様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月　 ９ 日現在

機関番号：８４４２０
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ１６４３６

研究課題名（和文）血漿タンパク結合データを利用した分布容積予測モデルの構築

研究課題名（英文）in silico prediction of volume of distribution using fraction unbound in plasma

研究代表者

渡邉　怜子（Reiko, Watanabe）

国立研究開発法人医薬基盤・健康・栄養研究所・医薬基盤研究所　ＡＩ健康・医薬研究センター・プロジェクト研究員

研究者番号：３０７２７３２６

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：まずChEMBLなどの複数のデータベースから分布容積のデータを収集しキュレーションを実施し、1314化合物の分布容積データセットを構築した。次に機械学習法を用いて回帰モデルにより分布容積を予測した。Test setの予測結果はR2=0.63, RMSE＝0.40であり、既存モデルと同等の精度を示した。さらに、血漿タンパク結合値を記述子として入力した場合とそうでない場合の精度比較を実施し、予測モデルは血漿タンパク結合値の入力に関わらず同等の精度を示すことが明らかとなった。

研究成果の学術的意義や社会的意義
　分布容積は、薬がヒトに投薬されたときに示す挙動を知るうえで重要なパラメータの１つである。創薬の初期に構造情報から分布容積を予測することで、上市までにかかる時間の短縮・臨床段階での逸脱確率の減少などの効果が期待されており、近年急速に発展するin silicoによる薬物動態パラメータ予測は実用化が進んでいる。血漿タンパク結合率は分布容積に影響を及ぼすパラメータであり、スクリーニングにおいて比較的簡便に収集可能であることから、血漿タンパク結合率を利用して分布容積をより高精度に予測することが可能となれば、創薬の効率化に貢献することができる。

研究成果の概要（英文）：The volume of distribution data were collected from multiple databases such as ChEMBL and curated to construct the data set containing 1314 compounds. Then, the distribution volume was predicted by a regression model using a machine learning method, and the prediction results in the test set were R2=0.63 and RMSE=0.40, showing accuracy equivalent to that of existing models. Furthermore, a comparison of the accuracy with and without the input of plasma protein binding values as descriptors was conducted, and it was shown that the prediction model showed equivalent accuracy regardless of the input of plasma protein binding values.

研究分野：薬物動態

キーワード：薬物動態予測　in silico　機械学習

## １．研究開始当初の背景

近年創薬の基礎研究から創薬シーズ創出までの研究を行うアカデミア創薬が推進されており、アカデミア研究者は創薬シーズの探索・同定から薬効・薬物動態・安全性の評価など、これまで主に企業が行ってきた創薬応用研究の実施が必要となっている。しかし、基礎研究を目的としたアカデミア研究者の志向性や研究資金の不足などが原因で、アカデミアで発見された新しい創薬ターゲットの研究成果は実用化に向けてのデータが不足する傾向がみられ、現状では医薬品の創出に結びついているとは言い難い。 アカデミア創薬における成功率の向上を目的とした支援の１つとして注目されているのが、in silico による ADMET（Absorption, Distribution, Metabolism, Excretion and Toxicity:吸収、分布、代謝、排泄、毒性）予測である。研究開発の早期に創薬シーズの ADMET プロファイルを判定することは効果的な研究開発に必要不可欠であるが、アカデミア創薬において実験による ADMET 判定を早期に行うことは現状としては非常に難しい。しかし近年は薬物動態や毒性に関するデータ量が大幅に増加しており、in silico で化学構造のみから薬物動態パラメータを直接予測し、低コストで ADMET 判定を行うことができるようになってきた。その一方で、無償で利用できる予測モデルは LogP などの比較的シンプルなパラメータに限られており、創薬において重要な薬物動態パラメータである血漿タンパク結合率、代謝安定性、クリアランス、分布容積などは高額な有償ソフトウェアでしか予測できず、アカデミア創薬への導入はほとんど進んでいない。

申請者はこれまで、血漿タンパク結合率の予測モデルを構築し公開した（Watanabe et al. Mol. Pharm. 2018）。この予測モデルは、これまでで最大のデータセット及び無償の記述子計算ソフトウェアを用いて構築され、化合物の構造情報だけで血漿タンパク結合率の予測を行うことが可能である。企業で一般的に利用されている有償ソフトウェアに収録されている血漿タンパク結合率予測モデルよりも高精度であり、特に血漿タンパク結合率の高い範囲での予測精度が向上している。この予測モデルの無償公開により、これまで有償のソフトウェアでしか予測できなかった因子の１つである血漿タンパク結合率の予測を誰もが無償で行うことが可能となった。

## ２．研究の目的

本研究の目的は、申請者がこれまでに構築した血漿タンパク結合予測モデルによる fu,p の予測値や学習データを利用して分布容積予測モデルを構築すること、さらには無償のソフトウェアを用いることでモデルの公開を可能にし、アカデミア創薬の加速化に貢献することである。これまでにいくつかの分布容積予測モデルが報告され、ある程度の精度が示されてはいるものの、そのほとんどが有償の記述子計算ソフトに依存しており誰もが利用できる形で公開された例はこれまでにない。申請者がこれまで構築したモデルを用いた血漿タンパク結合の予測値や学習データを利用し、今までに試みのない学習手法を用いて予測モデルを構築することで、予測精度を向上させることができると期待される。

## ３．研究の方法

■　データ収集及びキュレーションの実施

収集データは予測モデルの精度に大きく影響するため、質の高いデータ収集の実現を目指す。以下に示した公共及び有償のデータベースを対象とする。

・ChEMBL（医薬品と医薬品候補化合物の生物活性低分子の無償データベース）
・PubMed により検索した文献

1. それぞれのデータベースからの情報抽出法の検討を行い、データを抽出する。
2. 抽出したデータに対して、機械的なキュレーションと共にマニュアルキュレーションを実施する
   ① value が空欄、unit が空欄及び h、standard_type で分布容積（Vd）以外を削除
   ② assay_orgamism が Homo sapiens を選択
   ③ description で steady state の記載があるものを選択
   ④ description で健康な成人への単体単回投与の腎クリアランス意外であることが明らかなものを削除（疾病あり、多剤併用、放射線、肥満、やせ型、子供、新生児など）
   ⑤ 単位を L/kg に統一
   ⑥ １化合物に複数データが存在する場合は平均値と SD 値を計算し、SD が平均値の 20%以上の場合は元文献を調査し正しい値を選択する。
3. 異なったデータベース由来のデータは、構造情報に基づいて統合する。

■　予測手法の検討 及びモデル構築

2. 使用する機械学習法（LightGBM、 Random Forest、Support Vector Machine、Artificial Neural Network、k-Nearest Neighbors、Partial Least Square、Adaboost）を検討する。特徴量は Mordred 及び JCompoundMapper を用いて計算した（約 30000 個）。バラツキが０に近い特徴量及び相関が 0.9 以上特徴量は削除し、Boruta アルゴリズムを用いて特徴量の選択を行う。ハイパーパラメータチューニングは Optuna を用いて実施する。
3. 識別すべき特徴を自動的に学習できるディープラーニングの使用を検討する。精度を高

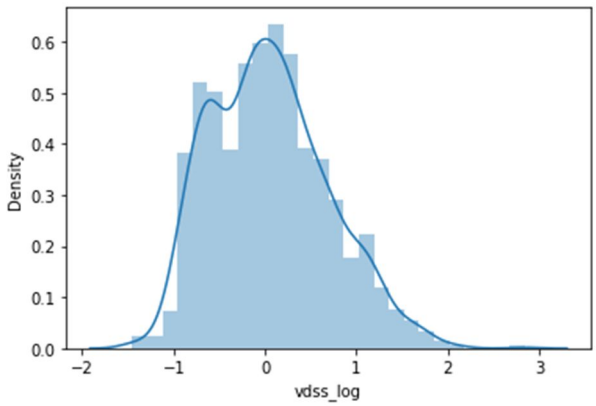
めるためには正確で大量のデータが必要であるため、収集できる分布容積のデータに応じて適正を検討する。

4. 血漿タンパク結合率は分布容積に影響を及ぼすパラメータであることが知られているため、血漿タンパク結合率と分布容積のベクトルやラベルの共通定義域を用いて精度の向上を図る。

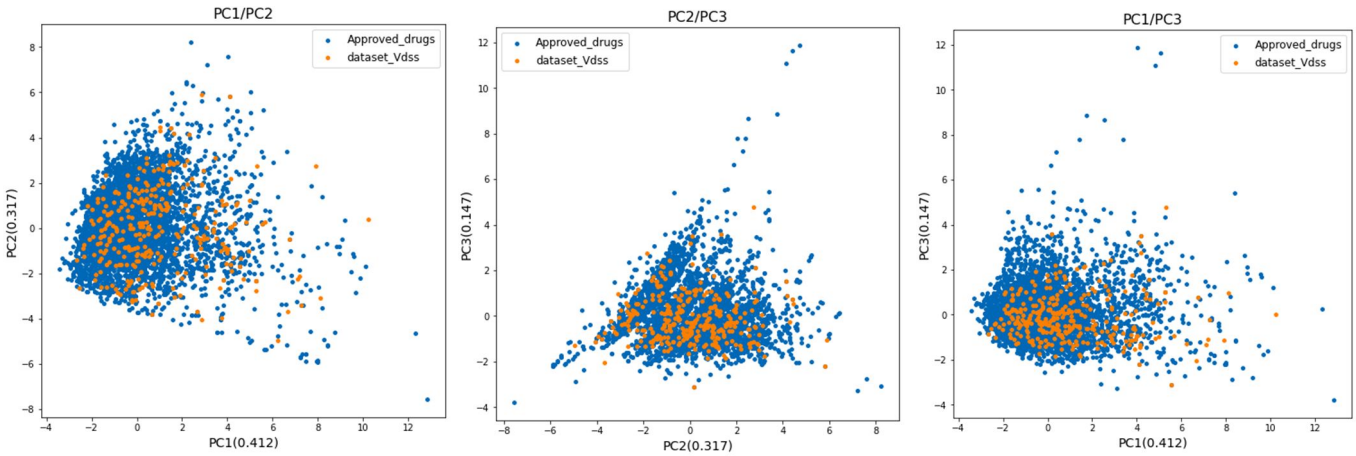
## 4．研究成果

■ データセットの構築

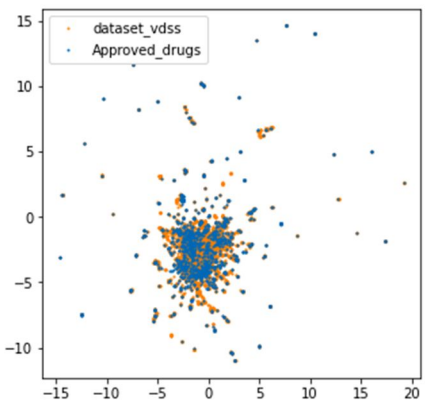
まずは ChEMBL_23 から抽出した 13,133 件のデータについて以下の手順でキュレーションを実施し、925 化合物の分布容積データを収集した。追加データ収集として、ChEMBL_28 より差分データ及び Lomberto et al のデータを収集した。差分データは molregno により、Lomberto のデータは構造情報より重複を調査し、重複化合物については統合した。最終的に 1314 化合物のデータセットを構築した。データセットの分布は以下に示す。




並行して、データセットのケミカルスペースを示すために、5858 の薬として市販されている化合物及び 1314 化合物のデータセットにおける PCA 解析を実施した。使用したのは MW,TopoPSA,nHBAcc, nHBDon,nRot,nAromAtom,nAromBond,SLogP の 8 個の特徴量で、一般的に構造展開の際に重要視されるパラメータを選択した。Principal component (PC)1,2 及び 3 により 9 割近い累積寄与率を示した。




構造情報（ECFP4）の示すケミカルスペースについても umap による解析を実施した。




２つの解析により、今回構築したデータセットは、薬として市販されている化合物と類似し

た特徴を持つことが明らかとなった。

■ 予測手法の検討

LightGBM、Random Forest、Support Vector Machine、Artificial Neural Network、k-Nearest Neighbors、Partial Least Square、Adaboost それぞれの機械学習法により予測モデルを構築した。ベースモデルとして最も精度の高かった LightGBM を用いることとした（$R^2$=0.538, 5-fold cross validation）。

また、ディープラーニングの使用を検討したが、データセット数が1314であったことから、万単位のデータがあることで性能を発揮するディープラーニングの使用は見送ることとした。

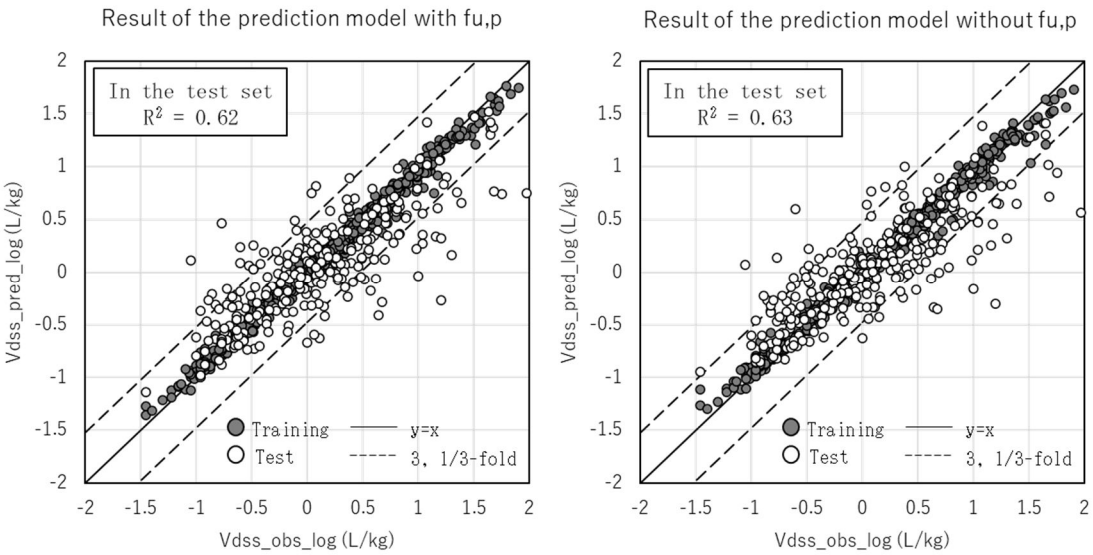
■ 血漿タンパク結合率の追加による精度検証

LightGBM を用いて血漿タンパク結合率を用いた予測の精度検証を実施した。分割バイアスを最小限にすることを目的として、Training set と Test set を 8:2 の割合で 10 パターンに分割し、予測血漿タンパク結合率（fu,p）を特徴量として追加するかしないかで精度比較を行った。実際の予測精度を以下に示す。fu,p ありなしで対応のあるサンプルの t 検定（Paired-samples t-test）を実施し、有意な差がないことが確認された（p>0.1）。

| | With fu,p | | Without fu,p | |
|---|---|---|---|---|
| | R2 | MSE | R2 | MSE |
| Train | 0.989±0.003 | 0.004±0.001 | 0.989±0.006 | 0.005±0.002 |
| Test | 0.549±0.061 | 0.187± 0.021 | 0.552±0.055 | 0.186±0.020 |

fu,p ありなしそれぞれの場合のベストモデルによる予測結果を以下に示す。fu,p ありの場合は $R^2$=0.62, なしの場合は $R^2$=0.63 であった。Test set においては両方のモデルで 8 割のサンプルが実測値の 3 倍以内に予測できていることから、構造情報のみから高精度に分布容積を予測可能なモデルであると判断した。

全て無償のソフトを用いて構造情報のみから高精度に分布容積を予測可能なモデルを構築されたことにより、創薬の効率化に貢献することが可能となると考えられる。



５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

| |
|---|
| １．発表者名<br>Reiko Watanabe |
| ２．発表標題<br>Development of in silico prediction models to evaluate pharmacokinetic profiles using chemical structure information |
| ３．学会等名<br>第48回日本毒性学会学術年会 |
| ４．発表年<br>2021年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|