（ 続紙 1 ）

京都大学	博士（情報学）	氏名	朱　見深
論文題目	Novel Methods for Chemical Compound Inference Based on Machine Learning and Mixed Integer Linear Programming （機械学習と混合整数線形計画法に基づく新しい化合物推定手法）		

（論文内容の要旨）

所望の物性値を持つ化合物構造を推定する手法は創薬や材料開発の領域で応用を持ち、Inverse QSARの名の下で研究が進められてきた。本論文では、化学グラフ（節点に元素のラベル、枝に多重度が割り当てられたグラフ）と物性値のデータが与えられたとき、化学グラフを特徴ベクトルに変換して予測関数を構築し、その逆解析を混合整数線形計画法（MILP）によって行い、逆解析によって得られた化学グラフの構造異性体を列挙するという手法の枠組みに関して研究を行っている。この枠組みで取扱可能な化学グラフは閉路指数が1以下のものに限定され、実用化に向けての課題が数多く残されていた。本論文ではそれらの課題を解決するための研究が展開されている。

第1章は序論であり、本研究の背景、先行研究、および目的について述べた後、結果の概要を説明している。第2章は準備として位置付けられ、本論文を読み進める上で必要な用語の定義や記法の導入を行っている。第3章では本研究の根幹を成す上記枠組みに関して概説を行っている。なお逆解析MILPの制約条件は、(C1) 予測関数の計算過程を模倣するためのものと、(C2) 特徴ベクトルと化学グラフを対応づけるためのものの二つから成る。

第4章および第5章の主題は、取扱可能なグラフクラスの拡張である。第4章では、閉路指数が2であるような化学グラフを取扱可能とするための方式を提案している。そのようなグラフのポリマートポロジーが3通りであることに着目し、これらすべてを部分グラフとして含むようなトポロジー構造（スキームグラフ）を導入することで、化学グラフのポリマートポロジーとして3通りが選択的に用いられるような制約を（C2）に組み込むことに成功している。第5章では、ρ-leanという新奇なグラフクラスを導入し、ρ-lean環状化学グラフを取扱可能とするための方式を提案している。ρ-lean環状グラフは核、内部、外部の三つの部分グラフに分割されるが、ユーザが指定した核の概形構造（種グラフ）を拡張し、内部と外部を接続することで化学グラフを構築するという制約を（C2）に組み込むことに成功している。またこの方式で得られる化学グラフの構造異性体を高速に列挙するため、従来アルゴリズムの改良版を設計している。

第6章から第8章の主題は、精度の高い予測関数の構築である。第6章では、予測関数として線形回帰（LR）を用いることを提案し、逆解析MILPの制約（C1）を定式化している。第7章では適応的線形回帰（Adjustive LR; ALR）という新しい予測関数のモデルを提案している。LRは、隠れ層が無く、入力素子と出力素子のすべてに線形活性化関数が与えられた人工ニューラルネットワーク（ANN）とみなすことができるが、ALRはより広いクラスの活性化関数を用いることが可能なモデルである。ある仮定の下で、ALRを構築する問題は線形計画問題として定式化できることを示している。またALRによって得られた予測関数に対し、逆解析MILPの制約（C1）を定式化している。第8章では二つの記述子の積で定まる二次記述子を特徴ベクトルの一部として用いることを提唱し、特徴選択アルゴリズムの設計、および二次記述子が存在する場合の逆解析MILPの制約（C1）を定式化している。

第4章から第8章の各章では、予測関数の精度、逆解析MILPおよび構造異性体列挙の計算時間等を評価基準とした、提案手法の有効性を検証するための計算実験が行われている。

第9章は結論で、本論文のまとめと今後の展望が触れられている。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、混合整数線形計画法（MILP）を用いたInverse QSARの枠組みに対して実用化に向けた研究を展開するものであり、得られた成果は以下のとおりである。

（１）Inverse QSAR、すなわち所望の条件を満たす化合物構造（化学グラフとして表現される）を推定するための研究は創薬や材料開発の領域に応用を持つ。この問題に対する一般的なアプローチは、化学グラフの物性値を予測するための予測関数を構築し、その逆解析を行うというものだが、困難であることが知られている。従来法にはある意味での最適性や厳密性が保証されておらず、その上、実用的な意味で取扱可能なグラフのサイズには限界があった。例えばある従来法は、数日にわたる計算を行ったとしても、（水素原子に加えて）高々20程度の非水素原子を持つ化学グラフしか推定することができない。逆解析をMILPとして定式化して解く枠組みでは、最適性と厳密性は保証されているものの、モデルとして取扱可能なグラフは閉路指数が1以下のものに限定されていた。本論文では、第一の成果として、閉路指数が2のグラフを取扱可能とするためのモデルを提案している。また第二の成果として、ρ-leanという新しい環状化学グラフのクラスを提案し、それを取扱可能とするためのモデルを提案している。いずれのモデルに対しても逆解析MILPが提案され、ソルバの計算を高速化するための工夫がなされている（たとえば第一のモデルについては変数や制約の数が線形オーダーで抑えられるように設計されている）。取扱可能なグラフの範囲を拡張したことは大きな進歩と言える。たとえば世界最大級の規模を誇るPubChemデータベースにおいて、閉路指数が1以下の化合物の割合は16%程度に過ぎないが、第一のモデルによって取扱が可能となった閉路指数が2の化合物の割合は28%程度に達する。また、同データベースにおける環状化学グラフの割合は97%程度だが、そのうち99%超が2-leanクラスに属し、第二モデルはこれらすべてを取扱可能とするもので、特筆に値する。また実データを用いた計算実験では、両モデルとも実用的な時間の範囲内で30～50程度の非水素原子を持つ化学グラフを推定することに成功している。これは上で述べた従来法に比べると顕著な差である。

（２）予測関数の精度は逆解析の結果の信頼性につながる。一方、あらゆる物性に対して一様なアプローチで高精度の予測関数を構築することは難しいため、個々の物性に対して最も適した手法を用いることが求められる。従来、MILPを用いた枠組みでは人工ニューラルネットワーク（ANN）のみを予測関数のモデルとして取り扱っており、このような視点に立った研究にまで手が届いていなかった。本論文では予測精度向上のために線形回帰（LR）や特徴選択を適用することを提案し、さらに適応的線形回帰（ALR）という新しいLRのモデルを提案している。その結果いくつかの物性において、これまでANNでは成し得なかった予測精度を達成することに成功している。ALRは従来のLRとANNの中間に位置するという点で数理的に興味深い手法だが、線形計画法によって効率良く学習可能な点、物性によってはラッソLRやANNを凌ぐ予測精度を達成可能なことが示されている点は実用上大きな意義があり、化学物性に限らず、他領域のデータに対する適用が期待される手法である。

このように、本論文はMILPを用いたInverse QSAR手法に関して、実用化に向けた数理モデルの構築を行い、様々な物性に関する実データを用いた実験を通じてその有効性を示している。本論文には大きな学術的価値が認められ、博士（情報学）の学位論文としてふさわしい高度な内容を持っているものと認める。

論文内容とそれに関連した事項に関する口頭試問を令和５年７月１９日に実施し、合格と認めた。また、本論文のインターネットでの全文公表についても支障がないことを確認した。