Form 3

Abstract of Thesis

Name (MARTIN)	
Title	Applications of machine learning and literature data mining in drug discovery (創薬における機械学習と文献データマイニングの応用)

Abstract of Thesis

Decades of extensive research in drug discovery have generated a substantial amount of data, prompting a shift towards literature data mining and machine learning to uncover hidden relationships within individual studies for predictive purposes. The objective of this dissertation is to apply literature data mining and machine learning for constructing *in silico* models as a practical alternative to conventional *in vivo* and *in vitro* techniques. Case studies were presented to showcase the predictive capabilities of the models, focusing on the cellular toxicity prediction of amorphous silica nanoparticles (SiO_2-NPs) and the activity prediction of small molecules targeting urate transporter 1 (URAT1).

SiO_2-NPs, extensively used in various industries, are regarded as a safety issue by the Scientific Committee on Consumer Safety (SCCS) despite their large-scale production. A generic *in silico* prediction model was developed to enhance the safety assessment of SiO_2-NPs, a type of nanoparticle with a diameter of 100 nm or less. This involved meticulous data collection from over 100 scientific papers and the application of an advanced machine learning algorithm (CatBoost) to construct a predictive model. The model highlighted key factors influencing their safety, including concentration, serum presence, size, exposure time, and surface properties. The results indicated that modifying the surface of SiO_2-NPs and using them at low concentrations significantly improves their safety.

URAT1, a protein with an unknown experimental 3D structure, emerges as a promising target for the development of innovative anti-hyperuricemic drugs due to its crucial role in reabsorbing over 90% of uric acid. A bespoke ligand-based drug design (LBDD) pipeline was developed to discover novel compound skeletons targeting URAT1. This involved generating the largest dataset, surpassing the ChEMBL dataset by 4.5 times, to train a successful predictive model that provided insights into discriminating high from low active URAT1 inhibitors using key molecular descriptors and counteractively generated compounds. The pipeline, evaluated for attaining 95% precision from two perspectives—the model's probability and the principal component of SHapley Additive exPlanations values, was applied to screen the massive ZINC database. This led to the identification of 22 promising potential lead compounds with unique structures, all anticipated to exhibit favorable activity and pharmacokinetics.

In summary, this dissertation underscores a pivotal shift in drug discovery, leveraging extensive research data through literature data mining and machine learning. This paradigm shift is rooted in the understanding that relying solely on conventional methods is less humane, efficient, and economical, while the *in silico* models offer cost-effective and rapid assessment of the potential toxicity or activity of specific nanoparticles or small molecules.

様式７

論文審査の結果の要旨及び担当者

氏　　名　（ MARTIN ）		
	（職）	氏　　名
論文審査担当者	主 査　教授　水口　賢司 副 査　教授　井上　豪 副 査　教授　福澤　薫 副 査　准教授　東阪　和馬	

論文審査の結果の要旨

申請者は、文献データマイニングと人工知能（AI）・機械学習という2つの技術を組み合わせて、薬学分野の予測研究に寄与する一般的な枠組みを提唱すると共に、開発した手法を2つの特定の問題に応用し、精度の高い予測方法の確立と当該現象についての新たな生物学的知見の創出に成功した。

ナノ粒子は化粧品、塗料、繊維、電子機器などに幅広く利用されているが、従来の材料とは異なる形状や性質のため、安全性を懸念する指摘がある。一方で、ナノ粒子が細胞等に与える潜在的な影響を精度よく予測することはこれまで困難だった。幾つかの特定の種類のナノ粒子については、その設計に関するパラメータなどを入力として細胞毒性の有無を予測する（出力する）研究が発表されているが、1）予測モデル構築に用いたデータセットを分割した内部検証のみが行われており、予測の汎化性能が適切に評価されていない、2）ナノシリカ-コロナ複合体（ナノシリカが血液のような生体環境で生体分子と相互作用する際に形成される生体分子層）の影響が考慮されていない、という問題があった。

この問題に対して、申請者はまず、2004年から2016年の間に発表された100以上の科学論文を丹念に調査し、ナノ粒子の一種であるナノシリカ（SiO2-NPs, 二酸化ケイ素のナノ粒子）の安全性予測モデル構築に必要な情報を収集した。このデータセットは、物理化学的特性、実験条件、細胞種など、ナノシリカの安全性に関連する多様な属性をコンピュータフレンドリーな形式に整理し、4124という過去にない大規模なデータ数を持つもので、この作成自身が新規で価値ある研究成果と言える。次に、各種AI・機械学習アルゴリズムを用いて、ナノシリカの安全性を評価するための汎用的で説明可能なインシリコ予測モデルを確立した。この手法により、濃度、粒子径、血清の有無などを含む、ナノシリカの安全性に影響する13の主要な属性を特定した。さらに、AIモデルのshapley Additive exPlanations（SHAP）値により算出した属性の重要度を解析することにより、ナノシリカのサイズが大きいこと、ナノシリカの懸濁液中にFBS （10%のウシ胎児血清）が含まれていること、ナノシリカの表面修飾があること、などの条件が低い細胞毒性につながることを示した。特に、血清の有無に関する属性が特定されたことで、ナノシリカ-コロナ複合体が果たす役割が初めて明らかになった。また、申請者は2017年-2022年に発表された論文及び、共同研究者による新規の実験データからなる、学習データとは独立のデータセットを構築し、それを用いた外部検証を実施することで、構築した予測モデルが高い予測精度と汎化性能を示すことを実証した。

第２の応用として、同様なアプローチに基づき、申請者は腎臓での尿酸再吸収に関わるトランスポーターURAT1の阻害剤探索を行った。URAT1阻害剤は高尿酸血症治療に用いられているが、より有効性と安全性の高い医薬品の開発が望まれている。文献データマイニングによって公共のデータベースよりも大規模な学習データセットを構築し、化学構造から阻害活性を予測する機械学習モデルを開発した。これにより、22の新規の阻害剤候補の提案に成功した。

これらの成果は、上で述べた既存の研究の課題を克服し、新規な方法論の確立と生物学的知見の創出に成功しただけでなく、インシリコによる生物活性の予測という広い分野に対して重要な貢献をしたと位置付けることができる。

以上により、本研究は、博士（薬科学）の学位論文に値するものと認める。