設定を紹介する。

2.1 スペクトル関数の予測

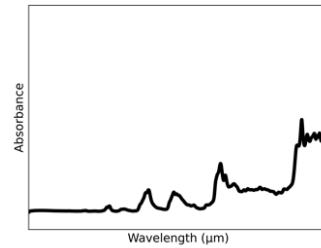
分子は、UV-Vis や NIR などの波長領域で光の吸収に応答し、基底状態から高エネルギー励起状態への一時的な励起を起こす。吸収波長はエネルギーの逆数に比例し、光吸収の強度を波長の関数として表す吸収スペクトルは励起エネルギーと分子の電子状態の遷移確率によって決定する。分子固有の吸光度スペクトルを正確に予測することは、有機発光ダイオード [39, 40] や有機太陽電池 [41], UV フィルター [42] 等の材料開発において重要なタスクである。通常、吸収ピーク波長は、時間依存密度汎関数理論 (time depending density functional theory, TD-DFT) 計算 [43] 等の第一原理計算に基づく電子の励起状態から計算・予測される。ただし、第一原理計算のコストが高いため、網羅的な分子スクリーニングに適用することは難しい。さらに、ピーク波長の位置と強度は第一原理計算から推定できるが、半値幅やある波長域における吸光度積分等、スペクトル全体に現れるその他の関数的特徴を求めることはできない。したがって、第一原理計算に依存せず、データ駆動型アプローチでスペクトル関数を予測できれば、材料研究の発展に大きく貢献で

(a)

Chemical structure



Spectral function



(b)

Composition



Process

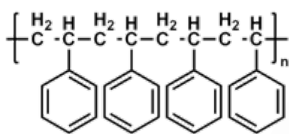
Deposition temperature
Deposition pressure
Average ion energy
Degree of ionization

Microstructure SEM image

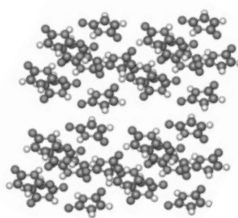


(c)

Chemical structure



Crystal structure



Temperature-dependent physical properties

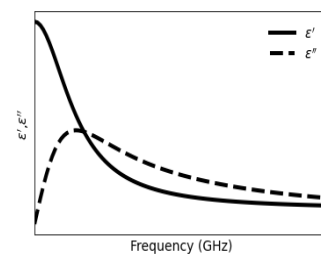


図 2.1 材料科学における関数出力変数の回帰問題の三つの事例；(a) 有機物の化学構造から分光スペクトルを予測するタスク，(b) 組成とプロセス条件から複合材料の微細構造電子顕微鏡像を予測するタスク，(c) 誘電特性の周波数応答曲線を予測するタスク。

きる。

ここで、関数出力変数の回帰の概要を示す。モデル f の入力変数は、分子の化学構造を表す記述子ベクトル X と波長 $t \in \mathbb{R}_+^1$ から成る。出力変数は、ある分子の波長における吸光度を表すスペクトル値 $Y(X, t)$ である。モデルは次のように表される。

$$Y(X, t) = f(X, t) + \epsilon. \quad (2.1)$$

ϵ は観測ノイズを表す確率変数である。 n 個の分子 $\{X_i | i = 1, \dots, n\}$ に対し、スペクトル $Y(X, t)$ は、 m 個の離散化された波長 $\{t_j | j = 1, \dots, m\}$ で観測されると仮定する。すなわち、ここでは波長の観測点は全ての分子に共通であると仮定する。したがって、 t_j は分子のインデックス i に依存しない。観測された波長が分子間で異なる場合は、欠損しているデータ点を補間することで、同じ観測点でのデータ集合を得ることができる。このとき、関数出力回帰の問題を、次のようにベクトル形式で表すことができる。

$$\mathbf{y}(X) = \mathbf{f}(X) + \mathbf{e}. \quad (2.2)$$

ここで、 $\mathbf{y}(X)^\top = (Y(X, t_1), \dots, Y(X, t_m))$, $\mathbf{f}(X)^\top = (f(X, t_1), \dots, f(X, t_m))$, $\mathbf{e}^\top = (\epsilon_1, \dots, \epsilon_m)$ である。

分子の光吸収スペクトルを入力として、その特徴量を抽出する研究は数多く存在する [44]。一方、スペクトル関数全体を出力として予測する研究はこれまであまり注目されてこなかったが、2021年に Urbina らが、化学構造を入力とし、紫外可視分光法により測定した UV-Vis スペクトルを予測する研究を発表した [17]。この研究では、 f を Long Short-Term Memory (LSTM) や注意機構を導入した Seq2Seq [30] でモデル化している。LSTM は自然言語処理や音声認識にも使用される方法で、系列データをモデリングするための再帰的ニューラルネットワークの一種である。この論文では、化学構造を表すトークン化された SMILES 文字列 [45] を直接入力変数とするケースと、事前に定義された分子記述子を入力するケースを考えている。Seq2Seq は、自然言語処理で広く使用されているニューラルネットワークであり、LSTM を利用したエンコーダーとデコーダーを介して入出力の系列データをモデル化する [30]。Seq2Seq では、エンコーダーからデコーダーに固定長の潜在変数ベクトルを渡すことになるため、例えば、入力する文が長文になると、必要な情報が削ぎ落とされてしまう欠点があった。そこで、情報を落とさずデコーダーに引き渡すために注意機構を導入することが有効であることが明らかになり [35]、近年では注意機構を実装した Seq2Seq が広く普及している。Urbina et al. [28] では、LSTM に加えて、注意機構を導入した Seq2Seq による光吸収スペクトルの予測も検証している。ただし、これらのモデルには多数のパラメータが含まれており、モデルの学習には一般に大量のデータが必要となる。

2.2 材料の微細構造の予測

材料研究は、材料の表面や界面構造を観察するために電子顕微鏡の画像データを取得することが多い。物質表面の微細構造は材料の物性や機能に大きな影響を及ぼす。例えば、液晶高分子は分子の配向構造が光の透過性や熱物性の支配因子になる [46]。また樹脂材料の場合、高分子鎖の絡み合いの状態により物質の強度が大きく変化することが知られている [47]。半導体デバイスの薄膜の表面微細構造は、そのモルフォロジーが電気伝導率に大きな影響を及ぼす [48]。特に複合材料では、微細構造の特徴が物性・機能に大きく影響を及ぼすため、表面やモルフォロジーを観察するために、走査電子顕微鏡 (SEM: Scanning Electron Microscope) や透過電子顕微鏡 (TEM: Transmission Electron Microscope) が広く利用されている [49]。SEM では、加速させた電子線を試料に当て、その照射点から発生する信号電子の強度を検出器で検出することで、各ピクセルが信号電子の量に比例した輝度を持つグレースケール画像を得ることができる。TEM では、加速した電子線を試料に照射する点は SEM と同じであるが、試料を透過した電子の干渉像を拡大して観察する点に違いがある。試料の表面構造や内部構造、組成の違いによって生じる電子線の割合の変化を利用して、画像データを得る。SEM の分解能はナノ～マイクロメートルサイズの構造を対象とする。TEM では、電子線が試料を透過する必要があるため、薄く切る必要があり、ナノメートルサイズの構造を対象とする。SEM や TEM の画像から構造的特徴量を定量化するには画像認識やコンピュータビジョンのテクニックを活用できる。一方、微細構造の画像を予測する研究は今のところ初期段階の域を出ていない。例えば、2018 年に発表された Yang et al.[32] や Li et al.[33] の研究では、GAN を用いて微細構造を予測している。また、2020 年に発表された Banko et al.[31] の研究では、プロセス条件を入力とし、条件付き GAN (Conditional GAN, cGAN) [29] を用いて SEM 画像を予測できることを実証している。画像を出力変数として予測する手法が材料開発に普及していけば、様々な応用が進んでいくことが期待される。

2.3 その他の問題設定

ここで、光吸収スペクトルや電子顕微鏡画像以外に材料科学における関数型出力変数を考えてみる。多くの物性は、外部電場の温度や圧力、周波数に依存して決定される。図 2.1 (c) に示すように材料の誘電率や誘電正接といった誘電特性は、周波数と温度の関数として与えられる [50]。この場合、 t は周波数と温度の二つの変数で構成される。さら

に、比体積、線膨張係数、弾性率、比熱、熱伝導率等の物性も温度に依存する。ガラス転移温度や結晶融点、結晶化温度は、物性の温度依存曲線の転移点から計算される。また、結晶系においては、エネルギー上の関数として与えられる状態密度 (density of states, DOS) が、熱や電気の輸送や超電導の決定に重要な役割を果たす。そこで近年、結晶構造の原子の配置情報を入力として DOS を予測する手法が提案されている [51]。また材料の構造解析では、観察対象の大きさに応じて多種多様な顕微鏡が使用される。特に、材料の内部構造を解析するために、様々な 3 次元計測技術が利用される。例えば 3 次元 TEM では、数十から数百ナノメートルの範囲の物体の形態を 3 次元で観察できる [52]。X 線コンピュータ断層撮影 (CT) は、医療用途でよく使用される非破壊検査技術である。この方法では、試料表面を研磨したりエッチングしたりすることなく、数マイクロから数ミリメートルの微細構造を観察できる [53]。さらに、高輝度の放射光を用いた放射光 X 線 CT は、X 線が透過しにくい金属であっても、数百ナノメートルから数マイクロメートルの高分解能で物質の内部構造を非侵襲的に計測できる [54]。このように材料科学には様々な多次元変数のデータが存在するが、これらを予測対象とする研究はほとんど進展がない。

2.4 提案手法の特徴

KRFO は、パラメータ数を抑制し学習プロセスを安定化させることで、十分な量のデータが利用できない場合でも高い予測精度を達成することを目的として、モデリングされている。また、 t の観測点が不均一な場合でも、欠損データ処理を施すことなく学習を実行できる。KRFO のもう一つの利点は、モデルの高い解釈可能性である。スペクトル関数の予測の場合では、記述子 X にエンコードされた分子フラグメントの有無に関連付けて特定の波長範囲でのピークの発生が直接表現される。微細構造の予測では、組成やプロセス条件の値によって画像の特定部位の構造的特徴が生成される。

ここで、KRFO と深層生成モデルの関係を考察する。第 2.2 節で紹介したように、材料の組成とプロセス条件を制御することで、形態的特徴や体積分率、粒子サイズの分布が異なる微細構造を設計できる [31, 32, 33]。例えば、材料組成や温度を制御して緻密で均質な粒子構造を得ることで、材料の機械的特性の向上を図る。微細構造をモデルの出力変数として扱うために、電子顕微鏡で得られた画像を使用することとする。こうすることで、画像認識やコンピュータビジョンの様々な手法を利用することができる。具体的には、入力変数 X は、材料を構成する組成とプロセス条件を表す実数値ベクトルによって与えられ、出力 $Y(X, t)$ はグレースケールあるいはカラー画像を表す行列またはテンソルとして定義される。変数 t は、2 次元または 3 次元の画像座標を示し、ピクセルまたはボクセ

ルの位置がこれに相当する。この多次元配列変数の回帰問題は、条件付き画像生成の機械学習と見なすことができる。

本研究は、最大波長 λ_{\max} といった関数のスカラー特徴量を予測するのではなく、関数全体をまるごと予測することで、データが限られた状況においても高い汎化性能を持つモデルを得ることができることを実験的に明らかにする。第5章の実験では、関数出力変数の回帰モデルで予測したスペクトル関数から計算された λ_{\max} と、スカラー出力変数 λ_{\max} の通常の回帰を比較し、前者の予測精度が後者を大幅に上回ることが確認された。微細構造の電子顕微鏡の予測においても同様の現象が確認されている。多次元出力変数 $\mathbf{y}(X) = (Y(X, t_1), \dots, Y(X, t_m))^T$ の各々の要素を生成する機構は当然ながら独立ではなく、互いに密接に関連し合っている。多変量出力変数の生成機構を単一のモデルで同時に学習するという事は、マルチタスク学習の特殊なケースになっている。マルチタスク学習 [36] では、関連する複数のタスクを同時に学習する過程でモデルがタスク間の共通性を適切に捉えることで、各タスクの予測精度の向上を図る。同様の学習メカニズムが関数出力変数の回帰にも働くことで、少数データに対する汎化性能の安定化が実現されていると予想される。モデルと実験結果の詳細については、第4章と第5章で説明する。

第3章

関数出力回帰の方法

本研究では、関数出力変数を回帰するための統一的なフレームワーク KRFO を構築する。KRFO に関する内容は、Iwayama et al.[38] の研究をまとめたものとなっている。第1章で述べたように、KRFO は、関数データ解析 [34, 55, 56, 57, 58, 59] の分野で研究されてきたモデリング手法と深層学習から着想を得て開発されたモデルとなっている。

以下では、KRFO とその関連手法である関数データ解析、画像生成手法として広く適用されている cGAN、そして、Urbina ら [28] がスペクトル関数の予測で用いた LSTM および注意機構を組み込んだ Seq2Seq を順に説明していく。また、KRFO の少数データに対する高耐性メカニズムに関連していると考えられるマルチタスク学習も本章で取り上げる。

3.1 提案手法：Kernel Regression with Functional Outputs

KRFO は、シンプルで解釈性の高いモデリングを意図して設計されている。KRFO において、関数出力変数 $y(X, t)$ は次のようにモデリングされる。

$$Y(X, t) = \sum_{i=1}^d k(t, s_i) \beta_i(X) + \mu(t) + \epsilon. \quad (3.1)$$

第1項は d 個のカーネル基底関数 $\{k(t, s_i) | i = 1, \dots, d\}$ は重み付きの和となっている。カーネルの中心 s_i は、波長空間あるいは画像の座標空間において等間隔に配置される。本研究では、カーネル基底関数としてガウス基底関数 (RBF) を用いる。

$$k(t, s_i) = l \exp\left(-\frac{\|t - s_i\|^2}{2\sigma^2}\right). \quad (3.2)$$

ここで、分散 $\sigma^2 > 0$ と振幅 $l > 0$ を表すハイパーパラメータは、交差検証における汎化性能の評価に基づき調整される。あるいは、経験的に分かっている実験装置の解像度や対象系の t に対する $Y(X, t)$ の変化に関する事前知識に基づいて、ハイパーパラメータを決定することも考えられる。回帰係数 $\beta_i(X)$ は入力変数 X の関数となっており、各位置に配置されたカーネルの重みを決める。本研究では $\beta_i(X)$ をニューラルネットワークでモデル化する。切片項 $\mu(t)$ は t のみに依存する。後に示すように、正則化として平滑性を課した上で $\mu(t)$ を推定する。 ϵ はノイズ項である。

このモデルは、入力変数 X 変数の値に応じて、 t 上に配置された各カーネル関数が活性化あるいは非活性化されるシステムを表現している。このモデリングの利点の一つは、配置されるカーネル関数の数 d を調整することで、パラメータの数を抑制できる点にある。第 3.3 節で述べる cGAN の生成器では、出力変数の次元数（例えば画像のピクセル数）が出力層のニューロン数と一致するため、関数出力変数を取り扱う場合は必然的にパラメータ数が増加してしまう。過剰パラメータリゼーションは学習の不安定性を引き起こす。KRFO は、カーネル関数の数を任意に調整できるため、パラメータ数を抑制して学習を安定化させることができる。第 5 章で示すスペクトル関数の予測では、スペクトルが観測される波長点の数は 171 や 181 であり、KRFO では配置するカーネル関数の個数を 128 個に設定した。そのため、cGAN の生成モデルのパラメータ数は約 270 万個であるのに対し、KRFO のパラメータ数は約 150~200 万に抑えられる。さらに近赤外スペクトル予測のケースでは、スペクトルが観測される波長点の数は 2,151 となり、KRFO で配置するカーネル関数の個数を 256 個に設定した。その結果、cGAN のパラメータ数は約 1,060 万個であるのに対し、KRFO はその約 1/6 に相当する約 180 万個に抑えられる。また、微細構造の 2 次元画像の解析では、cGAN の生成器のパラメータ数は約 580 万個であるのに対し、KRFO のパラメータ数は約 270 万個となる。

スペクトル関数の予測と微細構造の予測の二つのタスクにおいて、回帰係数 $\{\beta_i(X) | i = 1, \dots, d\}$ はニューラルネットワークを用いてモデリングする。KRFO のモデルの全体概要を図 3.1 に示す。二つのケースにおけるモデリングの詳細は第 4 章で述べるが、ここではモデルの概要を示す。入力 X は、Extended Connectivity Fingerprint（半径 3 の ECFP フィンガープリント）[12] を用いて分子の化学構造を数値化した長さ 1,024 のバイナリーベクトルとした。ベクトルの各要素はある部分構造の有無を表す。 X から d 個の回帰係数へのマッピングは、全結合層、バッチ正規化層、LeakyReLU 活性化関数から成る隠れ層を積層したニューラルネットワークでモデリングした。微細構造の画像予測では、入力 X は、材料組成とプロセス条件を 6 次元の実数ベクトルと 1 次元のランダムノイズとなる。 X から回帰係数へのマッピングは、転置畳み込み層、バッチ正規化レイヤー、

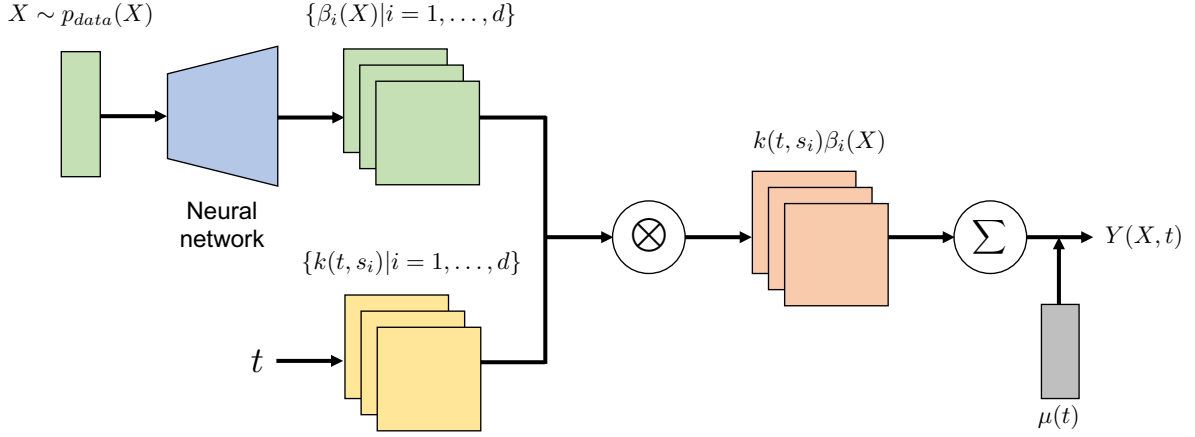


図 3.1 KRFO のモデル構造. 事前に定義された d 個のカーネル $\{k(t, s_i) | i = 1, \dots, d\}$ に対する回帰係数 $\{\beta_i(X) | i = 1, \dots, d\}$ は, 入力変数 X の関数となる. 入力変数 X から $\beta_i(X)$ への変換は, 全結合あるいは畳み込みニューラルネットワークによってモデル化される. カーネル関数の数 d は t の観測点数 m よりも小さくなるように選択する ($d \leq m$).

LeakyRelu 活性化関数からなる隠れ層を積層したモデルで表現される. 層数やニューロン数等のハイパーパラメータは交差検証に基づき調整した. 探索対象のモデルの構造やハイパーパラメータについては, 第 4 章で詳しく述べる.

モデルの学習では, $\{\beta_i(X) | i = 1, \dots, d\}$ のニューラルネットワークとベースライン $\mu(t)$ のパラメータが推定対象となる. 具体的には, 次の経験リスクを最小にするようにモデルを訓練する.

$$L(\beta, \mu) = \sum_{(X,t) \in \mathcal{D}_{\text{obs}}} C(Y(X,t), \hat{Y}(X,t)) + \lambda \sum_{i=1}^m \sum_{j \in \mathcal{A}_i} (\mu(t_i) - \mu(t_j))^2. \quad (3.3)$$

第 1 項は, 訓練データ集合 \mathcal{D}_{obs} に含まれる観測値 $Y(X,t)$ とその予測値 $\hat{Y}(X,t) = \sum_{i=1}^d k(t, s_i) \beta_i(X) + \mu(t)$ の間の乖離度を損失関数 C で測る. 後で示すスペクトル関数の予測と微細構造の予測では, C として, 二乗損失関数を用いる. 第 2 項は, ベースライン関数の推定における正則化項である. 正則化により, 観測点 t_i とその近傍 $t_j \in \mathcal{A}_i$ における $\mu(t_i)$ と $\mu(t_j)$ の間の乖離に罰則をかけることで, 滑らかなベースライン関数の推定を促す. 式 (3.3) より, t に対する $Y(X,t)$ のデータに欠損値があっても本手法を適用できることが分かる.

3.2 関連手法 1：関数データ解析における回帰

関数データとは、例えば、気温や降水量の時空間データのように、任意の観測点において本質的に連続的に変化する確率変数の観測値を意味する。観測点の順序や位置は物理的な意味を持つ。関数データ解析という分野は、1950年代の Grenander[55] や Rao[56] らの研究から始まり、1990年代に Ramsay と Silverman[57, 58] らにより体系化されたといわれている。

関数データ解析では、時間や空間上で離散的に観測されたデータに何らかの統計処理を施して関数に変換した後、関数化されたデータに対して回帰分析や分類等の統計解析を実施する。データに関数化を施すことで、観測ノイズの除去や欠損値補完による観測点数の水増しを図る。また、アプリケーションによっては微分を計算できるようになることも大きな利点になる。ここでサンプル j ($j = 1, \dots, n$) の観測点 t におけるデータ $Y_j(t)$ は、真の値 $Y_j(t)^*$ と加法ノイズ ϵ からなると仮定する。

$$Y_j(t) = Y_j(t)^* + \epsilon. \quad (3.4)$$

関数 $Y_j(t)^*$ を離散観測点上の $Y_j(t)$ のデータから推定するために、 d 個の基底関数 $\psi_1(t), \dots, \psi_d(t)$ と係数パラメータ $\beta_{1,j}, \dots, \beta_{d,j}$ を用いて、 $Y_j(t)^*$ を次のようにモデリングする。

$$Y_j(t)^* = \mu(t) + \sum_{i=1}^d \psi_i(t) \beta_{i,j}. \quad (3.5)$$

基底関数 $\psi_i(t)$ としては、B-スプライン関数やフーリエ基底などが使われることが多い。また、関数主成分分析から得られた固有関数を基底関数とすることもある。関数主成分分析では、標本共分散関数

$$c(t, s) = \frac{1}{n} \sum_{i=1}^n (Y_i(t) - \bar{Y}(t))(Y_i(s) - \bar{Y}(s)), \quad \bar{Y}(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \quad (3.6)$$

の固有関数 $\psi(t)$ を求める。

$$\int c(t, s) \psi(s) ds = \lambda \psi(t). \quad (3.7)$$

大きいほうから d 個の固有値に相当する固有関数を正規直交基底として、 $Y_j(t)^*$ をモデリングする。このとき、係数 $\beta_{1,j}, \dots, \beta_{d,j}$ の最小二乗推定量は、固有関数の直交性の条件

表 3.1 入出力変数の形式による関数回帰のモデルの分類.

目的変数\入力変数	スカラー	関数
スカラー	一般的な回帰モデル	スカラー—関数回帰
関数	関数—スカラー回帰	関数—関数回帰

の下, $Y_j(t) - \bar{Y}(t)$ と固有関数の内積となる. この表現は Karhunen–Loève 展開といわれることもある.

式 (3.4) と式 (3.5) をまとめると, 観測データ $Y_j(t)$ の生成過程は次のように表される.

$$Y_j(t) = \mu(t) + \sum_{i=1}^d \psi_i(t) \beta_{i,j} + \epsilon. \quad (3.8)$$

ここで, 係数とサンプル j の間に何らかの数学的構造が依存するケースを考える. サンプルを特徴付ける変数 (あるいはインデックス) を X と書けば, 式 (3.8) を次のように表すことができる.

$$Y(X, t) = \mu(t) + \sum_{i=1}^d \psi_i(t) \beta_i(X) + \epsilon. \quad (3.9)$$

ここで基底関数にカーネル $\psi_i(t) = k(t, s_i)$ を設定すれば, KRFO と同等の表現が得られる. すなわち, 基底展開による離散データの関数化において, KRFO は, サンプル依存的に変化する係数を外部変数 X を入力とする統計モデルで表現していることに相当する.

入力変数と出力変数のいずれかに関数データを含む回帰分析のことを関数回帰分析という. 関数回帰のモデルは, 表 3.1 のように入出力変数の形式により四つのカテゴリに分類される.

まずは, 入力変数が関数データ, 目的変数がスカラーの問題設定を考える. 入力変数のサンプル j の関数データを $X_j(t)$, 目的変数のスカラー値を Y_j とすると, 関数線形モデルは次式で与えられる.

$$Y_j = \mu + \int_{\mathcal{T}} X_j(t) \beta(t) dt + \epsilon_j. \quad (3.10)$$

μ は切片項, $\beta(t)$ は回帰係数関数, $\epsilon_j \sim N(0, \sigma^2)$ は観測誤差である. \mathcal{T} は観測点 t の定義域を表す.

次に、入力変数がスカラーを含む実数ベクトル、出力変数が関数データの問題設定を考える。本研究の問題設定はこのケースに該当する。サンプル j の d 個の入力変数を $X_{i,j}$ ($i = 1, \dots, d$)、目的変数の関数データを $Y_j(t)$ とすると、関数線形モデルは次式で与えられる。

$$Y_j(t) = \mu(t) + \sum_{i=1}^d X_{i,j} \beta_i(t) + \epsilon_j(t). \quad (3.11)$$

$\mu(t)$ は切片項の関数、 $\beta_i(t)$ は変数 i の回帰係数関数、 $\epsilon_i(t)$ は誤差関数である。目的変数 $Y_j(t)$ の t による変化を関数 $\beta_i(t)$ で表現している。

最後に、入力変数と出力変数が共に関数形式で与えられる問題設定を考える。入力変数を $X_j(s)$ 、目的変数を $Y_j(t)$ とする。このとき関数線形モデルは次式で与えられる。

$$Y_j(t) = \mu(t) + \sum_{i=1}^d \int_{\mathcal{T}} X_{i,j}(s) \beta_i(s, t) ds + \epsilon_j. \quad (3.12)$$

$\beta_i(s, t)$ は回帰係数関数、 $\epsilon_j(t)$ は誤算関数である。回帰係数関数は 2 変数関数として与えられる。また、関数—関数データ回帰では、入力変数と出力変数が同じ t の関数として与えられる場合も考えられる。このときモデルは次式で与えられる。

$$Y_j(t) = \mu(t) + \sum_{i=1}^d X_{i,j}(t) \beta_i(t) + \epsilon_j. \quad (3.13)$$

このモデルでは t の入力変数 $X_{i,j}(t)$ のみが目的変数 $Y_i(t)$ に影響を及ぼすことから、関数同時モデルと呼ばれる。

関数データ解析では、このように入力変数をスカラー、出力変数を関数データとする回帰手法が研究されてきた。ここで、本研究で取り扱うケーススタディとして、化学構造からのスペクトル関数の予測や組成・プロセス条件から微細構造の予測との対応を考えてみる。両ケース共に、関数回帰モデルは入力変数がスカラー（ベクトル）、目的変数が関数となり、問題形式は関数—スカラー回帰に分類される。化学構造からスペクトル関数を予測する場合、化学構造によってスペクトル関数が異なる。スペクトルは波長 t の関数である。また、組成・プロセス条件から微細構造を予測するケースでは、画像のピクセルの位置が t となる。しかしながら、通常の間数—スカラー回帰では、もう一つの説明変数である化学構造や組成・プロセス条件の影響をモデリングできない。そこで、関数データ解析のモデリングの自然な拡張として、第 3.1 節で述べた KRFO を考案した。

関数データ解析の中で、関数時系列解析という問題設定がある。今、ある時刻あるいは時間間隔 i における関数データの集合 $\{X_j(t) | j = 1, \dots, n\}$ が与えられているとする。こ

この関数時系列解析の目的は、このデータ集合を用いて将来の時点 $n+h$ ($h > 0$) の関数 $X_{n+h}(t)$ を予測することである。そこでまず、 n 個のサンプルを用いてデータ集合の関数化を行う。

$$X_j(t) = \mu(t) + \sum_{i=1}^d \psi_i(t) \beta_{i,j}. \quad (3.14)$$

例えば、 ψ として、関数主成分分析から得られた d 個の正規直交基底を用いる。ここで、予測対象の時点 $n+h$ の関数は、次のように表される。

$$X_{n+h}(t) = \mu(t) + \sum_{i=1}^d \psi_i(t) \beta_{i,n+h}. \quad (3.15)$$

このとき、予測すべき変数は係数 $\beta_{1,n+h}, \dots, \beta_{d,n+h}$ である。そこで、Hyndman et al.[60] や Aue et al.[61] は、 n 個の多変量時系列データ $\{(\beta_{1,i}, \dots, \beta_{d,i}) | i = 1, \dots, n\}$ を用いて時系列モデルを推定し、 $(\beta_{1,n+h}, \dots, \beta_{d,n+h})$ の予測値を式 (3.15) に代入して $X_{n+h}(t)$ を予測することを提案している。これらの関数時系列解析の手法と KRFO は、係数集合の変化に何らの数学的構造が存在すると仮定し、係数集合を予測する統計モデルを構築することで、将来の関数データを予測しようという点において思想的共通性を見出すことができる。

3.3 関連手法 2 : Conditional GAN

Mirza et al.[29] によって提案された Conditional GAN (cGAN) [29] は、GAN の生成器と識別器に生成対象の変数に紐づけられたラベルを導入することで、式 (3.16) を用いてラベルで条件づけられた画像を生成する手法である。cGAN は、生成器 (G) と呼ばれる生成モデルと、識別器 (D) と呼ばれる 2 値分類器から構成される。モデルの構造を図 3.2 に示す。詳細については、第 4 章で説明する。

生成器 $Y = G(X, Z)$ は、式 (2.2) に示すように、スペクトル関数の m 個の値や、グレースケール画像の行列 Y を出力する。入力変数は、ラベル X とランダムノイズ Z で構成される。cGAN では X は条件変数と呼ばれる。ノイズ $Z \sim N(0, 1)$ は平均が 0 で分散 1 の正規分布に従うと仮定される。例えば、入力 (X, Z) は、バッチ正規化層と Leaky ReLU 活性化関数、全結合埋め込み層を通して埋め込みベクトルに変換され、さらにスペクトルや微細構造画像を表すベクトル $Y = \mathbf{y}(X)$ に変換される。識別器 $D(X, Y)$ は二値分類器であり、条件変数 X と Y を入力とし、 Y が G から生成された偽のデータである

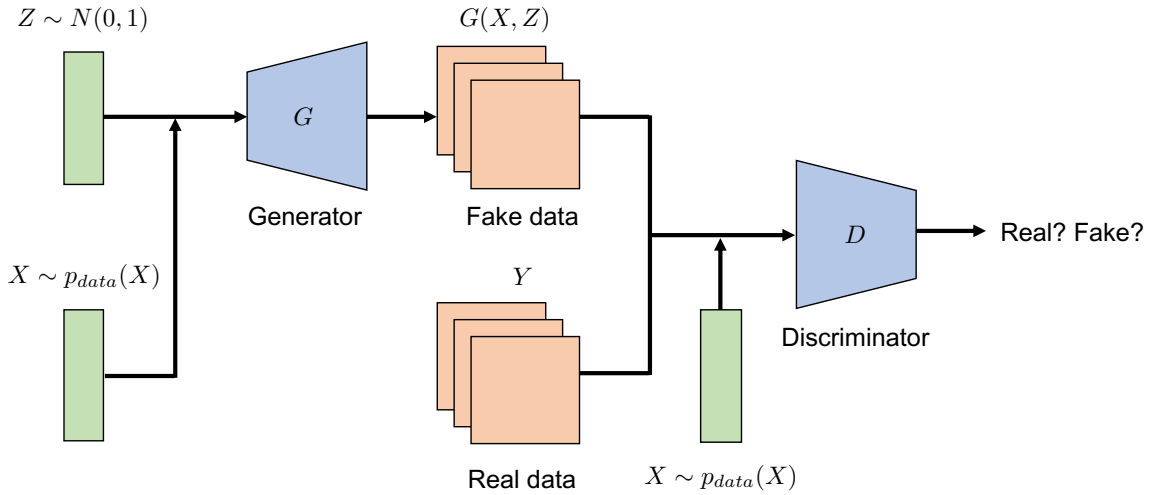


図 3.2 cGAN のモデルアーキテクチャー. 生成器 G は, 組成・プロセス条件や分子の記述子 X から走査電子顕微鏡像または UV-Vis-NIR 吸収スペクトル関数へのマッピングを行う. 入力 Z は, ランダムなガウスノイズ. 識別器 D は, X に加えて, 微細構造電子顕微鏡像あるいはスペクトル関数が与えられると, 入力されたデータが本物か偽物かを判定する.

か, 実際のデータであるかを判定する. 識別器 D はニューラルネットワークを用いてモデリングする. 層の数や各層のニューロン数などのモデル構造は, 基本的な形式を維持しながら, 検証データ集合に対する汎化性能に基づいて決定される. G と D は次のコスト関数を用いた敵対的学習により訓練される.

$$\min_G \max_D \mathbb{E}_{(X,Y) \sim p_{data}(X,Y)} [\log D(X,Y)] + \mathbb{E}_{Z \sim N(0,1), X \sim p_{data}(X)} [\log(1 - D(X, G(X, Z)))]. \quad (3.16)$$

第 1 項は, 識別器 $D(X, Y)$ が真の Y を正しく真であると判定できた場合大きくなる. 第 2 項は $1 - D(X, G(X, Z))$ が大きくなったとき, つまり, 識別器 D が生成器が生成した $G(X, Z)$ を偽物と判定したときに大きくなる. 識別器 D は, 分類誤差が最小になるように式 (3.16) を最大化するように学習される, 逆に生成器 G は D が誤認識するように, すなわち, コスト関数が大きくなるように学習される. G と D を交互に学習することにより, 任意の記述子 X に対して, 高品質の偽のスペクトル関数や微細構造画像を生成できる G が得られる. cGAN は, 多次元出力変数の教師あり学習の手法と見なすことができる. 特に cGAN は画像生成の手法として様々な研究がなされており, 画像を出力変数

として扱う場合、機械学習の研究で蓄積された様々な成果を利用できる。ただし、他の従来の敵対的生成ネットワークと同様に、cGAN は学習プロセスの不安定さという問題を持つ。特に、少数データセットに対しては脆弱であることが指摘されている。学習データが限定されている場合、識別器は簡単にデータに過適合し、本物と偽物を完全分類してしまう。これにより、モデルの訓練中に勾配が消失し、学習プロセスが停止してしまう。敵対的学習プロセスを安定させる解決策として、 G と D の更新頻度のバランスをとることが提案されているが、簡単ではない。主な理由は、 Y が高次元であるため生成器のパラメータ数が膨大になることにある。本研究でのケーススタディでは、 Y の次元は、スペクトル関数予測では 2,000 を超え、解像度 100×100 の微細構造顕微鏡画像では、最大 10,000 と高次元となり、それに伴い生成器のパラメータ数も大幅に増加する。

3.4 関連手法 3：Long Short-Term Memory

Urbina et al.[28] は、Long short-term memory (LSTM) と注意機構を用いたエンコーダー・デコーダーモデルを用いて、化学構造から UV-Vis スペクトル関数を予測している [28]。LSTM[62] は時系列データの解析のために、Recurrent Neural Network (RNN) の改良版として 1995 年に発表されたモデルである。時系列データの予測では、ある時点までのデータ集合から次に起こりうるデータを予測する。通常の入力変数とは異なり、入力ベクトルの隣り合う要素は互いに依存関係を持つ。RNN は、過去の情報から現在の値を計算する再帰的構造を有しており、中間層にて、ある時刻 t の入力層からの入力と前の時刻 $t-1$ の中間層からの情報を加算することにより、短期的な過去の情報を現在に引き継ぐ。具体的には、時系列データ $\{x_1, \dots, x_T\}$ を入力としたとき、 $t+1$ の予測値を y_{t+1} 、時刻 t での入力層から次の隠れ層の重みとバイアスを W_{xh} , b_{xh} 、中間の隠れ層の重みとバイアスを W_{hh} , b_{hh} 、出力層の一つ前の隠れ層から出力層の重みとバイアスを W_{hy} , b_{hy} とし、次のように $t+1$ の y_{t+1} を計算する。

$$z_t = f^{(\text{hidden})}(W_{xh}x_t + W_{hh}z_{t-1} + b_{xh} + b_{hh}), \quad (3.17)$$

$$y_{t+1} = f^{(\text{out})}(W_{hy}z_t + b_{hy}). \quad (3.18)$$

短期的な過去の情報は $W_{hh}z_{t-1}$ を介して継承される。RNN の逆誤差伝搬では、微分連鎖律に基づく勾配計算において、時点ごとに損失関数の微係数を計算し、伝搬対象の時間幅の分だけ乗算することになる。このとき、時間幅が大きくなると勾配の発散や消失が生じるため、長期的な情報を学習に反映できないという問題がある。LSTM は RNN の勾配の発散や消失を抑制するために拡張されたモデルである [62]。LSTM では、長期記憶

を学習に反映できるように、入力ゲートと忘却ゲート、出力ゲートと呼ばれる三つのゲートによって $t+1$ の予測値が計算される。RNN では一つ前の時刻からの情報伝播だけであったが、LSTM ではそれに加え、忘却ゲートと呼ばれるゲートにて、過去の情報から重要な情報と不必要な情報を取捨選択することにより、重みの発散を防ぎつつ、長期的な記憶を継承することができる。

3.5 関連手法 4：注意機構を用いた Seq2Seq

Seq2Seq は、自然言語処理で用いられる手法であり自動翻訳のために開発された手法である。エンコーダーと呼ばれる入力情報を圧縮していく工程と、その圧縮を復元していくデコーダーと呼ばれる工程で構成される生成モデルの一つである。Seq2Seq はエンコーダー、デコーダー共に RNN を利用したモデル構造を有している。例えば、日本語を英語に翻訳する場合には、エンコーダーは日本語文を入力として、RNN を通じて単語間の情報を受け取りつつ特徴量を出力する。次にエンコーダーの最後の隠れ層のパラメータ情報をデコーダーに渡す。デコーダーは、正解となる英語文を入力として、エンコーダーから受け取った隠れ層のパラメータを初期状態とし、入力された英語文が再現されるように学習を行う。Seq2Seq では長文を扱うことが難しいため、解決策の一つとして、注意機構と呼ばれるモデリング手法が提案され、学習精度が大幅に向上した [35, 63]。注意機構では、エンコーダーの全ての隠れ層の情報をデコーダーに渡し、スコアと呼ばれる重みをかけて重み付き平均で単語を予測する。単語ごとに異なる重みをかけることで、長文にも対応が可能となる。スペクトル関数に適用する場合は、エンコーダーへの入力は、化学構造を埋め込みベクトルに変換したフィンガープリントであり、デコーダーへの入力は、波長点ごとのスペクトル強度となる。

3.6 関連手法 5：マルチタスク学習

マルチタスク学習では、関連する複数のタスクを同時に学習することにより、各タスクを個別に学習するよりも優れた性能を引き出すことを目的とする [36, 37, 64]。人間の学習の仕組みと同様に、例えば英語を習得している日本人が他言語を習得する際に、外国語を何も習得していない日本人と比べて、容易に他言語を習得できる現象と同様のことが機械学習にも起こり得る。あるタスクから学習した知識を他の関連するタスクの学習に転用するため、マルチタスク学習は転移学習の一種ともいえる。教師あり学習におけるマルチタスク学習では、 p 個のタスク $\{T_i\}_{i=1}^p$ があり、各タスク i に対してデータセット

$\{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ が与えられる。 x_j^i は入力、 y_j^i はそのラベルである。 タスク $\{T_i\}_{i=1}^p$ は、 y_j^i が連続値の場合は回帰、離散値の場合は分類となる。 マルチタスク学習の目的は、 p 個の関数 $\{f_i(x)\}_{i=1}^p$ はタスク間共通成分と固有成分からなり、与えられたデータからこれらの成分を学習することである [65]。 深層学習によるアプローチでは、Hard Parameter Sharing と Soft Parameter Sharing の二つのアプローチがある。 Hard Parameter Sharing は、マルチタスク学習で一般的に使用されるアプローチである [66, 67]。 Hard Parameter Sharing は、全てのタスク間で入力から k 個の隠れ層を共有し、 $k + 1$ 個目の隠れ層からタスク固有のパラメータを学習する手法である。 もう一方の Soft Parameter Sharing のアプローチの一つである Cross-Stitch Networks では、あるタスク A とタスク B に対して、同じネットワークアーキテクチャを有するニューラルネットワークが与え、それぞれの i 番目の隠れ層の j 番目のユニットに含まれた隠れた特徴量を $x_{i,j}^A$, $x_{i,j}^B$ とする。 このとき、タスクの関連性は次式で計算される。

$$\begin{pmatrix} \tilde{x}_{i,j}^A \\ \tilde{x}_{i,j}^B \end{pmatrix} = \begin{pmatrix} \alpha_{AA} & \alpha_{AB} \\ \alpha_{BA} & \alpha_{BB} \end{pmatrix} \begin{pmatrix} x_{i,j}^A \\ x_{i,j}^B \end{pmatrix}. \quad (3.19)$$

二つのネットワークのパラメータと重みを学習することにより、タスク間の特徴量を共有した学習ができる。

第 4 章

モデル設計と性能指標

第 3 章で紹介した KRFO と cGAN について、具体的なモデルの設計方法とハイパーパラメータの調整方法の詳細を示す。また、スペクトル関数の予測および微細構造の電子顕微鏡像予測において用いた性能検証指標の詳細を示す。

4.1 Kernel Regression with Functional Outputs のモデル設計

化学構造を入力とした UV-Vis-NIR 吸収スペクトルの予測と組成やプロセス条件を入力とした微細構造の予測の二つのケースにおける KRFO のモデルの基本的な形は同じである。

スペクトル関数を予測する場合の入力変数 X は、半径 3 の ECFP フィンガープリントで化学構造を数値化した 1,024 ビット ($p = 1,024$) の二値ベクトルである。モデルは全結合層、バッチ正規化層、および LeakyReLU 活性化関数を積層した構造をとる。KRFO に用いるカーネル基底関数はガウス基底関数 (RBF) を選択した。RBF カーネル以外にも、tricube カーネル関数、Epanechnikov カーネル関数、一様カーネル関数を用いて検証を行ったが、予測精度にほとんど差がなかったため、本論文では RBF カーネルの結果のみを報告する。探索したハイパーパラメーターセットは、隠れ層の層数と、RBF の高さ $l \in \{0.5, 1, 5, 10\}$ と分散 $\sigma^2 \in \{0.0005, 0.0001, 0.00005\}$ である。隠れ層の数は 1 から 4 の範囲で探索を行なった。各層数のモデルのニューロン数は、図 4.1 に示すように、隠れ層の数に基づき、入力と出力変数の次元数を線形補間するように決定した。

微細構造を予測場合の入力変数 X は、化学組成とプロセス条件を表す 6 次元ベクトルである。スペクトル関数の場合とは異なり、画像の場合、同一被写体だとしても撮影位置や角度によって出力画像が変化するため、ある組成とプロセス条件の組み合わせに対

応する SEM 画像は一意に決まらない。このような画像特有の多様性を表現するためにノイズ Z を入力変数に加える。なおノイズ Z を単純にガウス分布からランダムにサンプリングしてしまうと、誤差逆伝播法による微分が不可能になるため、変分オートエンコーダー [68] の学習に広く適用されている reparameterization trick を用いる。ガウス分布に従う乱数 $\kappa \sim N(0, 1)$ を用いてノイズ Z は $Z_i = \tau_i(X) + \kappa\eta_i(X)$ で定めた。平均 τ_i と標準偏差 $\eta_i(X)$ は X の関数で、ニューラルネットワークで記述される。なお、 X は平均が 0、分散が 1 になるように正規化した。スペクトル関数の予測とは異なり、モデル構造は、転置畳み込み層、バッチ正規化層、Leaky ReLU 活性化関数を積層した構造により構成した。スペクトル関数の場合と同様に、RBF カーネル以外にも、tricube カーネル関数、Epanechnikov カーネル関数、一様カーネル関数を用いた検証を実施したが、同じく予測精度への影響は小さかったため、本論文では RBF カーネルのみ結果のみを報告する。探索したハイパーパラメータセットは隠れ層の数、RBF カーネルの高さ $l \in \{0.5, 1, 5, 10\}$ と分散 $\sigma^2 \in \{0.0005, 0.0001, 0.00005\}$ とし、交差検証法によりこれらの値を決定した。隠れ層の数は 1 から 4 の範囲で探索を行なった。各層数のモデルのニューロン数は、図 4.2 に示すように、隠れ層の数に基づき、入力と出力変数の次元数を線形補間するように決定した。隠れ層の数に対するニューロン数詳細は、表 4.1, 表 4.2 にて示す。スペクトル関数の予測と微細構造の予測共に、検証データセットに対して、平均平方二乗誤差 (root mean square error, RMSE) を最小にするハイパーパラメータを選択した。

4.2 Conditional GAN のモデル設計

cGAN においても、二つのケーススタディにおけるモデルの基本構造は共通である。生成器は入力変数 $X \in \mathbb{R}^p$ とガウスノイズ $Z \in \mathbb{R}^q \sim N(0, 1)$ から出力変数 $Y \in \mathbb{R}^m$ にマッピングする。入力変数は、全結合層、バッチ正規化層および LeakyReLU 活性化関数を介して埋め込み空間に転送され、複数の隠れ層と全結合層またはシグモイド層を介して出力変数に変換される。識別器の入力は本物あるいは偽物の Y と条件変数 X であり、 Y が本物か偽物かを識別する。識別器は二つのタスクに対してそれぞれ、全結合のニューラルネットワークと畳み込みニューラルネットワークでモデルを構築した。

スペクトル関数予測の入力変数 X は、半径 3 の ECFP フィンガープリントで化学構造を 1,024 ビット ($p = 1,024$) の二値ベクトルに定めた。生成器のノイズ Z の次元は $q = 100$ に設定した。したがって、生成器の入力変数の次元は、化学構造の 1,024 ビット ($p = 1,024$) とノイズ Z ($q = 100$) を合わせた 1,124 となる。埋め込み層と出力層を固定し、隠れ層の数を 1 から 4 まで変化させ、各繰り返し単位を全結合層、バッチ正規化、

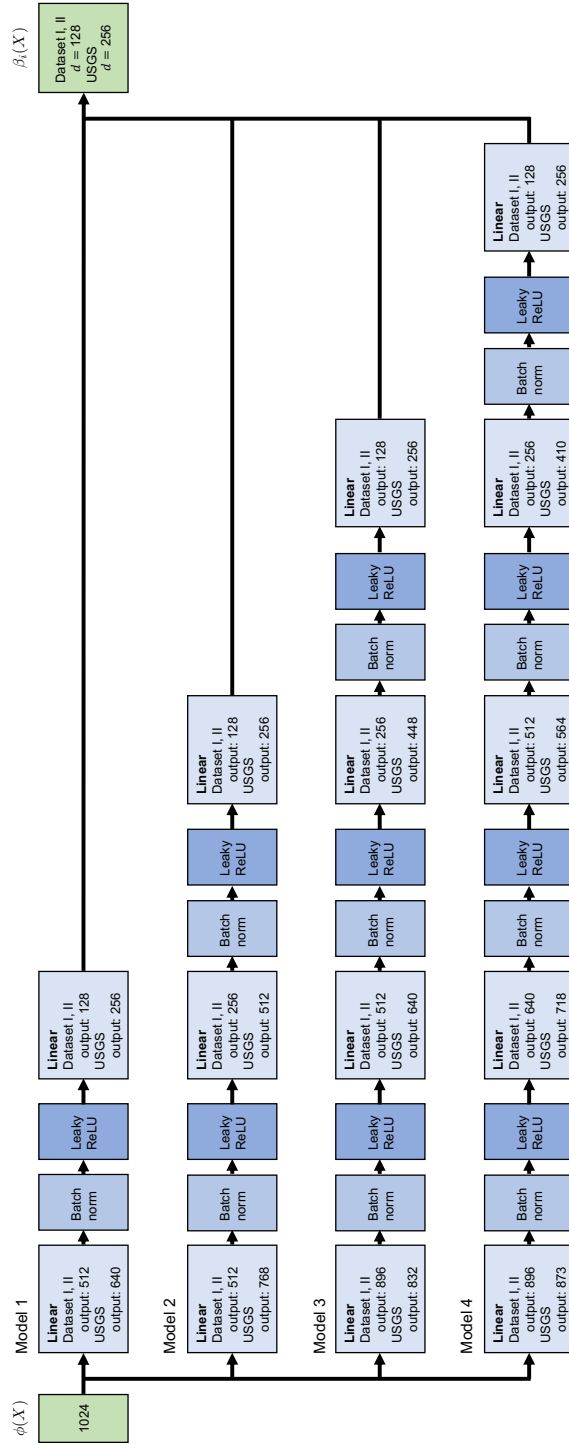


図 4.1 スペクトル関数予測における KRFO 候補モデルのネットワーク設計.

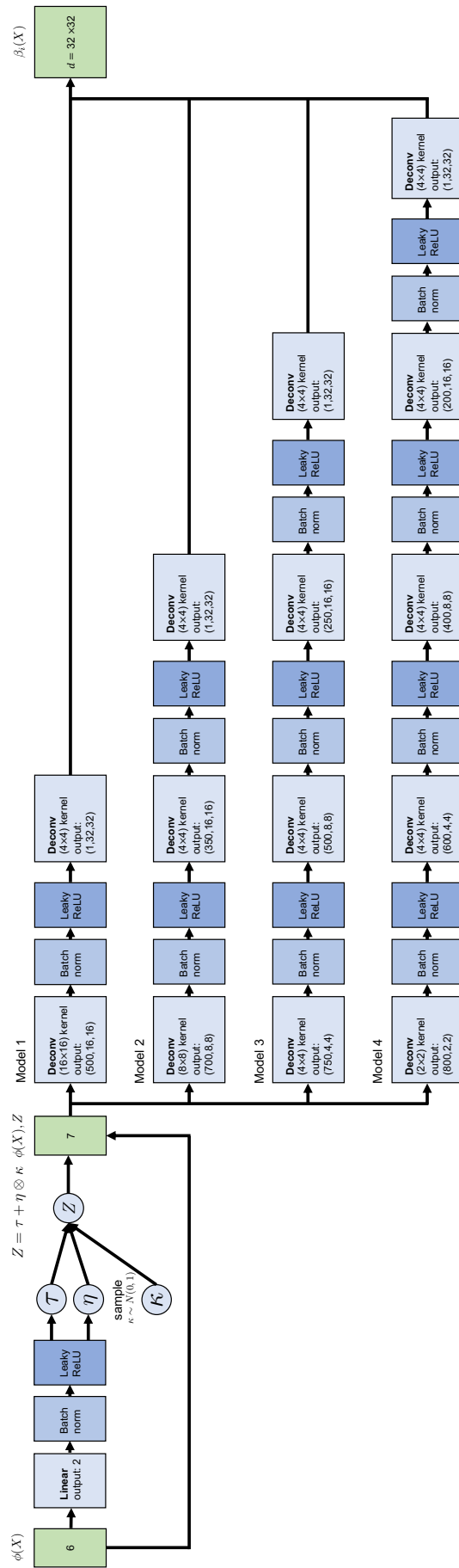


図 4.2 微細構造予測における KRFO 候補モデルのネットワーク設計.

表 4.1 スペクトル関数予測における KRFO 候補モデルのネットワーク構造. アスタリスクは, モデルの検証によって選ばれた最も性能が高いモデルを表す.

Module	Dataset	Input	Layer1	Layer2	Layer3	Layer4	Output
1	Dataset I	1024	512				128
	Dataset II	1024	512				128
	USGS	1024	640				256
2	Dataset I	1024	512	256			128
	Dataset II	1024	512	256			128
	USGS	1024	768	512			256
3	Dataset I	1024	896	512	256		128
	Dataset II*	1024	896	512	256		128
	USGS*	1024	832	640	448		256
4	Dataset I*	1024	896	640	512	256	128
	Dataset II	1024	896	640	512	256	128
	USGS	1024	873	718	564	410	256

表 4.2 微細構造予測における KRFO 候補モデルのネットワーク構造. アスタリスクは, モデルの検証によって選択された最も性能が高いモデルを表す.

# of layers	Input	Layer0 (Linear)	sampling + noise	Layer1 (Deconv)	Layer2 (Deconv)	Layer3 (Deconv)	Layer4 (Deconv)	Output
1	6	2	7	(500, 16, 16)				(1, 32, 32)
2	6	2	7	(700, 8, 8)	(350, 16, 16)			(1, 32, 32)
3	6	2	7	(750, 4, 4)	(500, 8, 8)	(250, 16, 16)		(1, 32, 32)
4*	6	2	7	(800, 2, 2)	(600, 4, 4)	(400, 8, 8)	(200, 16, 16)	(1, 32, 32)

LeakyReLU 活性化関数で構成した上で各モデルの予測性能を検証した。各層のニューロンの数は、隠れ層の数に応じ、入力と出力の長さを丸め込んだ整数で線形補間するように設定した。スペクトルの予測では、データセット I, データセット II, USGS データセットという三つのデータを解析する。このケースの Y の次元は 181, 171, 2,151 となる。識別器の入力変数の次元は、化学構造フィンガープリントの 1,024 ビット ($p = 1,024$) と Y の次元を合わせて、データセットごとに、1,205, 1,195, 3,175 となる。識別器のネットワーク構造は、基本的に生成器と同じ方法で設計したが、図 4.3 および図 4.4 に示すように、バッチ正規化層はドロップアウト層に置き換えられている。

微細構造予測の入力変数は、組成とプロセス条件を表す 6 次元ベクトル X とノイズ Z から成る。ノイズ Z の次元は 100 とした。 X は、訓練データにおいて各要素が平均が 0, 分散が 1 になるように正規化した。スペクトル関数の予測とは異なり、モデルの構造は、転置畳み込み層、バッチ正規化層、LeakyReLU 活性化関数を積層して構築した。モデルの性能評価は、スペクトル関数のケースと同様に、隠れ層の数を 1 から 4 まで変化させて、最良のモデル構造を決定した。二つのタスクにおいて、検討したパラメータの組み合わせを図 4.3, 図 4.4, 図 4.5, 図 4.6, 表 4.3, 表 4.4, 表 4.5 に示す。スペクトル関数の予測と微細構造の予測共に、検証データセットに対して RMSE を最小にするハイパーパラメータを選択した。

4.3 性能評価指標

n 個のテスト分子 $\{X_i | i = 1, \dots, n\}$ のスペクトル関数の m 個の波長 $\{t_j | j = 1, \dots, m\}$ に対し、RMSE, 平均絶対誤差 (mean absolute error, MAE), 決定係数 (R^2)), 予測ス

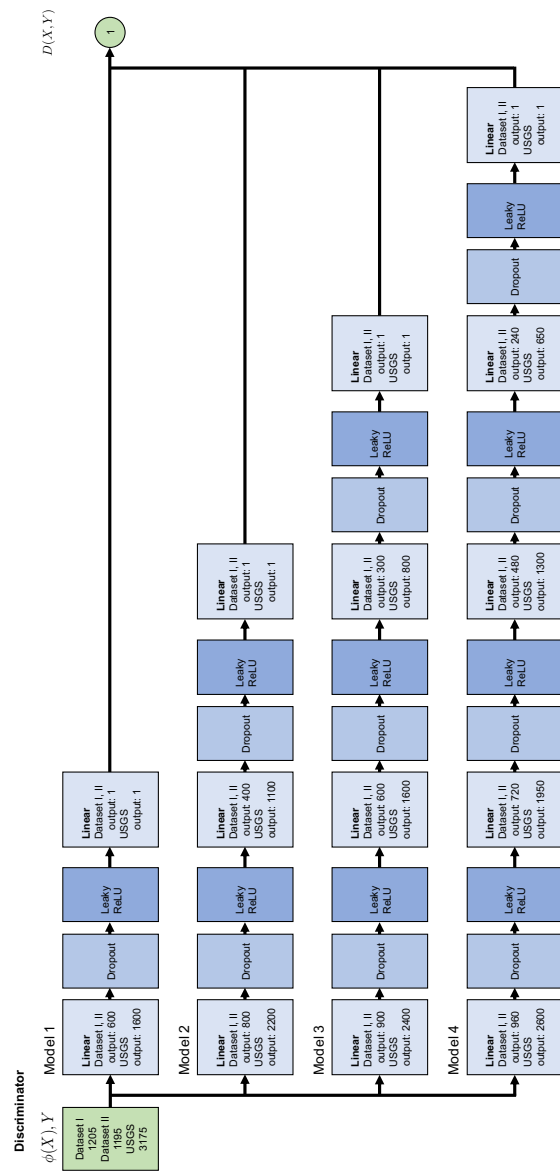


図 4.3 スペクトル関数予測における cGAN 識別器候補モデルのネットワーク設計.

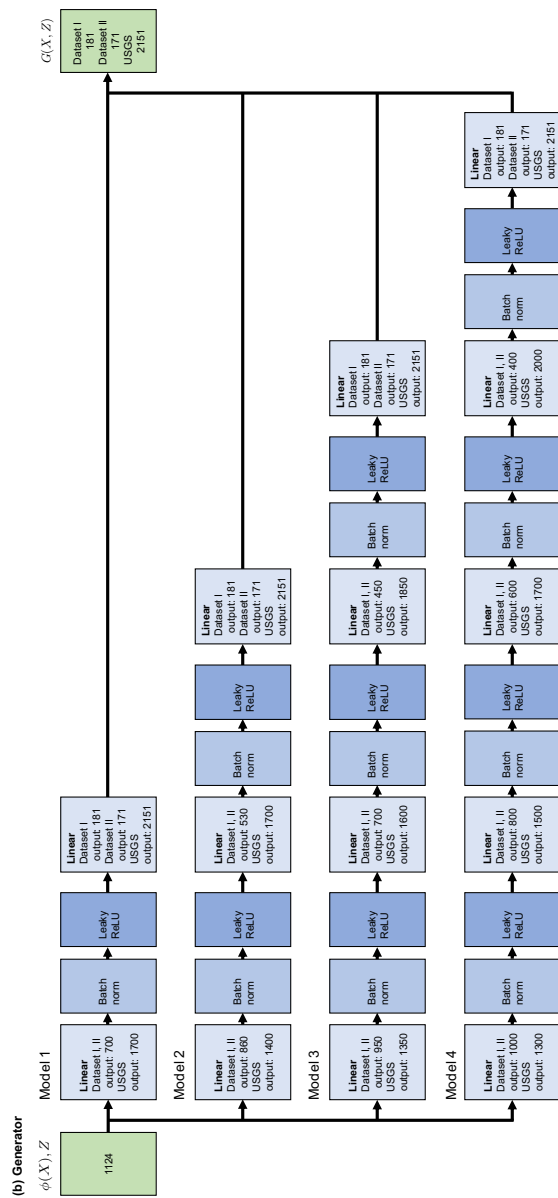


図 4.4 スペクトル関数予測における cGAN 生成器候補モデルのネットワーク設計.

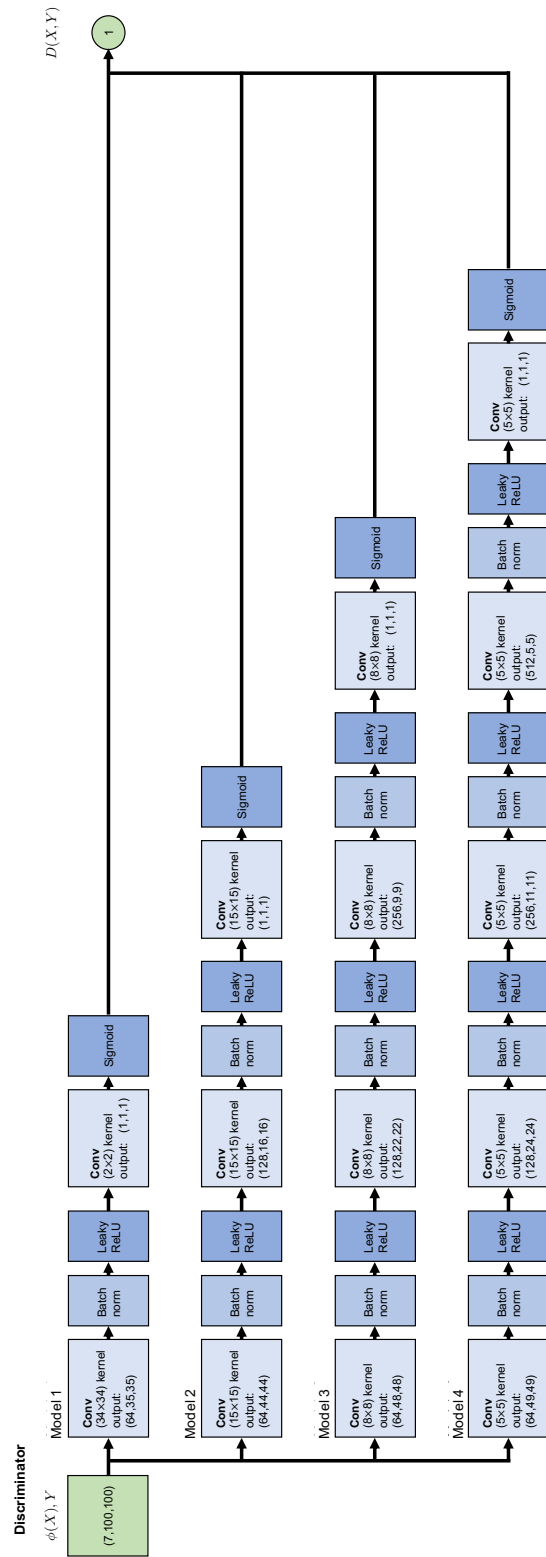


図 4.5 微細構造予測における cGAN 識別器候補モデルのネットワーク設計.

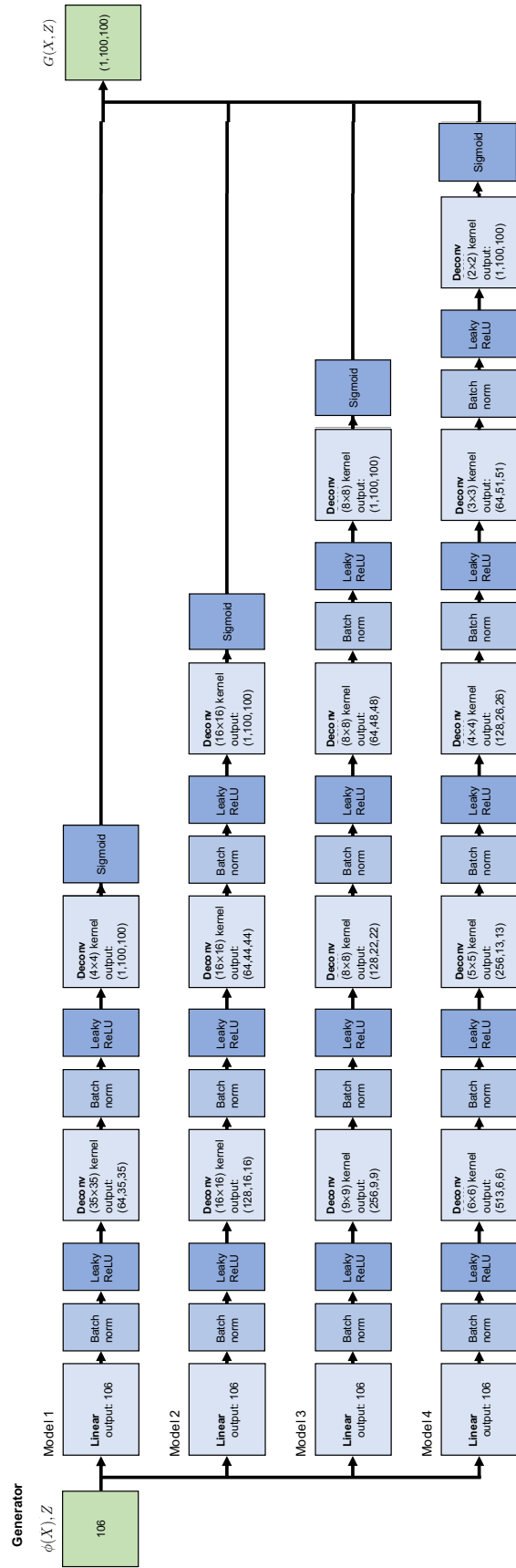


図 4.6 微細構造予測における cGAN 生成器候補モデルのネットワーク設計.

ベクトルの微分に対する $\text{RMSE}_{\text{derivative}}$ を用いて、モデルの予測性能を評価した。

$$\text{RMSE}(i) = \sqrt{\frac{1}{m} \sum_{j=1}^m \left(Y(X_i, t_j) - \hat{Y}(X_i, t_j) \right)^2}, \quad (4.1)$$

$$\text{MAE}(i) = \frac{1}{m} \sum_{j=1}^m \left| Y(X_i, t_j) - \hat{Y}(X_i, t_j) \right|, \quad (4.2)$$

$$R^2(i) = 1 - \frac{\sum_{j=1}^m \left(Y(X_i, t_j) - \hat{Y}(X_i, t_j) \right)^2}{\sum_{j=1}^m \left(Y(X_i, t_j) - \bar{Y} \right)^2} \quad \left(\bar{Y} = \frac{1}{m} \sum_{j=1}^m \hat{Y}(X_i, t_j) \right), \quad (4.3)$$

$$\begin{aligned} \text{RMSE}_{\text{derivative}}(i) &= \sqrt{\frac{1}{m-1} \sum_{j=1}^{m-1} \left[\left(Y(X_i, t_j) - Y(X_i, t_{j+1}) \right) - \left(\hat{Y}(X_i, t_j) - \hat{Y}(X_i, t_{j+1}) \right) \right]^2}. \end{aligned} \quad (4.4)$$

ここで、 $\hat{Y}(X_i, t_j)$ は分子 i と波長 t_j の予測スペクトル値を示す。テストデータセットの n 個の分子の各性能指標の中央値を計算し、三つの独立した実験の統計量を計算した。

微細構造予測では、予測画像と実験の SEM 画像との間の類似性を評価する指標として、Structural SIMilarity (SSIM) [69] と位置・スケール・回転に対する不変性を保持する Oriented FAST and Rotated BRIEF (ORB) [70, 71] を使用した。ORB は、二つの画像の局所的な特徴点を検出した上で、それらの画像特徴量の類似度を計算する。FAST アルゴリズムは、画像内のコーナー特徴点を検出し、それらのバイナリ特徴記述子を計算するための方法である。検出した特徴点に対してマッチングを行い、二つの画像間の近さを計算する。スケール不変性を導入するために、ORB は画像の解像度を変えた画像ピラミッドを作成し、各レイヤーに FAST を適用して全てのピクセルの特徴点を計算する。二つの画像が完全に一致する場合はゼロになる。本研究で使用したもう一つの画像類似度である SSIM は、明るさ、コントラスト、構造の観点から画像を比較し、それらを乗算することにより、類似度を計算する。実験画像 o と予測画像 p の特定のサイズのウィンドウをそれぞれ x と y と表せば、SSIM は次式で定義される。

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \quad (4.5)$$

μ_x と μ_y は、実験画像と予測画像のウィンドウ内の平均強度を示す。 σ_x と σ_y は、分母が非常に小さい場合に、SSIM の出力値を安定させるための定数である。画像の類似性は、ウィンドウが画像座標上を移動するときに、SSIM 値を平均化することによって計算される。SSIM は 0 から 1 の間をとり、二つの画像が完全に一致する場合は 1 になる。

表 4.3 スペクトル関数予測における cGAN 識別器候補モデルの層数とニューロン数の設定. アスタリスクは, モデル検証において汎化性能が最も高かったモデルを表す.

Category	# of layers	Dataset	Input	Layer1	Layer2	Layer3	Layer4	Output
		Dataset I	1205	600				1
1		Dataset II	1195	600				1
		USGS	3175	1600				1
		Dataset I*	1205	800	400			1
2		Dataset II	1195	800	400			1
		USGS*	3175	2200	1100			1
Discriminator								
		Dataset I	1205	900	600	300		1
3		Dataset II*	1195	900	600	300		1
		USGS	3175	2400	1600	800		1
		Dataset I	1205	960	720	480	240	1
4		Dataset II	1195	960	720	480	240	1
		USGS	3175	2600	1950	1300	650	1

表 4.4 スペクトル関数予測における cGAN 生成器候補モデルの層数とニューロン数の設定. アスタリスクは, モデル検証において汎化性能が最も高かったモデルを表す.

Category	# of layers	Dataset	Input	Layer1	Layer2	Layer3	Layer4	Output
1		Dataset I	1124	700				181
		Dataset II	1124	700				171
		USGS	1124	1700				2151
2		Dataset I	1124	860	530			181
		Dataset II	1124	860	530			171
		USGS	1124	1400	1700			2151
Generator		Dataset I	1124	950	700	450		181
		Dataset II	1124	950	700	450		171
		USGS*	1124	1350	1600	1850		2151
		Dataset I*	1124	1000	800	600	400	181
		Dataset II*	1124	1000	800	600	400	171
		USGS	1124	1300	1500	1700	2000	2151

表 4.5 微細構造予測における cGAN 識別器および生成器の候補モデルにおける層数とニューロン数の設定. アスタリスクは, モデル検証において最も高い汎化性能を示したモデルを表す.

Category	# of layers	Input	Layer0	Layer1	Layer2	Layer3	Layer4	Output
Discriminator	1	(7, 100, 100)		(64, 35, 35)				1
	2*	(7, 100, 100)		(64, 44, 44)	(128, 16, 16)			1
	3	(7, 100, 100)		(64, 48, 48)	(128, 122, 22)	(256, 9, 9)		1
	4	(7, 100, 100)		(64, 49, 49)	(128, 24, 24)	(256, 11, 11)	(512, 5, 5)	1
Generator	1*	106	106	(64, 35, 35)				(1, 100, 100)
	2	106	106	(128, 16, 16)	(64, 44, 44)			(1, 100, 100)
	3	106	106	(256, 9, 9)	(128, 22, 22)	(64, 48, 48)		(1, 100, 100)
	4	106	106	(513, 6, 6)	(256, 13, 13)	(128, 26, 26)	(64, 51, 51)	(1, 100, 100)

第 5 章

結果と考察

第 3 章で示した提案手法である KRFO の予測性能を三つのケーススタディを通して検証する。また、KRFO の少数データに対する学習・予測メカニズムを数値的に明らかにし、検証・考察を行う。具体的な問題設定として、有機分子の化学構造を入力から UV-Vis スペクトルおよび近赤外スペクトルの関数を予測するタスクと、マグネトロンスパッタリング法を用いて作成された金属薄膜を対象に、組成とプロセスパラメータから材料の微細構造を表す電子顕微鏡像を予測するタスクを取り上げる。UV-Vis スペクトルの事例では実験データの数は数百～数千であるが、近赤外スペクトルの事例で得られた実験データの数はわずか 68、微細構造予測の実験データ数は 123 とデータ量が非常に少ない。

5.1 スペクトル関数の予測

5.1.1 UV-Vis スペクトルの予測

Urbina et al.[28] は、異なる化学空間を包含する有機分子の UV-Vis スペクトルの実験データセットを 2 種類作成した。本研究では、これらのデータセットをデータセット I およびデータセット II と呼ぶことにする。データセット I は、高速液体クロマトグラフィーを用いて測定された 949 種類の市販化合物の光吸収スペクトルで構成されている。データセット II には、マイクロプレート分光光度計で測定された 2,222 の異なる市販の医薬品分子のスペクトルが含まれている。データセット I およびデータセット II においてそれぞれ、UV-Vis 領域 (220~400nm) で等間隔に配置された 181 点および 171 点の波長での吸収スペクトルが測定されている。測定されたスペクトル関数は最小値が 0、最大値が 1 となるように標準化を行なった。データセット I と比較して、データセット II はデー

タ量が多く、実験対象の化学構造も多様である。モデルの学習と予測性能を評価するために、データセットの約 70% を訓練データとしてランダムに選択し、約 15% をテストデータセット、残りの約 15% を検証用データセットとして使用した。データセットの分割は、同じ化合物の複数のスペクトル関数がテストデータセットに含まれないように実行した。Urbina et al.[28] は、このデータセットに対して LSTM と注意機構を用いた Seq2Seq を適用した結果を報告している [28]、本研究で新たに実施した cGAN と KRFO の予測結果を論文の報告値と比較する。

KRFO では、観測された波長範囲 (220 または 230~400nm) を等間隔に 128 個のセグメントに分割し、カーネル中心 s_1, \dots, s_d ($d = 128$) を配置した。分割数は任意に決定することができる。カーネル基底関数の種類も任意に設定することができるが、カーネル基底関数の種類は予測精度に特に影響を及ぼさなかったため、ここでは RBF カーネルの結果のみを報告する。ハイパーパラメータは、カーネル基底関数の分散 σ^2 と高さスケール l 、ニューラルネットワークの隠れ層の数で構成される。前章で述べたように、ハイパーパラメータの探索候補として σ^2 では三つの異なる値、 l では四つの異なる値を設定した。ニューラルネットワークの隠れ層の候補数は 1~4 とした。各層のニューロン数は出力の数 (128) と入力変数の次元数の間で線形に増加あるいは減少するように決定した。最終的なハイパーパラメータの候補は、 σ^2 の 3 個、 l の 4 個、隠れ層の 4 個の組み合わせになり、合計で 48 パターンの組み合わせを探索した。各組み合わせのモデルの性能を検証データセットの平均平方二乗誤差 (RMSE) で測定して最適な組み合わせを決定した。

cGAN のハイパーパラメータは、生成器と識別器のネットワーク構造で構成される。KRFO と同様に、生成器と識別器のネットワーク構造は、全結合層、バッチ正規化、LeakyReLU 活性化関数から成るブロックの繰り返しにより構成される。隠れ層のブロック数の候補は 1~4 とし、各層構造のニューロン数はカーネル回帰と同様に入出力間で線形に次元数が増加・減少するように定めた。モデルの予測性能の評価では検証データセットに対する RMSE を計算し、最適なハイパーパラメータの組み合わせを決定した。

ハイパーパラメータ探索の結果、データセット I では、KRFO, cGAN の生成器・識別器のニューラルネットワークの隠れ層の数はそれぞれ 4, 4 および 2 が選択された。データセット II では、KRFO, cGAN の生成器・識別器のニューラルネットワークの隠れ層の数は、それぞれ 3, 4 および 3 となった。分散 σ^2 の値は両方のデータセットで 0.0005、 l はデータセット I とデータセット II のそれぞれ 0.5 および 5 となった。いずれのハイパーパラメータも予測精度に影響を与えうるが、特に等間隔に配置するカーネル基底関数の幅を決める分散 σ^2 の値の影響は大きい。分散 σ^2 の値は、配置するカーネル基底関数の数に応じて、各カーネル基底関数が十分に重なるように調節しなければならない。重

なりが不十分であれば、モデルの滑らかさが足りず予測精度の低下が起きる。そのため、KRFO を用いてスペクトル関数を予測する際には、カーネル基底関数の種類や高さ、隠れ層の数よりも、カーネル基底関数の幅に注意してハイパーパラメータ調整を行う必要があることが分かった。

調整されたハイパーパラメータを用いて訓練されたモデルの予測性能を評価するため、各テスト分子の m 個の予測スペクトル値と観測値の間の RMSE, 決定係数 (R^2) および MAE を計算し、150 あるいは 342 個のテスト分子の各性能指標の中央値の平均と標準偏差を比較した。また、関数データの予測では勾配の予測も重要であることから、スペクトル関数の勾配の RMSE (RMSE derivative) に基づいて予測性能を評価した。訓練データ, 検証データ, テストデータへの分割を独立に 3 回実施し、性能指標を計算した。

表 5.1 に、三つの独立した数値実験の性能指標の平均と標準誤差と共に、Urbina et al.[28] が報告した LSTM と注意機構を用いた Seq2Seq の予測精度を示す。データ数はデータセット II が最も大きく、学習に用いたデータ数は 1,552 個である。データセット I では、649 個のデータが訓練に用いられた。データセット II では、Urbina et al.[28] の Seq2Seq の性能が最も優れていたものの、KRFO と cGAN もほぼ同等の性能を示している。データセット I では、KRFO が他の三つの手法よりも優れていたが、cGAN との違いはそれほど顕著ではない。また、cGAN の RMSE derivative は他の手法と比較して精度が低い。これは、cGAN の予測スペクトルの計算ではランダムにサンプリングされた Z のアンサンブル平均を使用しているため、スペクトルの勾配の滑らかさが失われることが原因である。ただし、この問題は事後処理として予測スペクトルの平滑化を実施することである程度解決することができる。なお LSTM や Seq2Seq のコードが配布されていないため、モデルの性能比較をさらに深めることができなかったが、少なくともこのケーススタディではほぼ同じ性能を有していると結論付ける。

図 5.1 と図 5.2 は、データセット I とデータセット II からランダムに選択した 27 個のテスト分子の予測結果を示す。なおここでは 27 個のみの結果を示したが、予測精度について包括的な理解を得るために、付録 A にランダムに選択した 60 個のテスト分子に対する予測結果や推定された切片項 $\mu(t)$, 訓練時のフィッティング結果も示している。

図 5.1 と図 5.2 から、KRFO はスペクトルの細かい凹凸のパターン含めて予測できていることがわかる。また多くの分子について、吸収スペクトルの複数のピーク位置と関数の形状がほぼ完全に予測されている。ケースによっては、小さなピークや緩やかな丘状のピークといった目視では分かりづらい特徴も適切に捉えることができている。この結果は、化学的には部分構造の有無が分子の光吸収スペクトルの主決定要因であることを示唆する。スペクトル関数の大まかな傾向を捉えるという点では、KRFO と cGAN の間に大

表 5.1 UV-Vis-NIR 吸収スペクトルの予測精度の比較. データセット I とデータセット II は UV-Vis スペクトル, USGS は近赤外スペクトルの予測. LSTM と注意機構を用いた Seq2Seq の結果は [28] の報告値である.

	RMSE	R ²	MAE	RMSE derivative
Dataset I	LSTM	0.169 ± 0.132	0.626 ± 1.166	0.119 ± 0.106
	Attention	0.154 ± 0.144	0.680 ± 1.230	0.091 ± 0.120
	Kernel	0.111 ± 0.009	0.798 ± 0.043	0.075 ± 0.006
	cGAN	0.112 ± 0.014	0.786 ± 0.057	0.071 ± 0.010
Dataset II	LSTM	0.064 ± 0.062	0.710 ± 0.472	0.047 ± 0.075
	Attention	0.055 ± 0.071	0.699 ± 0.259	0.044 ± 0.052
	Kernel	0.093 ± 0.004	0.655 ± 0.007	0.066 ± 0.001
	cGAN	0.085 ± 0.002	0.718 ± 0.022	0.058 ± 0.003
USGS	Kernel	0.076 ± 0.024	0.602 ± 0.200	0.053 ± 0.009
	cGAN	0.074 ± 0.003	0.493 ± 0.102	0.528 ± 0.043

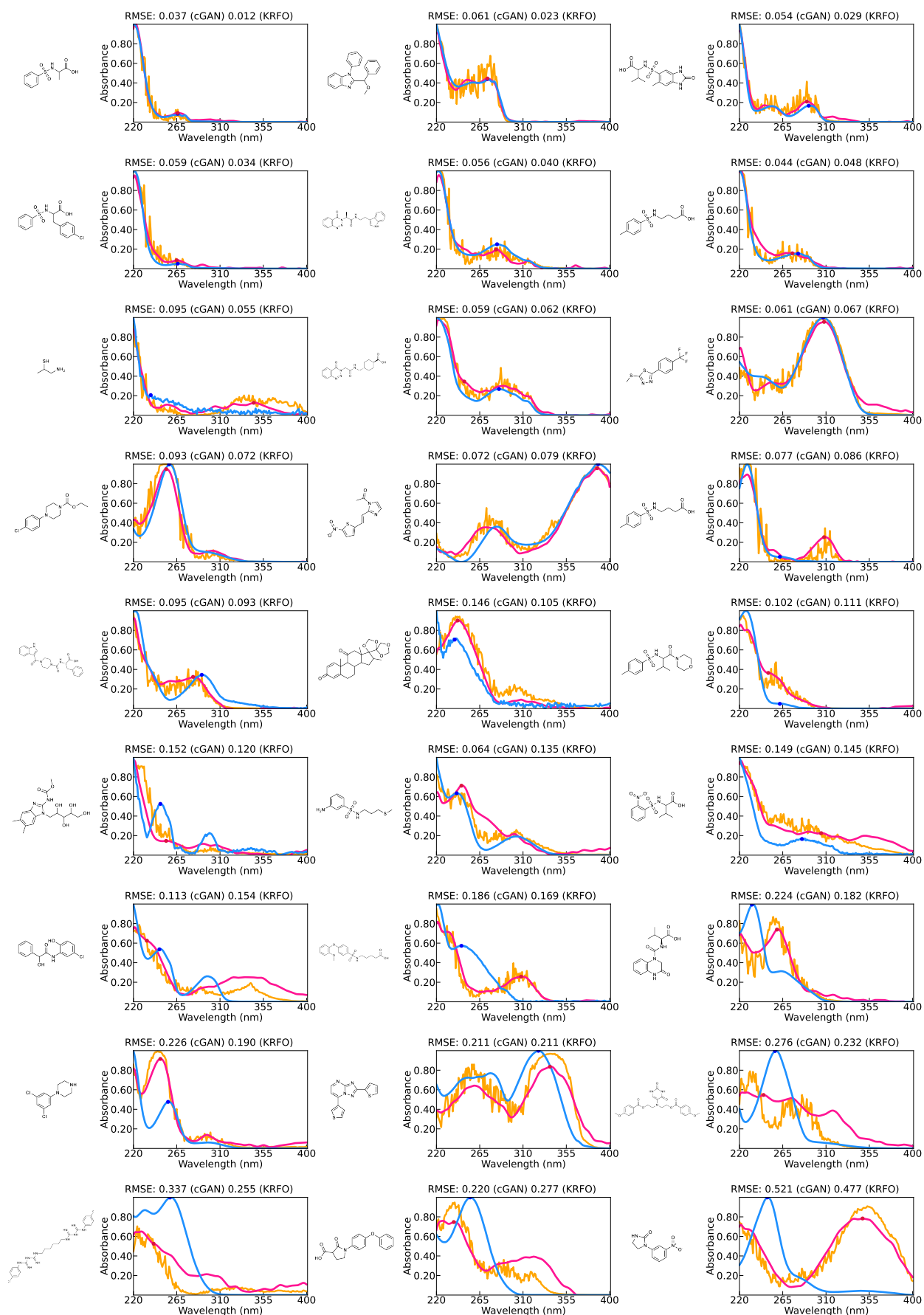


図 5.1 データセット I からランダムに選択された 27 個のテスト分子の予測結果. (a) KRFO (ピンク), (b) cGAN (オレンジ), (c) UV-Vis スペクトルの実験値 (青).

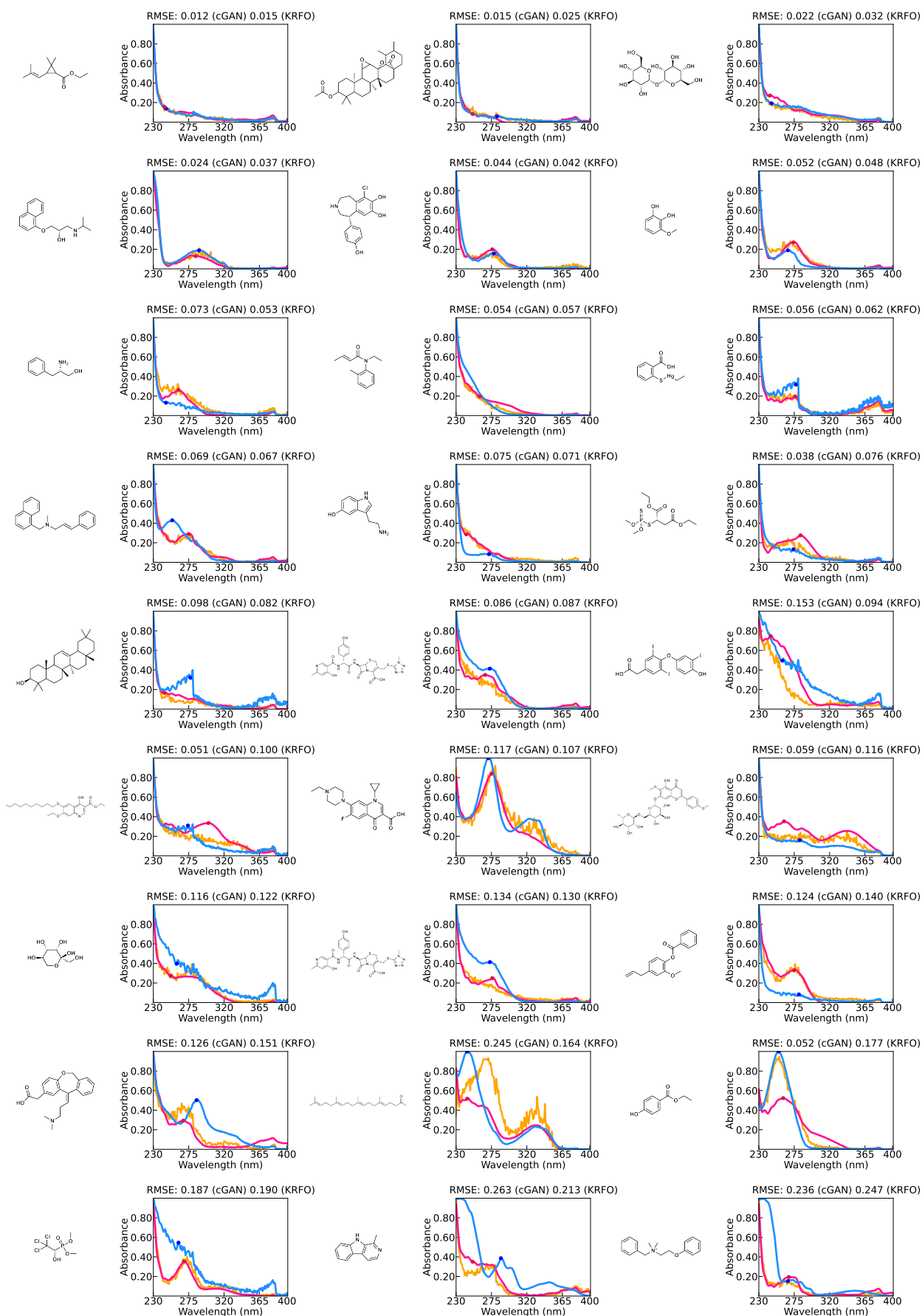


図 5.2 データセット II からランダムに選択された 27 個のテスト分子の予測結果。
(a) KRFO (ピンク), (b) cGAN (オレンジ), (c) UV-Vis スペクトルの実験値 (青).

きな違いはない。ただし、cGAN は予測式の Z のランダムネスにより、予測されたスペクトル関数に揺らぎが生じるため、マイナーなピーク位置を捉えることは難しい。平滑化処理を施したとしても、偽のピークが出現してしまう可能性がある。したがって、スペクトル予測に関しては KRFO が大きな利点を有していると結論付ける。

5.1.2 近赤外スペクトルの予測

米国地質調査所 (USGS) スペクトルライブラリのデータセットには、1,000 種類以上の分子の反射光のスペクトルが収録されている。データセットには実験室や野外で空中分光器で測定されたサンプルが含まれている [72]。また無機物から有機物までさまざまな物質のデータが収録されている。本研究では有機物に解析対象を限定した。また、同一分子に複数のスペクトルデータが記録されている場合には室温で測定したスペクトルデータを選択した。全分子に対し、紫外線から遠赤外線までの波長領域 $0.35\sim 2.5\mu\text{m}$ において、ほぼ等間隔で配置された 2,151 点の波長でスペクトルの値が記録されている。UV-Vis のデータ数と比べると、取得できたデータの数 68 個とサンプル数はかなり少ないデータセットとなる。なおデータとして、反射率 R が記録されていたため、反射率と吸収率の関係式から $A = \log_{10} 100/R$ という式に従い吸光度 A に変換し、最小値が 0、最大値が 1 となるように標準化を実施した上でモデルの学習を行なった。cGAN と KRFO のモデル構成とハイパーパラメータの選択方法は、UV-Vis スペクトル予測のケースと全く同じである。KRFO と cGAN の生成器・識別器のニューラルネットワークの隠れ層の数はそれぞれ、3, 3, 2 が最適値として選択された。また、カーネル基底関数の σ^2 と l の値は、それぞれ 0.0005 と 1 となった。訓練と検証、テストには、ランダムに選択した 80%, 10%, 10% のデータ集合を使用した。また、データ分割を独立に 3 回実施し、性能指標の平均とばらつきを算出した。

選択されたハイパーパラメータを用いて学習した cGAN と KRFO の予測精度を表 5.1 に示す。UV-Vis スペクトルの予測結果とは異なり、KRFO は cGAN の精度を大幅に上回ることが分かった。例えば、cGAN と KRFO の R^2 の値は、それぞれ 0.493 ± 0.102 , 0.602 ± 0.200 , MAE は 0.058 ± 0.009 , 0.053 ± 0.009 となった。UV-Vis の二つの予測タスクに比べてデータ数がかなり少ないが、KRFO は高い予測精度を保持している。データが少ない場合、cGAN の生成器の過剰なパラメタリゼーションは予測性能の低下を引き起こす原因となる。一方、KRFO では、わずか 54 サンプルで学習されたにもかかわらず、十分に高い予測精度を達成できた。図 5.3 は、ランダムに選択された 9 個のテスト分子の予測スペクトルと実験スペクトルを示す。UV-Vis の場合と同様に、観測されたスペ

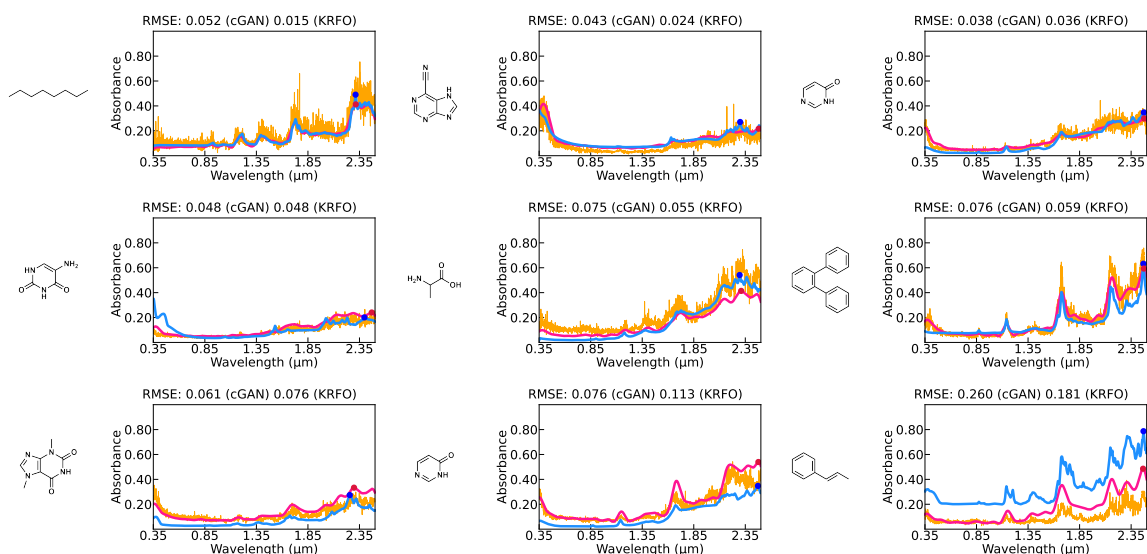


図 5.3 USGS データセットの 9 個のテスト分子の予測結果. (a) KRFO (ピンク), (b) cGAN (オレンジ), (c) UV-Vis スペクトルの実験値 (青).

クトルのピーク位置と関数の形状の大まかな特徴を予測することに成功していることが分かる. また, 多くの分子において, マイナーなピークを含むスペクトル関数の細かい凹凸パターンも適切に捉えられていることも確認できる. KRFO には少数データに対する高い学習性能が備わっていることが示唆される. 付録 A では, 図 5.3 で示したテスト分子以外の予測結果と切片項 $\mu(t)$ の推定結果, 訓練分子に対するフィッティングの結果を網羅的に示している.

5.2 電子顕微鏡像の予測

ここでは, マグネトロンスパッタリング法を用いてアルミニウム (Al) をコーティングしたクロム (Cr) を基盤とする金属板から構成される薄膜材料を解析対象とする. Al と Cr の金属板を並べた状態で, マグネトロンスパッタリング装置で窒素 (N) とアルゴン (Ar) の混合ガスを Al 板に吹き付け, 放出された Al 原子を Cr 板に吸着させて材料を作製する. 高温条件下では, 装置の残留ガスの放出により, 金属板は酸素 (O) によって汚染される. モデルの入力変数 $X \in \mathbb{R}^6$ は, 組成 $\text{Cr}_{1-x}\text{Al}_x\text{O}_y\text{N}$ と加工条件を示す 6 次元の実数ベクトルである. 入力変数ベクトルの要素は以下の通りである.

1. Cr と Al の組成
2. O の組成
3. Al を吹き付ける際の温度

4. Al を吹き付ける際の圧力
5. Ar イオンを金属板に発射する際のエネルギー
6. Ar ガスのイオン化エネルギー

入力ベクトル X の各要素は、訓練データのサンプル平均が 0、分散が 1 になるように正規化した。出力変数 Y は、微細構造を観測した解像度 100×100 の SEM 画像である。

KRFO では、 32×32 の RBF カーネル ($d = 32 \times 32$) の中心を 2 次元画像座標空間上に等間隔に配置した。ハイパーパラメータ探索の条件は、スペクトル関数の予測とほぼ同様である。画像予測でも、カーネル関数の種類は予測精度に大きな影響を及ぼさなかったため、ここでは RBF カーネルの結果のみを報告する。ハイパーパラメータの探索では、ネットワークの構造とカーネル基底関数の分散 σ^2 と高さスケール l を探索対象とした。ハイパーパラメータ探索の結果、KRFO と cGAN の生成器・識別器のニューラルネットワークの隠れ層の数は、それぞれ 4, 1, 2 が選択された。 σ^2 と l の値は、0.0001 と 5 が選択された。画像予測のケースでも、スペクトル関数と同様に、予測精度への影響度が大きいハイパーパラメータはカーネル関数の幅 σ^2 の値であった。KRFO ではデータの特性に依じた滑らかさを有する関数クラスを表現するために、分散パラメータを調整して、カーネルが適切に重視するように配置させることが重要である。データセットには、組成および処理条件が紐づけられた 123 個の SEM 画像が含まれている。そのうちの 90%, 5%, 5% をそれぞれ、訓練、検証、テストデータセットとしてランダムに選択した。データ分割を独立に 3 回繰り返し、性能指標の平均とばらつきを評価した。モデル設計や性能評価指標の詳細については、第 4 章を参照せよ。

図 5.4 に、ランダムに選択した 7 個のテストケースに対する KRFO と cGAN の予測結果と実験の SEM 画像を示す。また予測性能を直観的に理解するために、付録 A において、ランダムに選択された 14 個のテストデータに対する予測結果と切片項 $\mu(t)$ の推定結果、ランダムに選択された 60 個の訓練データに対するフィッティングの結果をまとめた。KRFO では、訓練データ数が 109 個と非常に少ないにもかかわらず、粒子サイズや形状といった微細構造の形態的特徴の違いが適切に捉えられていることがわかる。一方、cGAN の予測は実験データの微細構造のパターンを全く捉えることができていない。このように、訓練データ数が限られている状況では、cGAN による画像生成は性能が大きく低下する。図 5.5 には、予測・実験画像の類似度として SSIM と ORB を用いた KRFO と cGAN の比較結果をまとめている。図 5.4 の解析例よりほぼ明らかではあるが、定量的にも KRFO が cGAN よりも圧倒的に優れた予測性能を有することが確認された。

スペクトル予測と同様に、微細構造の電子顕微鏡像の予測においても、少量データに対する KRFO の高い学習性能が確認された。ただし、予測画像は微細構造の形態的特徴を

上手く捉えているが、実験画像のような鮮明な画像は得られていない。鮮明度の低下は、カーネル関数を用いることによるぼかし効果が原因である。KRFO は、カーネル基底展開を用いたパラメータの節約により少数データに対する高い耐性を実現しているが、その代償として画像のシャープさに低下が生じる。ただし、材料研究における画像予測タスクでは、一般画像の生成に要求されるような高品質の画像を生成する必要はないことが多い。組成やプロセス条件ごとの微細構造の大まかな差異を識別できれば、材料設計の目的は十分に達成される。また画像の鮮明さが要求される場合には、様々な工夫を施せばよい。例えば、損失関数の設定において、画像変換用の生成モデルとして有名な pix2pix[73] のように、 l_1 損失を適用することで画像のブラー効果を抑制できると考えられる。

5.3 少数データに対する KRFO の予測メカニズム

スペクトル予測のデータ数は約 50、微細構造予測のデータ数は約 100 というように非常に少ない訓練データ数にも関わらず、KRFO が高い予測能力を獲得することに成功した要因を分析する。KRFO には、マルチタスク学習と同じような学習メカニズムが備わっていると考えられる。関数の各値を予測するタスクは、当然ながら独立ではなく互いに関連し合っている。マルチタスク学習では、複数の関連するタスクを同時に学習することで、モデルがタスク間に共通する特徴とタスク固有成分を捉えることを期待する、また、複数のタスクのデータを同時に利用することで、個々のタスクのデータ量の不足を補い、タスク固有の観測ノイズにモデルが過適合することを抑制する。関数出力回帰は、異なるタスク（すなわち異なる関数の値）のデータが同時に観測されるという点で、マルチタスク学習の特殊ケースであると解釈できる。このようなメカニズムが働くことで、関数出力回帰は少数のデータしかないにも関わらず、高い予測性能を引き出すことに成功した可能性がある。ここで、この仮説を検証するために、データセット I の光吸収スペクトルの予測タスクについて、以下の三つの数値実験を実施した。

- (a) KRFO によって予測されたスペクトル関数から、最大吸収波長 λ_{max} を求め、その位置と強度の値を算出し、実験値と比較した。
- (b) λ_{max} の位置と強度の実験値を用いて、通常のスカラー出力変数の回帰を行った。入力変数は化学構造のフィンガープリント記述子とし、MLP でモデルを構築した。
- (c) 等間隔に配置された m 個の波長点におけるスペクトルの値をそれぞれ独立に予測する m 個の MLP を学習し、予測されたスペクトル関数から最大吸収波長 λ_{max} の位置と強度を算出した。

表 5.2 少数データセットに対する最大吸収波長 λ_{max} の位置と強度の予測結果. 報告されている数値は, 本文で示した三つのケース (a) - (c) の λ_{max} の位置・強度の正解率を表す. 訓練, 検証, テストデータの比率を 649:151:150, 100:425:425 と設定した場合の結果を確認した.

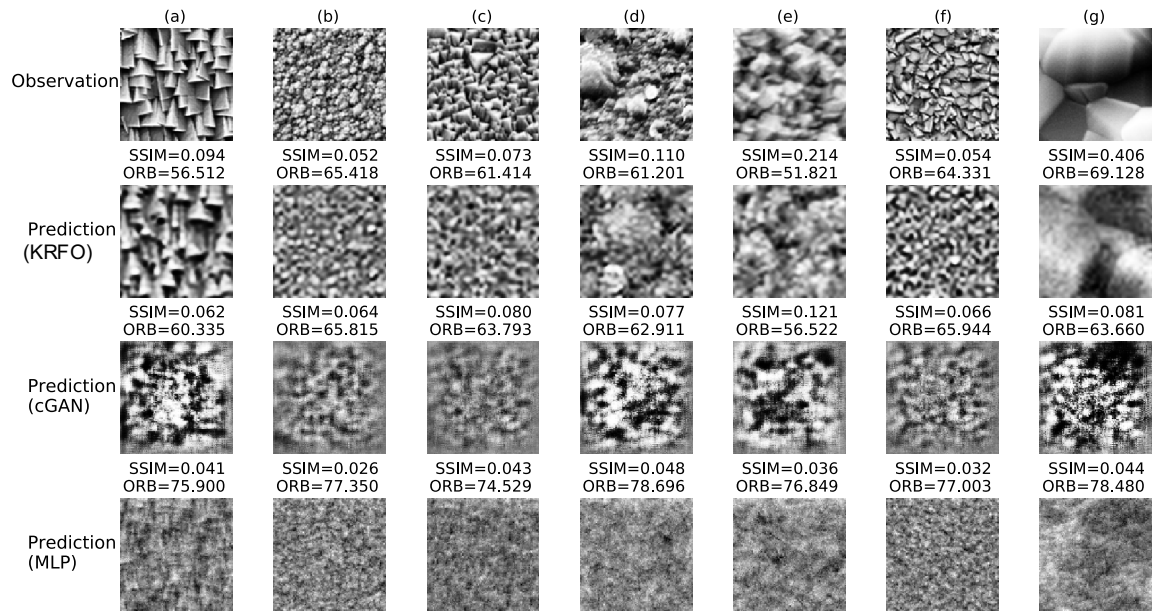
Method	Data Split	Position of λ_{max} (%)	Intensity of λ_{max} (%)
(a) Kernel	649:151:150	46.3	43.9
(b) MLP for λ_{max}		23.5	11.7
(c) MLP for each wavelength		45.9	45.5
(a) Kernel	100:425:425	36.1	32.6
(b) MLP for λ_{max}		22.4	11.1
(c) MLP for each wavelength		32.9	25.5

スペクトル関数からのピーク検出では, 左端から 232nm に現れる極大値のうち, 最大強度の波長を λ_{max} として選択した. これら三つの実験では, 訓練, 検証, テストのデータはそれぞれ 649:151:150 および 100:425:425 と設定した. データの分割を独立に 10 回繰り返し, モデルの訓練とテストを実施した. テストセットに対する予測精度の評価では, 予測されたピーク位置が実験で観測されたピーク位置の 10nm 以内にある場合を正解とし, さらに予測されたピーク強度が観測されたピーク強度の 0.1 以内にある場合を正解とした. 表 5.2 に訓練データのデータ量を 649 と 100 とした場合の実験結果を示す. 表 5.2 より, 訓練データ数が異なる両方のケースにおいて, 関数全体を同時に学習したケース (a) の KRFO は, ケース (b) とケース (c) の予測性能を大幅に上回ることが確認された. 特に, スペクトルの値を個別に学習・予測したケース (c) と比較して, ケース (a) は訓練データ数が 100 の場合でも予測性能がそれほど低下しなかった.

さらにケース (c) と同様のテストを微細構造の画像予測に対しても実施した. 画像のピクセル毎に独立に画像強度を学習した MLP の予測結果を図 5.4 に示す. 図 5.4 から, ピクセル毎に独立に学習された MLP は明らかに微細構造の予測に失敗していることが確認できる.

これらの観測事実は, 我々に MI における機械学習のアプローチの仕方を再考する機会を与えてくれる. 例えば, 温度依存性を示す物性の予測では, 多くの場合, 測定温度を室温に限定してスカラー出力変数をモデリングの対象とすることが多いが, このようなアプローチは問題をより難しくしており, 逆に予測精度の低下を引き起こしている可能性がある

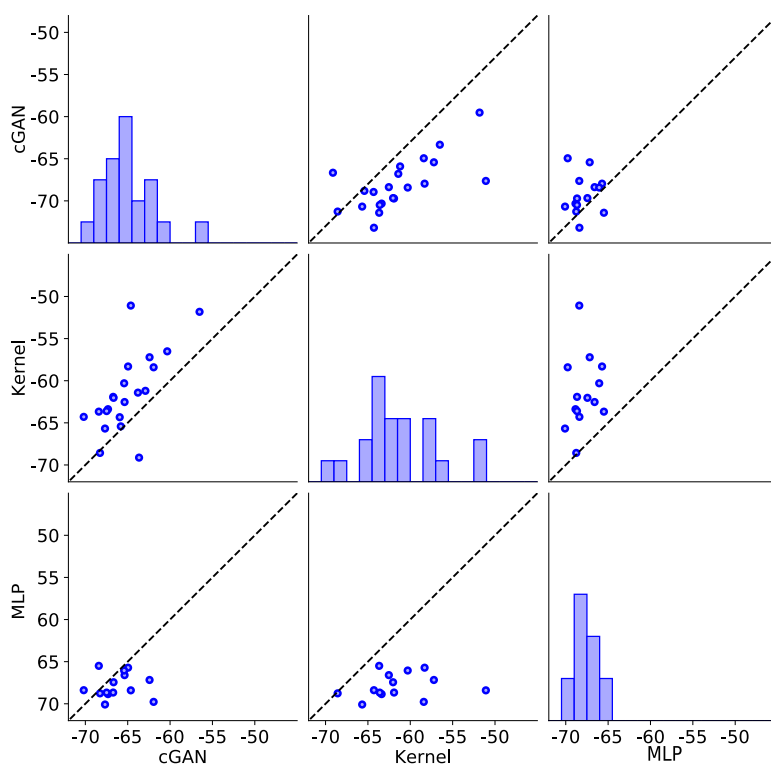
る。実際に、ピクセル毎の学習・予測では画像を全く予測できなかったのに対し、KRFOで画像全体を同時に学習することで微細構造の予測に成功した。関数データを利用できる場合は、関数全体を直接予測することで予測精度を大幅に改善できる可能性がある。



Parameters	(a)	(b)	(c)	(d)	(e)	(f)	(g)
T_d [°C]	20	200	300	500	500	600	800
P_d [P_a]	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Al [at. %]	0	27	0	57	0	33	0
O [at. %]	0	10	3	0	0	10	3
E_l [eV]	1.43	7.23	6.40	41.4	202	6.44	2.80
I_d [a. u.]	0.14	0.50	0.22	1.0	0.20	0.50	0.12

図 5.4 KRFO と cGAN, ピクセルごとに独立に学習された多層パーセプトロン (MLP) による微細構造の予測結果. (i) 実験の SEM 画像, (ii) KRFO の予測画像, (iii) cGAN の予測画像, (iv) MLP の予測画像.

(a) Negative ORB



(b) SSIM

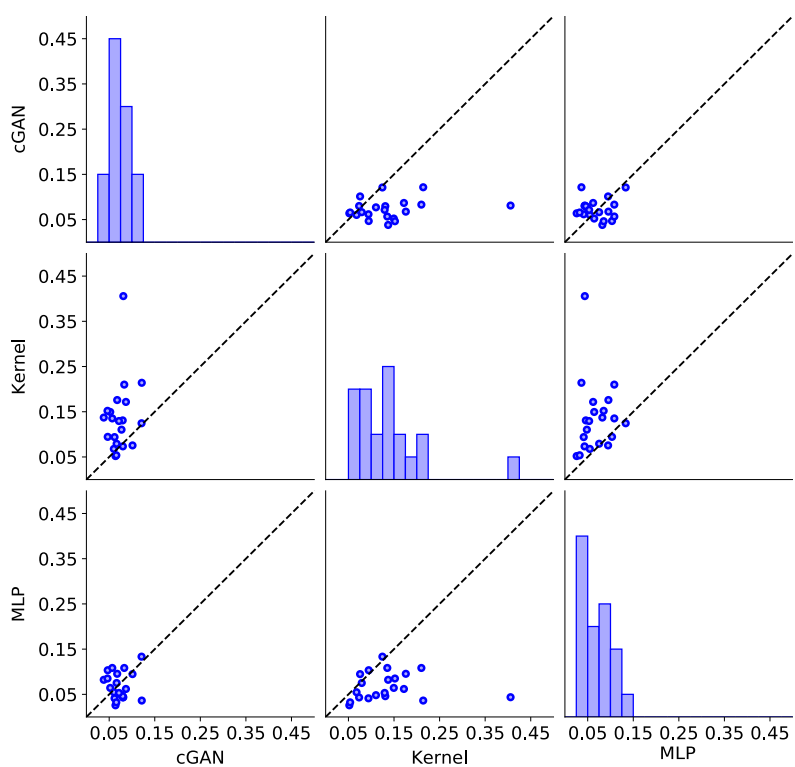


図 5.5 cGAN, KRFO, ピクセルごとに独立に学習された多層パーセプトロン (MLP) の予測画像と実験画像の画像類似度 (SSIM と負の ORB) の散布図.

第6章

まとめ

ここ数年で機械学習の様々な技術が材料科学に導入されてきた。それに伴い、材料種やプロセス条件等の入力変数からスカラー量定義された材料特性を予測するための方法論や解析技術の研究が急速に進展し、実践展開が進められてきた。その一方で、超多次元出力変数や関数型出力の予測に着目した研究はまだほとんど進んでいない。本研究では、材料研究に存在する関数出力変数の回帰問題に着目し、関数データ解析の分野で研究されてきた手法に深層学習による特徴抽出のアイデアを融合する形で KRFO というモデルと統計的方法論を構築した。このような観点に基づいた MI の研究は、これから発展していく段階にあり、今後、本研究の提案手法以外にもより優れた方法論の開発が進んでいくに違いない。本研究の最も重要な学術的意義は、材料科学における関数出力回帰の未開拓問題の発掘、新しい視点に基づく統計的方法論の構築とその概念実証を行った点であることを強調したい。

本研究では、材料科学での関数出力回帰の研究の出発点として、化学構造から光吸収スペクトルの関数を予測するタスクと、組成とプロセス条件から材料の微細構造の電子顕微鏡画像を予測するタスクをテストケースとした [38]。これらのケーススタディから浮かび上がった最も興味深い点は、関数出力変数回帰の少数データに対する高い学習性能である。本研究で学習に用いたデータの量は、一般的な教師あり学習のデータよりもはるかに少ない。光吸収スペクトルの予測では、最小で 54 個のサンプル、微細構造画像の予測では、109 個のサンプルをモデルの訓練に使用した。データ量が非常に少ないにも関わらず、KRFO を用いることで十分に高い予測精度を持つモデルを得ることに成功した。一般に、物性予測の MI では、関数データを直接予測せず、関数データを何らかの方法で特徴量に変換した後、スカラー出力変数を予測することが多い。しかしながら、本研究で示した実験より、スカラー出力変数の予測よりも、関数出力変数をまるごと予測するほうが

高い汎化性能を獲得できるケースが観測された。直感的には、出力変数の次元を高くするとモデルの過学習による汎化性能の低下が生じることも考えられる。しかしながら、複数の関数の値を同時に予測するタスクは相互に強く関連しているため、複数タスクのデータを同時に学習に使用することで、データ拡張によるタスク固有のノイズの悪影響を抑制できる。さらに、関数学習では、通常は1次元あるいは2次元の入力変数 t のみがモデルに追加されるため、モデルの複雑さが大幅に増大することはない。モデルの複雑化とデータ拡張のトレードオフでは、後者の利点が前者の利点を上回る傾向がある。KRFO とマルチタスク学習との間にある理論的な関係性や学習メカニズムを明確化することは今後の課題である。また、材料研究では、関数出力変数の予測が重要な役割を担う問題設定が他にも数多く存在している。ほとんどの材料特性は、温度や周波数の関数として与えられる。誘電特性は、温度と周波数の関数として定義される。比体積、線膨張係数、体積弾性率、比熱、熱伝導率、 χ パラメータなどのポリマー特性も、温度に依存する物性である。さらに、これらの温度依存曲線からガラス転移温度や結晶融点、結晶化温度などの重要な特性が計算される。材料特性や構造の製造・加工条件に対する依存性も多次元出力変数の回帰問題の典型例である。さらに、材料構造を表す2次元および3次元の画像は画像座標空間の多次元関数として与えられる。現時点では利用できるデータが不足しているため、スペクトル関数と微細構造の電子顕微鏡像という限られたケースのみを研究対象としたが、本研究の提案手法は、このような材料研究の多くの未踏課題に適用できる可能性がある。

付録 A

実験結果の補足

A.1 スペクトル予測の実験結果補足

データセット I とデータセット II それぞれのデータセットに対しランダムに分割し、約 70%, 15%, 15% をそれぞれ訓練データ、検証データ、およびテストデータに割り当てた。データの分割は独立して 3 回繰り返した。全てのテストデータの予測結果は示さないが、図 5.1 と図 5.2 に示したテスト分子以外の 60 個のテスト分子の予測結果を図 A.1.1, 図 A.1.3 に示す。また、図 A.1.2 と図 A.1.4 は、訓練データセットに対する当てはまり結果を示す。

少ないデータ量に対する KRFO のメカニズムの検証実験では、データセット I を対象に最大吸収波長 λ_{max} とその強度に対し以下の三つのケースを実験した。

- (a) KRFO によって予測されたスペクトル関数から、最大吸収波長 λ_{max} を求め、その位置と強度の値を算出し、実験値と比較した。
- (b) λ_{max} の位置と強度の実験値を用いて、通常のスカラー出力変数の回帰を行った。入力変数は化学構造のフィンガープリント記述子とし、MLP でモデルを構築した。
- (c) 等間隔に配置された m 個の波長点におけるスペクトルの値をそれぞれ独立に予測する m 個の MLP を学習し、予測されたスペクトル関数から最大吸収波長 λ_{max} の位置と強度を算出した。

データ量による違いを比較するため、訓練データ、検証データ、およびテストデータ数は、それぞれ 649:151:150 (第 5.1.1 項と同条件)、100:425:425 に設定した。データの分割は、独立して 10 回繰り返した。ケース (c) の MLP については、四つの候補の中で検証セットで最も低い RMSE を示すネットワーク構造を選択し、ケース (b) の場合四つ

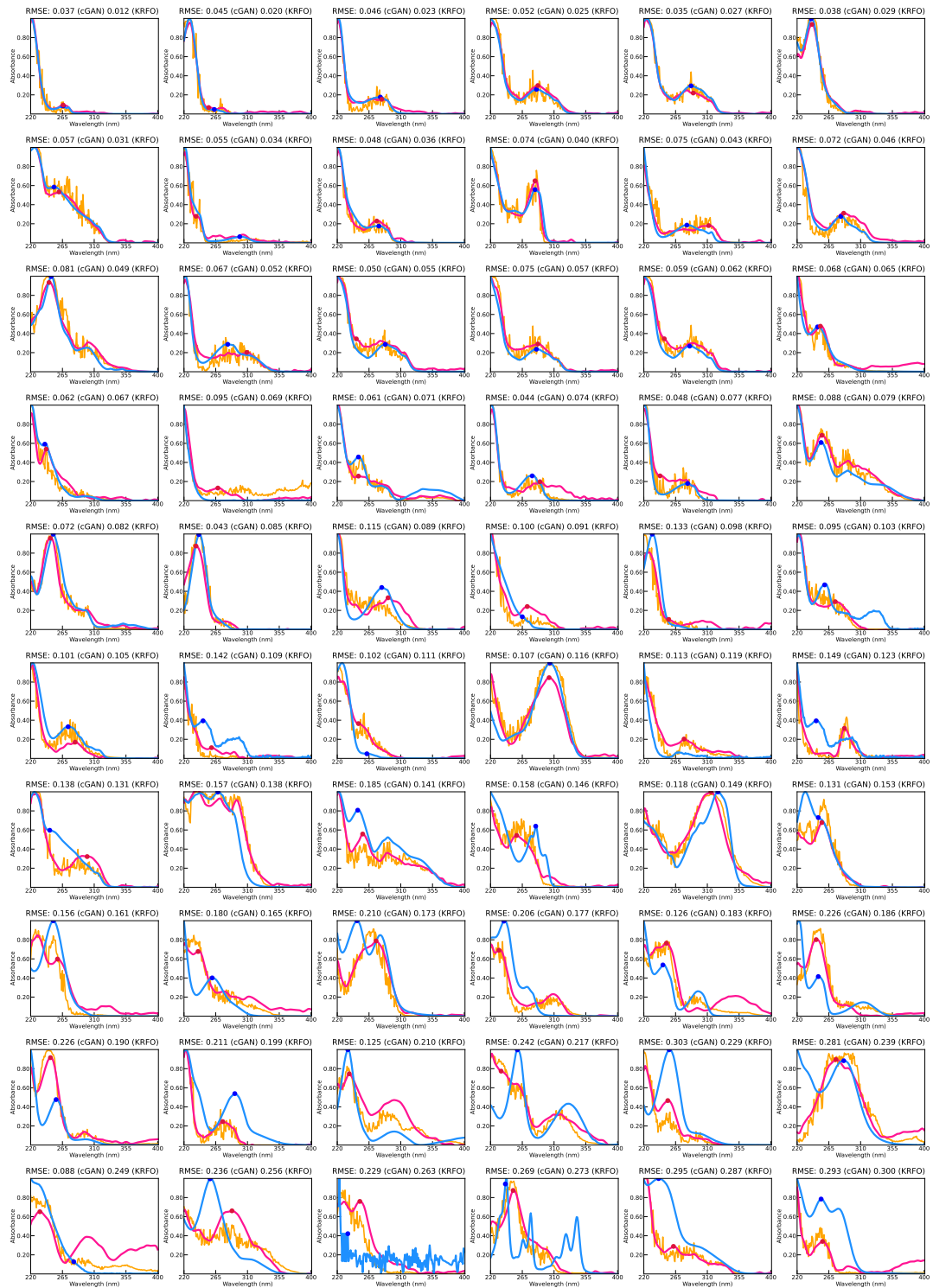


図 A.1.1 DatasetI: DatasetI でランダムに選択された 60 個のテストデータの実験 UV-Vis スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) の予測結果.

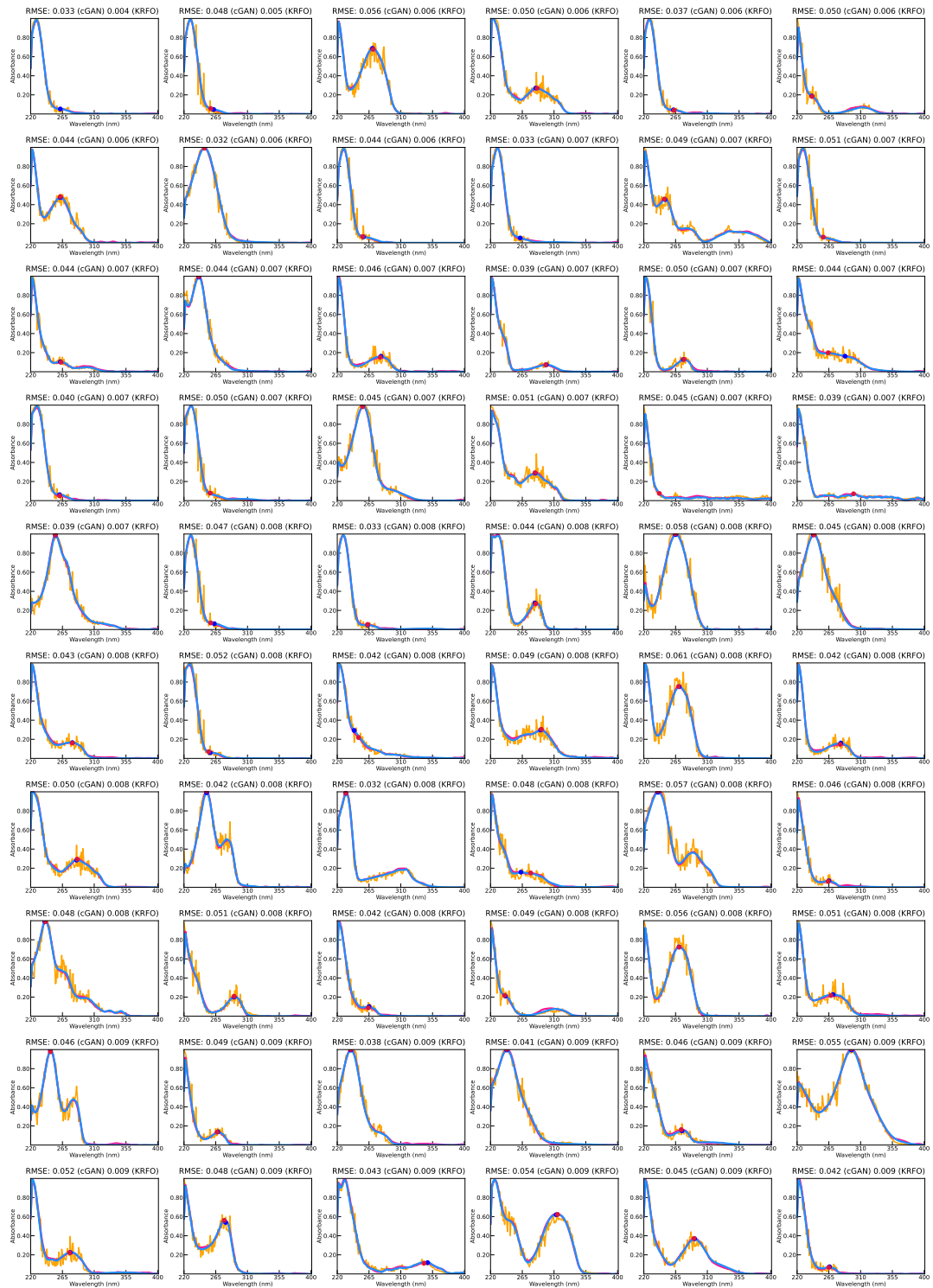


図 A.1.2 Dataset I: Dataset I でランダムに選択された訓練データの実験 UV-Vis スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) のフィッティング結果.

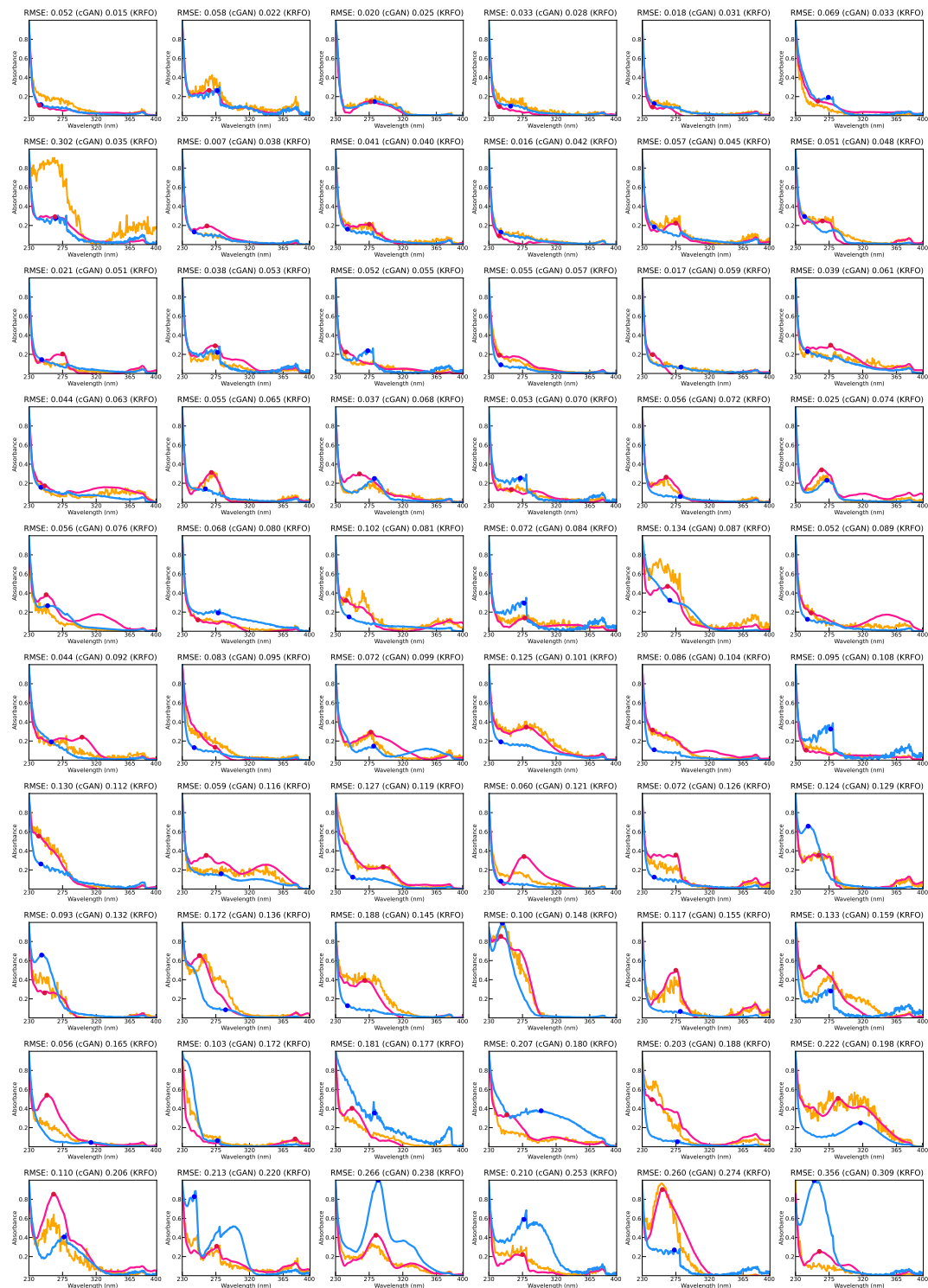


図 A.1.3 DatasetII: DatasetII でランダムに選択された 60 個のテストデータの実験 UV-Vis スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) の予測結果.

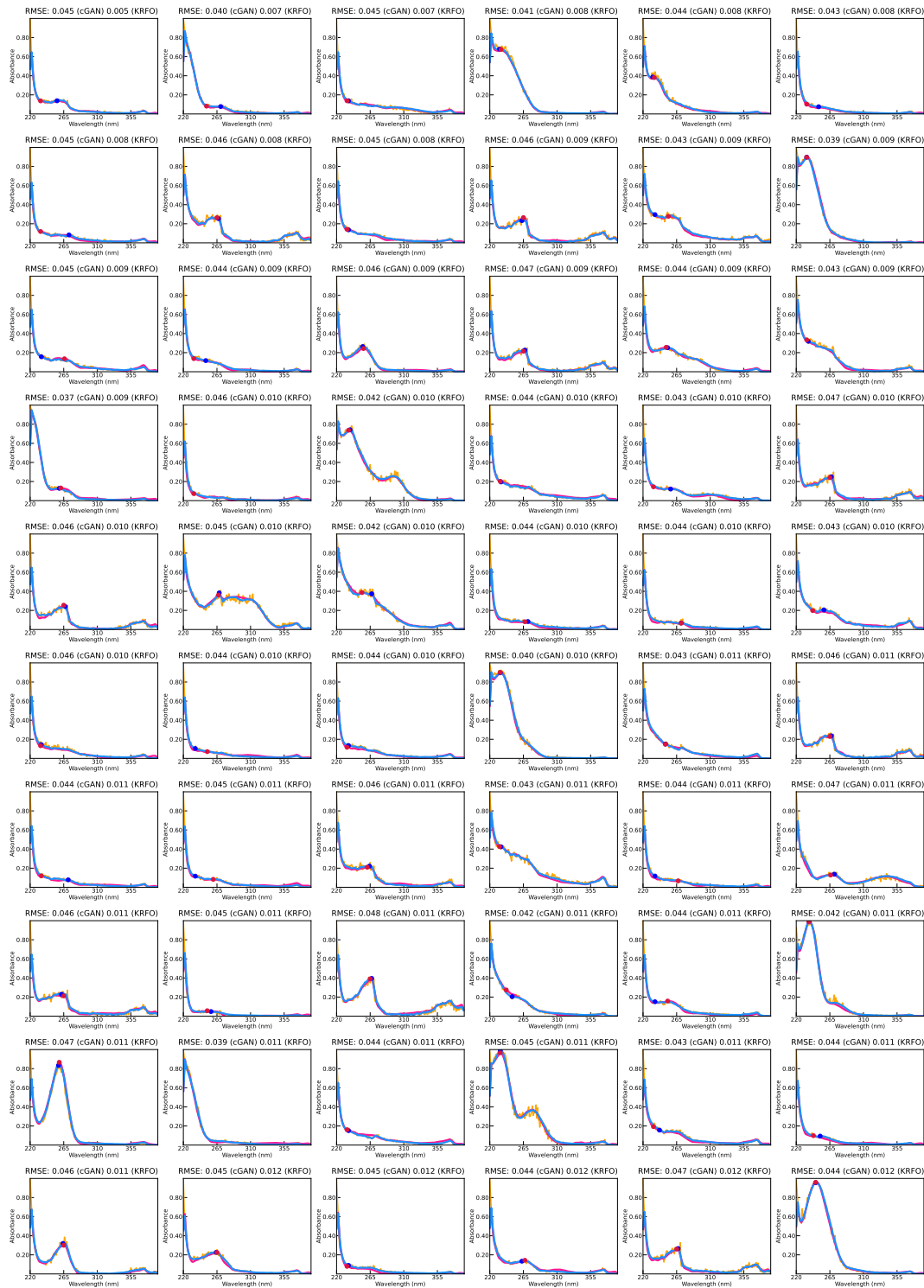


図 A.1.4 DatasetII: DatasetII でランダムに選択された訓練データの実験 UV-Vis スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) のフィッティング結果.

の候補の中で最も高い正解率を示すネットワーク構造を選択した: 四つの候補はそれぞれ (1) $1024 \rightarrow 500 \rightarrow 1$, (2) $1024 \rightarrow 660 \rightarrow 330 \rightarrow 1$, (3) $1024 \rightarrow 768 \rightarrow 512 \rightarrow 256 \rightarrow 1$, (4) $1024 \rightarrow 800 \rightarrow 600 \rightarrow 400 \rightarrow 200 \rightarrow 1$. ケース (b) では, λ_{max} の位置と強度の予測における検証データで, ピーク位置と強度の正解率が最良なモデルは (2) と (4) であった. ケース (c) では最小の RMSE を示すモデルは (3) であった.

USGS スペクトルライブラリからの図 5.3 以外のテスト分子の予測結果を示す. 合計で 68 個の分子をランダムに分割し, 80%, 10%, 10% をそれぞれ訓練データ, 検証データ, およびテストデータに割り当てた. データの分割は独立して 3 回繰り返した. 図 A.1.5 は, 図 5.3 以外の 21 個のテストデータの cGAN および KRFO の予測スペクトルと, 実験から得られたスペクトル関数を示す. 図 A.1.6 は, 訓練データに対する学習済みモデルのフィッティング結果を示す.

データセット I, データセット II, USGS スペクトルライブラリに対する KRFO の予測でのベースライン $\mu(t)$ を図 A.1.7 に示す. 図 A.1.7 から, ベースラインとしての $\mu(t)$ の絶対値は低いものの, データセットに共通して強度が高く出やすい箇所はベースラインの値も増加している. ベースラインの絶対値が低くなった要因は, データセットに含まれるスペクトル関数の形状が多様であり, 共通する値の表現が困難であったためと考えられる. スペクトル関数に関しては, 本研究では $\mu(t)$ としてベースラインを求めたが, 化学構造の種類によってベースラインもいくつかの種類に分けることが可能である. そのようにベースラインの変数に t だけでなく, 種類を区別する変数をさらに加えることによって, 各グループ毎に共通するベースラインを抽出することができ, 予測精度が向上する可能性もあると考えられる.

A.2 微細構造予測の実験結果補足

SEM 画像からの図 5.4 以外のテストデータの予測結果を示す. 合計で 123 個のサンプルをランダムに分割し 90%, 5%, 5% をそれぞれ訓練データ, 検証データ, およびテストデータに割り当てた. データの分割は独立して 3 回繰り返した. 図 A.2.9 は, 実験データを含む図 5.4 以外の全てのテストケースの cGAN および関数出力カーネル回帰の予測画像を示す. 図 A.2.10 は, 訓練データからランダムに選択された KRFO による 60 個のフィッティング結果を示す. また KRFO の予測でのベース画像 $\mu(t)$ を図 A.2.8 に示す. 図 A.2.8 から, μ を足す前のベースラインの画像は微細構造を表していないものの, カーネル基底関数にカーネル回帰係数を掛けて足し合わせただけの画像と合わせることによって, 微細構造が鮮明化している. ベースラインの画像はノイズ様であるが, 全微細構造に

共通する成分を示していると考えられる。

ピクセルごとに独立に MLP を行なった微細構造画像予測では、異なるネットワーク構造を持つ四つの候補を探索した。四つの候補はそれぞれ (1) $6 \rightarrow 64 \rightarrow 1$, (2) $6 \rightarrow 64 \rightarrow 32 \rightarrow 1$, (3) $6 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 1$, (4) $6 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 8 \rightarrow 1$. モデル (4) が、検証データにおいて最も低い RMSE を示した。

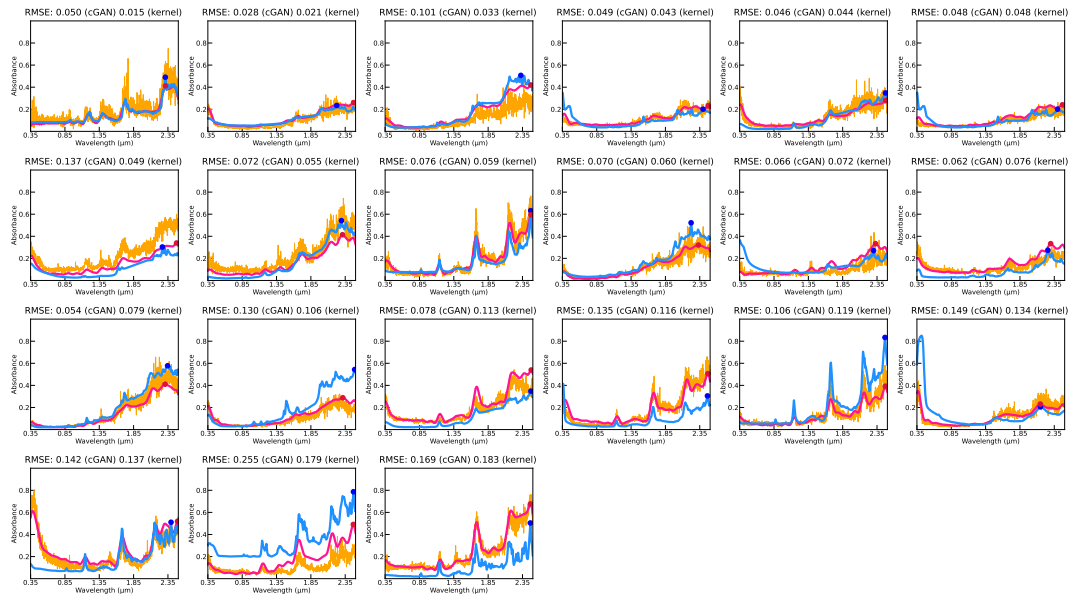


図 A.1.5 USGS:21 個のテストデータの実験 UV-Vis スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) の予測結果.

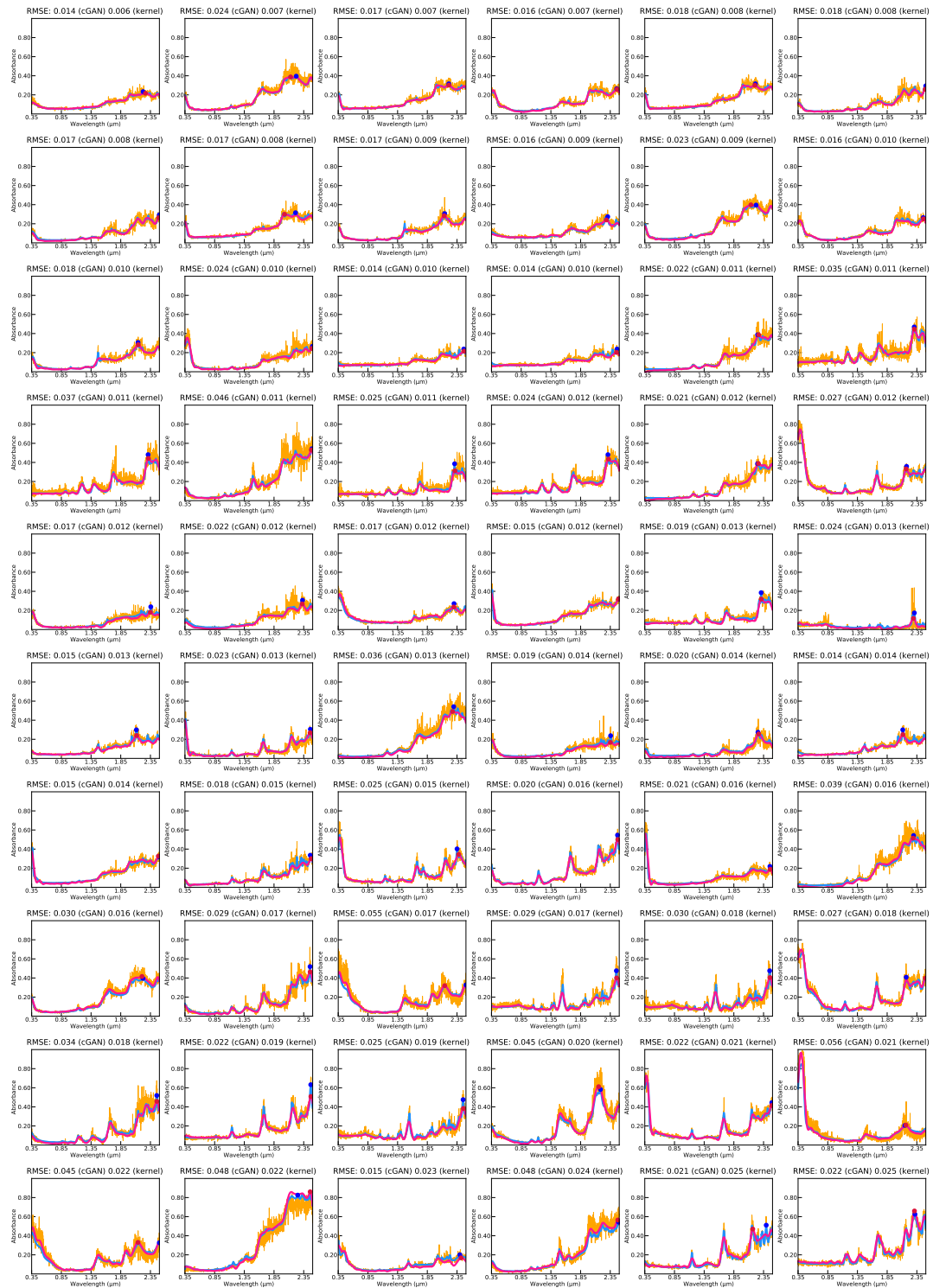


図 A.1.6 USGS: USGS でランダムに選択された訓練データの実験近赤外スペクトル (青) と KRFO (ピンク) および cGAN (オレンジ) のフィッティング結果.

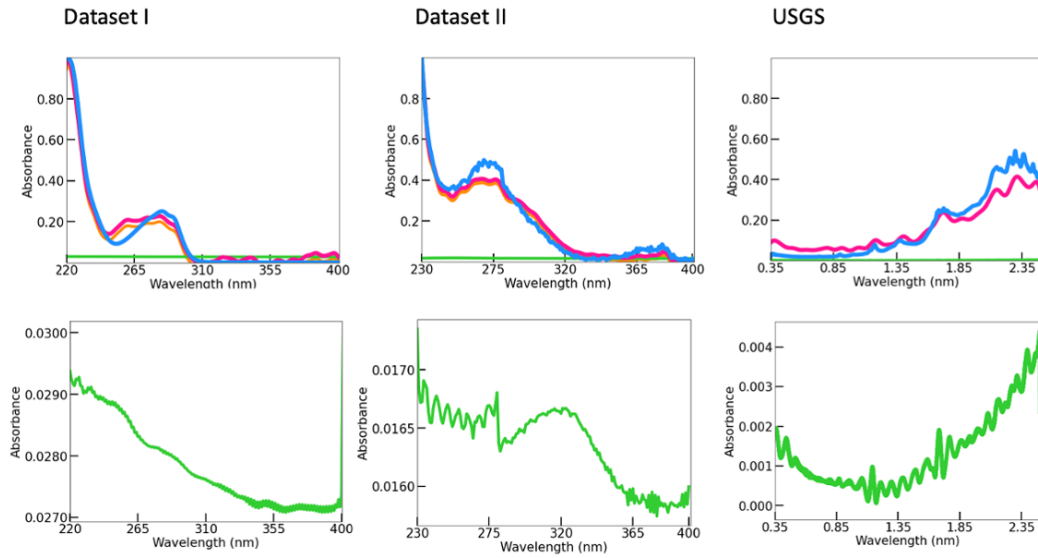


図 A.1.7 $\mu(t)$ の値 (緑) $\mu(t)$ を不足前の $\sum_{i=1}^d k(t, s_i) \beta_i(X)$ の値 (オレンジ) と、KRFO のスペクトル関数 (ピンク) と実験のスペクトル関数 (青) の結果。



図 A.2.8 μ による画像 (左) と、 $\mu(t)$ を不足前の $\sum_{i=1}^d k(t, s_i) \beta_i(X)$ による画像 (中) と、KRFO の予測画像 (右)。

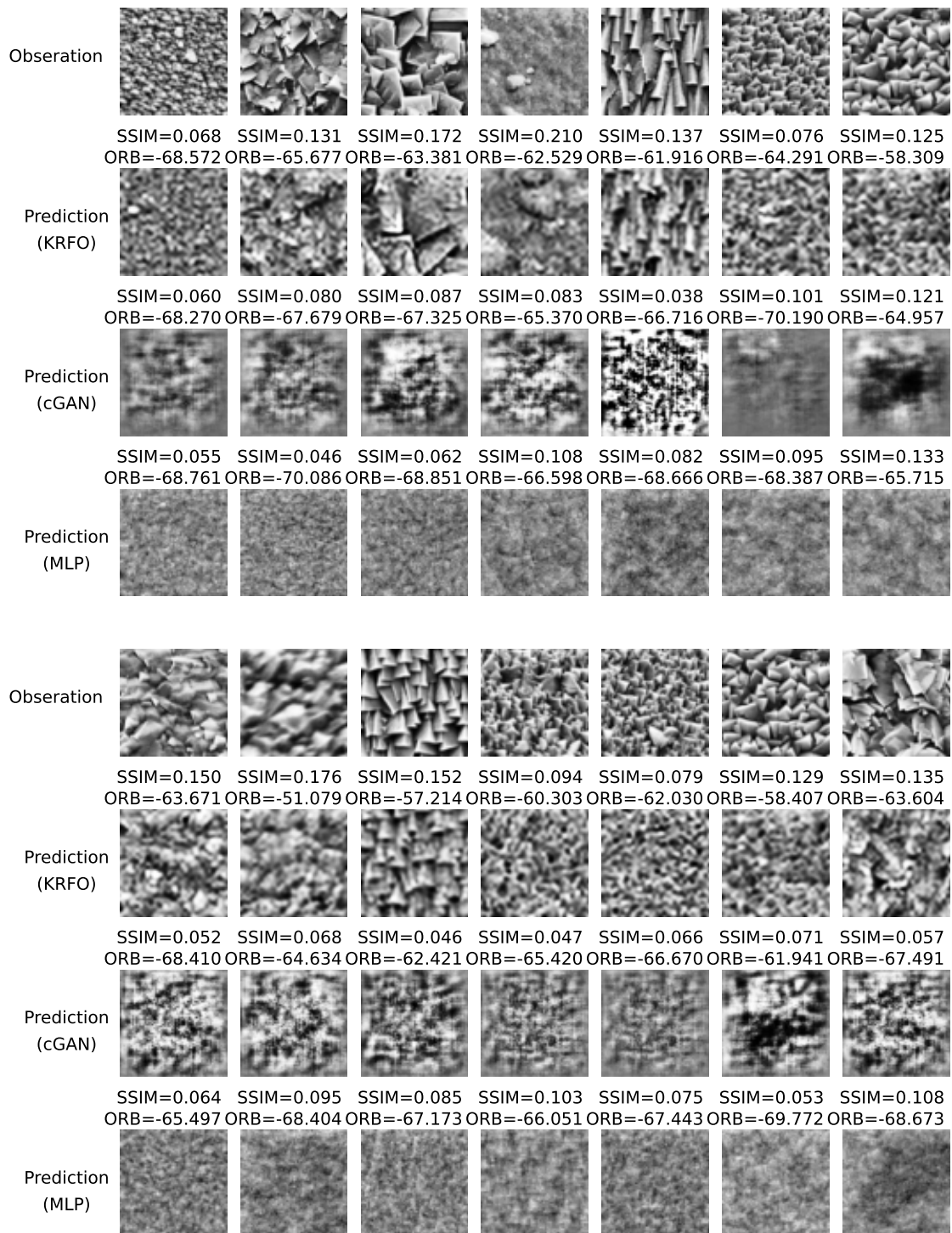


図 A.2.9 微細構造の予測結果: 学習済みの cGAN, KRFO, および MLP の 14 個のテストデータに対する予測画像と実験 SEM 画像.

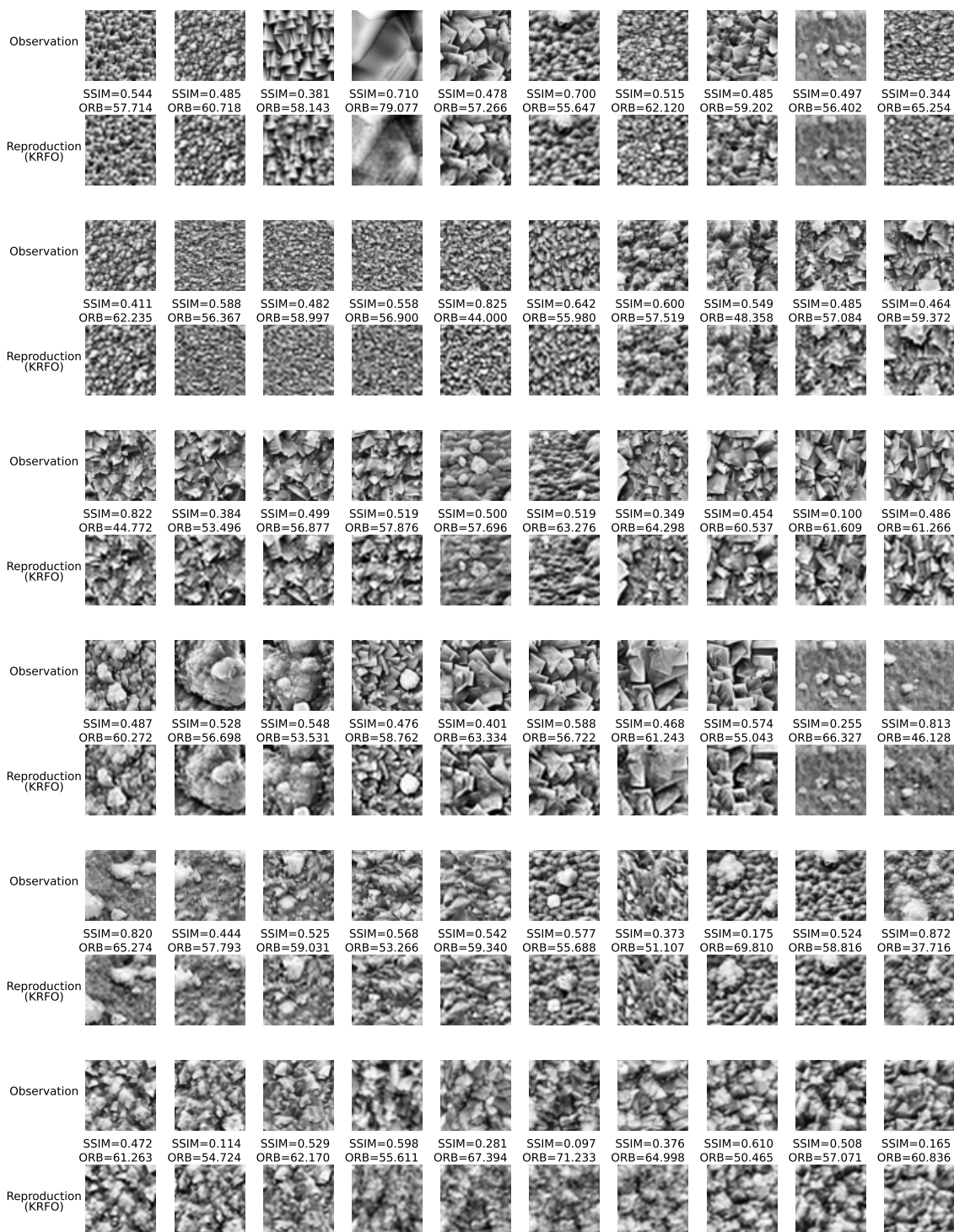


図 A.2.10 訓練データのフィッティング結果: 学習済みの KRFO により生成した 60 ケースの画像と対応する実験 SEM 画像.

謝辞

本学位論文は総合研究大学院大学複合科学研究科統計科学専攻にて、2019年10月から2022年9月までの研究成果をまとめたものです。まずは、本研究を進めるにあたり、懇切丁寧に御指導いただきました吉田亮教授に厚く心より御礼申し上げます。博士研究の構想からディスカッション、論文作成に至るまで、時間を惜しまず多大なご尽力をいただき本学位論文を書き上げるに至りました。言葉では表現し切れない思いがありますが、ここに深謝の意を表します。Wu Stephen 准教授には、研究に関しディスカッションや論文作成、発表のフォローをいただきました。Liu Chang 博士には、コード開発に多大なご助力をいただきました。お二方のお力添え無くしては本学位論文を書き上げることはできませんでした。重ねて深謝の意を表します。

研究生生活において研究室の皆様方にも支えられ、本論文を完成することができ深く感謝しております。また、授業やセミナーを通じご指導いただいた統計数理研究所の先生方にも心より御礼申し上げます。

勤務先である株式会社ダイセルでは、博士研究を行う機会を与えてくださった旧研究開発本部の西村久雄本部長、高部昭久本部長、高橋郁夫所長、博士研究を見守り続けていただいた生産本部の川口尚孝本部長、花村克幸所長、伊奈智秀リーダーに心より感謝申し上げます。また、基礎研究を続ける中で社会実装の視点も交えながら親身に相談にのっていただいた、中山隆文所長補佐、小園英俊氏、奥田貢一博士、井上慶三氏にも深く感謝申し上げます。他にもここには書き切れない多くの方々に支えていただき本学位論文を書き上げられました。ここに謹んで感謝申し上げます。

最後に支えていただいた家族の皆様はこの場を借りて御礼申し上げます。

2022年10月
岩山めぐみ

参考文献

- [1] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. Vol. 3, No. 54, 2017.
- [2] Christopher M Dobson. Chemical space and biology. *Nature*, Vol. 432, pp. 824–828, 2004.
- [3] 吉田亮. マテリアルズインフォマティクス概論. 統計数理, Vol. 69, No. 1, pp. 5–33, 2021.
- [4] 吉田亮, 山田寛尚, Liu Chang, Zhongliang Guo, Stephen Wu. マテリアルズ・インフォマティクス概説. Vol. 36, No. 1, 2018.
- [5] 吉田亮, 岩山めぐみ, グォチョンリャン. 材料研究の逆問題と機械学習. 応用物理, Vol. 90, No. 7, pp. 428–432, 2021.
- [6] 岡崎進, 吉井範行. コンピュータ・シミュレーションの基礎 (第2版) 分子のミクロな性質を解明するために. 化学同人, 東京, 2011.
- [7] Chang Liu, Erina Fujita, Yukari Katsura, Yuki Inada, Asuka Ishikawa, Ryuji Tamura, Kaoru Kimura, and Ryo Yoshida. Machine learning to predict quasicrystals from chemical compositions. *Adv. Mater.*, Vol. 33, No. 36, p. 2102507, 2021.
- [8] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B*, Vol. 95, No. 14, p. 144110, 2017.
- [9] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, Vol. 87, No. 18, p. 184115, 2013.
- [10] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties

- of inorganic crystals. *Nat. Commun.*, Vol. 8, No. 1, pp. 1–12, 2017.
- [11] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.*, Vol. 2, No. 1, pp. 1–7, 2016.
- [12] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model*, Vol. 50, No. 5, pp. 742–754, 2010.
- [13] Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. Pubchem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, Vol. 4, pp. 217–241, 2008.
- [14] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, Vol. 42, No. 6, pp. 1273–1280, 2002.
- [15] Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, Vol. 25, No. 2, pp. 64–73, 1985.
- [16] Ramaswamy Nilakantan, Norman Bauman, J Scott Dixon, and R Venkataraghavan. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *J. Chem. Inf. Model.*, Vol. 27, No. 2, pp. 82–85, 1987.
- [17] Kamal Choudhary, Brian DeCost, and Francesca Tavazza. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.*, Vol. 2, No. 8, p. 083801, 2018.
- [18] Hironao Yamada, Chang Liu, Stephen Wu, Yukinori Koyama, Shenghong Ju, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.*, Vol. 5, No. 10, pp. 1717–1730, 2019.
- [19] 船津公人, 佐藤寛子, 増井秀行. ケモインフォマティックス-予測と設計のための化学情報学. 丸善出版, 東京, 2005.
- [20] Hiroto Moriawaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *J. Cheminf.*, Vol. 10, No. 1, pp. 1–14, 2018.
- [21] Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, Vol. 5, No. 3, pp. 115–146, 1954.

- [22] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, Vol. 37, No. 1/2, pp. 17–23, 1950.
- [23] Gilles Moreau and P Broto. The autocorrelation of a topological structure: A new molecular descriptor. *New J. Chem.*, Vol. 4, No. 6, pp. 359–360, 1980.
- [24] Artem R Oganov and Mario Valle. How to quantify energy landscapes of solids. *J. Chem. Phys.*, Vol. 130, No. 10, p. 104504, 2009.
- [25] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, Vol. 120, No. 14, p. 145301, 2018.
- [26] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems 28*, pp. 2224–2232, Montreal, 2015.
- [27] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30*, pp. 992–1002, California, 2017.
- [28] Fabio Urbina, Kushal Batra, Kevin J Luebke, Jason D White, Daniel Matsiev, Lori L Olson, Jeremiah P Malerich, Maggie AZ Hupcey, Peter B Madrid, and Sean Ekins. Uv-advisor: Attention-based recurrent neural networks to predict uv–vis spectra. *Anal. Chem.*, Vol. 93, pp. 16076–16085, 2021.
- [29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112, Montreal, 2014.
- [31] Lars Banko, Yury Lysogorskiy, Dario Grochla, Dennis Naujoks, Ralf Drautz, and Alfred Ludwig. Predicting structure zone diagrams for thin film synthesis by generative machine learning. *Commun. Mater.*, Vol. 1, No. 1, pp. 1–10, 2020.
- [32] Zijiang Yang, Xiaolin Li, L Catherine Brinson, Alok N Choudhary, Wei Chen, and Ankit Agrawal. Microstructural materials design via deep adversarial learning methodology. *J. Mech. Des.*, Vol. 140, No. 11, 2018.

- [33] Xiaolin Li, Zijiang Yang, L Catherine Brinson, Alok Choudhary, Ankit Agrawal, and Wei Chen. A deep adversarial learning methodology for designing microstructural material systems. In *Volume 2B: 44th Design Automation Conference*, Quebec, 2018.
- [34] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*, Vol. 77. Springer, New York, 2002.
- [35] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.00473*, 2014.
- [36] Rich Caruana. Multitask learning. *Machine Learning*, Vol. 28, No. 1, pp. 41–75, 1997.
- [37] Yu Zhang and Qiang Yang. An overview of multi-task learning. *Natl. Sci. Rev.*, Vol. 5, No. 1, pp. 30–43, 2018.
- [38] Megumi Iwayama, Stephen Wu, Chang Liu, and Ryo Yoshida. Functional output regression for machine learning in materials science. *Journal of Chemical Information and Modeling*, Vol. 62, No. 20, pp. 4837–4851, 2022.
- [39] Marc A Baldo, David F O’Brien, Y You, A Shoustikov, S Sibley, Mark E Thompson, and Stephen R Forrest. Highly efficient phosphorescent emission from organic electroluminescent devices. *Nature*, Vol. 395, No. 6698, pp. 151–154, 1998.
- [40] R Gómez-Bombarelli, J Aguilera-Iparraguirre, T. D Hirzel, D Duvenaud, D Maclaurin, M. A Blood-Forsythe, Hyun Sik Chae, M Einzinger, Dong-Gwang Ha, Tony Wu, Georgios Markopoulos, Soonok Jeon, Hosuk Kang, Hiroshi Miyazaki, and Masaki Numata. Kim, s.; and huang, w.: and hong, s. i.; and baldo, m.; adams, r. p.; aspuru-guzik, a. design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.*, Vol. 15, No. 10, pp. 1120–1127, oct 2016.
- [41] Katherine A Mazzio and Christine K Luscombe. The future of organic photovoltaics. *Chem. Soc. Rev.*, Vol. 44, No. 1, pp. 78–90, 2014.
- [42] Nadim A Shaath. Ultraviolet filters. *Photochem. Photobiol. Sci.*, Vol. 9, No. 4, pp. 464–469, 2010.
- [43] Miguel Marques, Angel Rubio, Eberhard KU Gross, Kieron Burke, Fernando Nogueira, and Carsten A Ullrich. *Time-dependent density functional theory*, Vol. 706. Springer Science & Business Media, Berlin, 2006.

- [44] Rafael Mamede, Florbela Pereira, and João Aires-de Sousa. Machine learning prediction of uv – vis spectra features of organic compounds related to photoreactive potential. *Sci. Rep.*, Vol. 11, pp. 1–11, 2021.
- [45] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, Vol. 28, No. 1, pp. 31–36, 1988.
- [46] Birendra Bahadur. Liquid crystal displays. *Molecular Crystals and Liquid Crystals*, Vol. 109, No. 1, pp. 3–93, 1984.
- [47] B.H. Bersted and T.G. Anderson. Influence of molecular weight and molecular weight distribution on the tensile properties of amorphous polymers. *Journal of Applied Polymer Science*, Vol. 39, No. 3, pp. 499–514, 1990.
- [48] J. Tersoff and D. R. Hamann. Theory of the scanning tunneling microscope. *Phys. Rev.B*, Vol. 31, No. 2, pp. 805–813, 1985.
- [49] 佐藤公隆. 材料開発を支える電子顕微鏡技術—表面・断面 (界面) 観察技術におけるその位置づけ—. *表面技術*, Vol. 54, No. 1, pp. 3–9, 2003.
- [50] Lihua Chen, Chiho Kim, Rohit Batra, Jordan P Lightstone, Chao Wu, Zongze Li, Ajinkya A Deshmukh, Yifei Wang, Huan D Tran, Priya Vashishta, Gregory A. Sotzing, Yang Cao, and Rampi Ramprasad. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput. Mater.*, Vol. 6, No. 1, pp. 1–9, 2020.
- [51] Zhantao Chen, Nina Andrejevic, Tess Smidt, Zhiwei Ding, Qian Xu, Yen-Ting Chi, Quynh T. Nguyen, Ahmet Alatas, Jing Kong, and Mingda Li. Direct prediction of phonon density of states with euclidean neural networks. *Advanced Science*, Vol. 8, No. 12, p. 2004214, 2021.
- [52] AJ Koster, Ulrike Ziese, AJ Verkleij, AH Janssen, and KP De Jong. Three-dimensional transmission electron microscopy: a novel imaging and characterization technique with nanometer scale resolution for materials science. *J. Phys. Chem. B*, Vol. 104, No. 40, pp. 9368–9370, 2000.
- [53] SC Garcea, Ying Wang, and PJ Withers. X-ray computed tomography of polymer composites. *Compos. Sci. Technol.*, Vol. 156, pp. 305–319, 2018.
- [54] Akihisa Takeuchi and Yoshio Suzuki. Recent progress in synchrotron radiation 3d–4d nano-imaging based on x-ray full-field microscopy. *Microscopy*, Vol. 69, No. 5, pp. 259–279, 2020.

- [55] U. Grenander. Stochastic processes and statistical inference. *Arkiv for Matematik*.
- [56] C.R. Rao. Some statistics methods for comparison of growth curves. *Biometrics*, Vol. 14, pp. 1–17.
- [57] J.O. Ramsay. When the data are functions. *Psychometrika*, Vol. 47, pp. 379–396.
- [58] James O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 1997.
- [59] 松井秀俊. 関数データに基づく統計的モデリング. *統計数理*, Vol. 67, No. 1, pp. 73–96, 2019.
- [60] Rob J Hyndman and Md Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, Vol. 51, No. 10, pp. 4942–4956, 2007.
- [61] Alexander Aue, Diogo Durbart Norinho, and Siegfried Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, Vol. 110, No. 509, pp. 378–392, 2015.
- [62] Hochreiter Sepp and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.
- [63] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
- [64] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- [65] Misra Ishan, Shrivastava Abhinav, Abhinab Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. *arXiv preprint arXiv:1604.03539*, 2016.
- [66] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. *ICML*, 1993.
- [67] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, Vol. 28, pp. 7–39, 1997.
- [68] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations*, 2014.
- [69] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.
- [70] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an

- efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564–2571, Barcelona, 2011. IEEE.
- [71] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, pp. 430–443, Graz, 2006. Springer.
- [72] RF Kokaly, RN Clark, GA Swayze, KE Livo, TM Hoefen, NC Pearson, RA Wise, WM Benzel, HA Lowers, RL Driscoll, and Annna J. Klein. Usgs spectral library version 7 data: Us geological survey data release. *United States Geological Survey (USGS): Reston, VA, USA*, 2017.
- [73] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.