



Title	Enhancing QSAR-based Predictions of Natural Product Bioactivity through Multi-Task Learning with Protein Evolutionary-Relatedness Metrics
Author(s)	Donny, Ramadhan
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105652
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Abstract of Thesis

Name (Donny Ramadhan)

Title

Enhancing QSAR-based Predictions of Natural Product Bioactivity through Multi-Task Learning with Protein Evolutionary-Relatedness Metrics

(タンパク質の進化的類縁性指標を用いたマルチタスク学習による天然物生物活性予測性能の向上)

Abstract of Thesis

Natural products possess diverse and generally three-dimensional structures that may play crucial roles in their interactions with drug targets. Considering the scarcity of natural product bioactivity data in public databases, multi-task learning (MTL) emerges as a valuable approach to enhance quantitative structure-activity relationship (QSAR) prediction performance. This study aimed to enhance MTL by incorporating protein evolutionary-relatedness metrics to improve the natural product bioactivity prediction when data are limited, while also determining the conditions in which MTL achieves optimal performance. Using binary classification filtering, a curated dataset of predicted natural products with bioactivity toward enzymes from ChEMBL was constructed. Single-task learning (STL) was used as the baseline model, while feature-based MTL (FBMTL) was trained on all proteins within each group. Instance-based MTL (IBMTL), which extends FBMTL, incorporated protein evolutionary-relatedness metrics as additional features. The results indicated that IBMTL consistently surpassed STL and FBMTL in most protein groups, highlighting the benefit of including protein evolutionary relatedness. Significant improvements were observed in the kinase and cytochrome P450 groups, where proteins are categorized into specific levels within ChEMBL's six-level protein classification. Within the kinase group, IBMTL achieved its best performance at the target parent level, suggesting a trade-off between dataset size and protein evolutionary relatedness. Overall, these findings demonstrate that integrating protein evolutionary information into MTL is a valuable strategy for natural-product drug discovery, even under data-limited conditions.

Keywords: protein evolutionary relatedness, feature-based multi-task learning (FBMTL), instance-based multi-task learning (IBMTL), single-task learning (STL)

Bioactivity Data of Compounds Against Proteins Under the Enzyme Group in ChEMBL's Protein Hierarchy

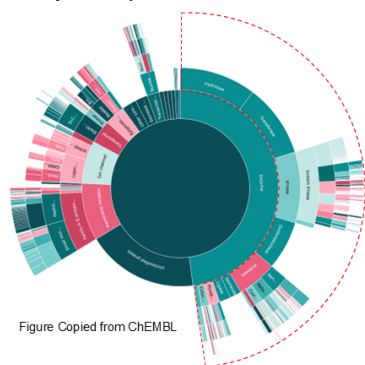
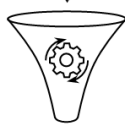
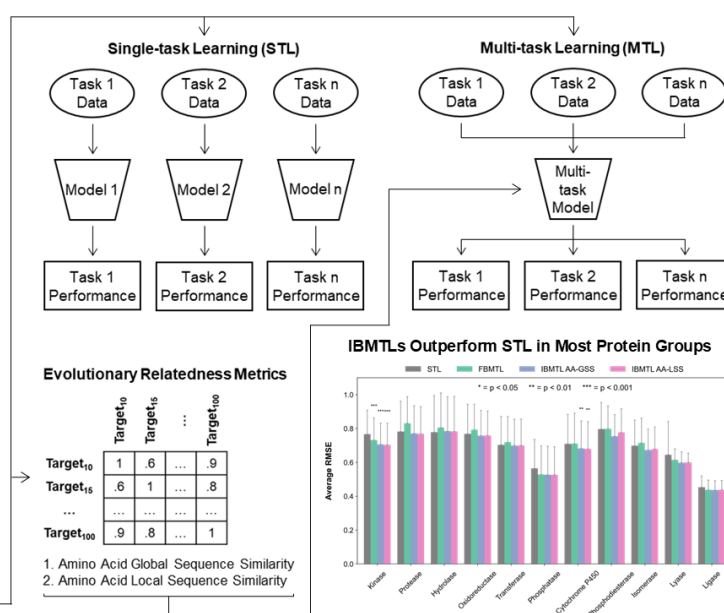


Figure Copied from ChEMBL

Optimized Natural Product Filter



IBMTL: MTL with Protein Similarity Features; FBMTL: MTL without Protein Similarity Features



論文審査の結果の要旨及び担当者

氏 名 (Donny Ramadhan)			
論文審査担当者	(職)		氏 名
	主 査	教授	水口 賢司
	副 査	教授	船橋 靖博
	副 査	教授	水谷 泰久

論文審査の結果の要旨

天然物由来化合物（以下、天然物という）は、複雑で多様な立体構造を持つことから、酵素などの主要な創薬標的への結合性に優れ、医薬品候補の重要な源泉である。しかし、天然物に関する生物活性の実験データは非常に限られていることが多く、コンピュータモデルを用いた新薬発見は困難であった。既知の実験データを統計的に解析し、化合物の化学構造だけからその生物活性を予測する定量的構造活性相関（Quantitative structure–activity relationship; QSAR）という計算手法が知られているが、従来の QSAR モデルは、1) 単一のタンパク質標的のみに焦点を当てており、2) 学習に用いる実験データが少ない場合には予測性能が著しく低下するという問題があった。

この問題に対して、申請者は、複数の関連するタンパク質について同時に学習を行う「マルチタスク学習」という手法を導入することで、データ不足の問題を克服しようと試みた。本研究の独創性は、このマルチタスク学習に「タンパク質の進化的情報」を組み込んだ点にある。共通祖先に由来し進化的に関連するタンパク質群は機能が似ていることが多いという観察事実に着目し、進化情報をモデルの学習過程の指針として利用した。これにより、標的タンパク質間の関連性をより効率的に利用できるモデルの構築が可能になった。申請者は、厳密なベンチマークを実行し、本アプローチが従来法と比較して、一貫して高い予測性能を発揮することを示した。特に、創薬における重要な標的であるキナーゼやシトクロム P450 といった酵素ファミリーにおいて、その有効性が顕著であった。さらに、本研究は、利用可能なデータ量とタンパク質間の進化的類似性のバランスをとることで、マルチタスク学習が最も効果的に機能する条件も明らかにした。

これらの成果は、実験データが限られているという天然物創薬における長年の課題を克服し、タンパク質の進化情報を機械学習モデルに統合するという新規な方法論を確立したものである。これは、インシリコ（コンピュータ内）での医薬品開発、特にデータが乏しい状況下での活性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（理学）の学位論文に値するものと認める。