

博 士 論 文

液体クロマトグラフィー質量分析法における  
イオン化効率の機械学習による予測に関する研究と  
医薬品原薬中の遺伝毒性不純物分析への応用

本論文は静岡県立大学薬食生命科学総合学府薬学研究院博士論文である

2021 年 3 月

静岡県立大学薬食生命科学総合学府  
薬学研究院 薬科学専攻  
生体機能分子分析学講座

宮本 浩平

Studies on Machine Learning-Guided Prediction of  
Liquid Chromatography-Mass Spectrometry Ionization Efficiency  
and Application for Genotoxic Impurities in Drug Substance

March 2021

Laboratory of Analytical and Bioanalytical Chemistry  
Graduate Division of Pharmaceutical Sciences

University of Shizuoka

Kohei Miyamoto

# 目次

略語表及び代数記号表 .....	I
緒言 .....	1
第 1 章 探索的データ分析と LC/MS パラメーターを用いたイオン化効率予測モデルの開発 .....	4
第 1 節 探索的データ分析 .....	4
第 2 節 LC/MS パラメーターによるイオン化効率の重回帰分析 .....	10
第 3 節 RIE の導入と LC/MS パラメーターによる化合物横断的イオン化効率の予測 .....	14
第 4 節 小括 .....	18
第 2 章 機械学習を用いた LC/MS パラメーター及び化合物の物理化学的特性によるイオン化効率の 予測モデルの確立 .....	19
第 1 節 分子記述子への変換と遺伝的アルゴリズムを用いた変数選択 .....	19
第 2 節 LC/MS パラメーター及び分子記述子によるイオン化効率予測モデルの開発 .....	22
第 3 節 小括 .....	24
第 3 章 イオン化効率予測モデルの医薬品中遺伝毒性不純物への適用 .....	25
第 1 節 GTI のイオン化効率予測 .....	25
第 2 節 HILIC モードを使用した医薬品原薬中遺伝毒性不純物の分離 .....	27
第 3 節 HILIC モードで分離した医薬品原薬中遺伝毒性不純物のイオン化効率予測 .....	30
第 4 節 小括 .....	33
総括 .....	34
実験の部 .....	36
試薬 .....	36
装置 .....	36
ソフトウェア .....	37
R パッケージ .....	37

第1章第1節の実験操作 .....	37
第1章第2節の実験操作 .....	38
第1章第3節の実験操作 .....	39
第2章第1節の実験操作 .....	40
第2章第2節の実験操作 .....	42
第3章第1節の実験操作 .....	46
第3章第2節の実験操作 .....	49
第3章第3節の実験操作 .....	49
<b>謝辞 .....</b>	<b>53</b>
<b>引用文献 .....</b>	<b>54</b>

## 略語表及び代数記号表

(アルファベット順)

略語	英語名称	邦語名称
APCI	atmospheric pressure chemical ionization	大気圧化学イオン化
DNA	deoxyribonucleic acid	デオキシリボ核酸
EDA	exploratory data analysis	探索的データ分析
EMA	european medicines agency	欧州医薬品庁
ESI	electrospray ionization	エレクトロスプレーイオン化
FDA	U.S. food and drug administration	アメリカ食品医薬品局
FIA	flow injection analysis	フローインジェクション分析法
GA	genetic algorithm	遺伝的アルゴリズム
GTI	genotoxic impurity	遺伝毒性不純物
HILIC	hydrophilic interaction liquid chromatography	親水性相互作用クロマトグラフィー
HPLC	high performance liquid chromatography	高速液体クロマトグラフィー
ICH	international council for harmonisation of technical requirements for pharmaceuticals for human use	医薬品規制調和国際会議
MS	mass spectrometry	質量分析法
PLS	partial least squares	部分的最小二乗法
QSAR	quantitative structure-activity relationship	定量的構造活性相関
QSPR	quantitative structure-property relationship	定量的構造物性相関
RIE	relative ionization efficiency	相対的イオン化効率
RMSE	root mean square error	平均平方二乗誤差
SVM	support vector machine	サポートベクターマシン
TFA	trifluoroacetic acid	トリフルオロ酢酸
UV-Vis	ultraviolet-visible	紫外可視

## 緒言

遺伝毒性とは、外来性の化学物質や物理化学的要因、もしくは内因性の生理的要因などによりデオキシリボ核酸(Deoxyribonucleic acid: DNA)や染色体、あるいはそれらと関連するタンパク質に変化を与え、癌など個体に悪影響をもたらす性質をいう [1] [2]。欧州医薬品庁 (European medicines agency: EMA) [3]や、アメリカ食品医薬品局 (U.S. food and drug administration: FDA) [4]は医薬品中の遺伝毒性不純物 (Genotoxic impurity: GTI) 管理の重要性を強調したガイドラインを発行し、その制限と管理を義務付けてきた。更に、医薬品規制調和国際会議 (International council for harmonisation of technical requirements for pharmaceuticals for human use: ICH) M7 ガイドライン “Assessment and Control of DNA Reactive (mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk”[5]には、医薬品中の GTI の評価と管理の枠組みが提示されている。このように、GTI の制限と管理に対し規制当局は高い関心を示している。

医薬品中の不純物量は安全性が確保された量まで下げる必要があり、特に GTI はわずかな量でも DNA 損傷を直接引き起こす可能性があるため、高感度な定量法は特に重要になる[6]。加えて、医薬品の品質試験で信頼できる正確なデータを取得することは、製品の安全性を確保するための鍵となる。品質試験の現場で汎用される液体クロマトグラフィー紫外可視吸光光度法 (Liquid chromatography and ultraviolet-visible spectrophotometry: LC/UV-Vis) に比べ、液体クロマトグラフィー質量分析法 (Liquid chromatography-mass spectrometry: LC/MS) は、より高い特異性と感度が必要な定量測定に適した手法であり、医薬品中 GTI の定量において強力なツールとして期待される。

しかしながら、LC/MS で汎用されるエレクトロスプレーイオン化 (Electrospray ionization: ESI) における分析対象物のイオン化効率の変動は大きな問題であり、LC/MS 条件の最適化は煩雑で時間がかかることが多い[7]。GTI の測定は、規制当局からの要請により迅速な分析法開発を求められるケースがあり、LC/MS 分析法開発のニーズは高い。化合物のイオン化効率は、LC/MS の条件パラメーター、化合物特性などの要因によって影響を受けることが知られている。過去にも種々の検討がなされ、Liigand 等 [8]は、ESI に使用されるアセトニトリル含有量と水系移動相のモディファイヤー (添加剤) を変更することにより、化合物のイオン化効率の変化を報告している。Kiontke 等 [9]は、さまざまな pH 条件下での LC/MS 信号強度と測定対象の分子サイズ、揮発性および極性との間の相関を報告している。更に、Caetano 等 [10]は、統計的手法を利用して、より効果的なイオン化法を選択するために、化合物の分子記述子から ESI および大気圧化学イオン化 (Atmospheric pressure chemical ionization: APCI) の応答を予測したが、その研究ではイオン化法と分子記述子との相関に注目するため LC/MS の条件パラメーターは固定されていた。化合物のイオン化効率に関する多くの研究から ESI のイオン化メカニズムは複雑であり、多くの要因の影響を

受けることが分かってきた。しかしながら、LC/MS パラメーターと化合物特性を同時に考慮したイオン化効率予測に関する研究報告は調査した限りこれまで無かった。

以上を踏まえ、本研究では医薬品中 GTI の効率的な高感度分析法開発を達成し、医薬品開発を効率化することを最終目標とした簡便かつ迅速な LC/MS イオン化効率予測法の開発を目的とした。

多種多様な低分子化合物全般を対象とした LC/MS イオン化効率予測は極めて困難なため、本研究では医薬品中 GTI の化学的特徴に焦点を当てることとした。医薬品中 GTI は、原料、中間体、反応試薬、副生成物の形で混入するリスクが高く、比較的分子量が小さくシンプルな構造を有する。また、医薬品原薬の一般的な官能基、例えばアミノ基やカルボニル基も不純物構造に含まれている場合が多い。これら医薬品中 GTI の化学的特徴を踏まえることで、イオン化効率予測法の構築を目指した。また、ハイエンドの LC/MS 機器は高感度の分析を実現できるが、購入コスト、高価なメンテナンス、および高度な操作スキルが求められ、日常的な医薬品製造における品質試験環境での使用は現実的ではない[12]。したがって、本研究では、実運用を考慮し、ユーザーフレンドリーで比較的安価なシングル四重極 LC/MS を使用することとした。

本研究では、簡便かつ迅速な LC/MS イオン化効率予測を実現するため、イオン化効率と LC/MS パラメーターおよび化合物の物理化学的特性の数学的関係性を定義した定量的構造物性相関 (Quantitative structure-property relationship: QSPR) モデルを機械学習で開発した。新薬の開発においてここ数十年にわたって実績があり、よく知られた定量的構造活性相関 (Quantitative structure-activity relationship: QSAR) [13]は化学物質の構造と生物学的な活性との間に成り立つ量的関係を指す、一方で QSPR は、化学構造と物理化学的性質との間に成り立つ量的関係のことを指す[13]。分析研究でも注目を集めている手法[14]で、新しい化合物のイオン化効率を予測するための強力な統計ツールになりうる。また、機械学習とは、コンピューターがデータから反復的に学習し、そこに潜むパターンを見つけ出すことで、大量のデータから自動的にアルゴリズムを構築する。そして、見つけたパターンを新しいデータに適用することで新しいデータの予測を行うことが出来る[11]。これらの技術を組み合わせた新たな簡便かつ迅速な LC/MS イオン化効率予測手法の開発について3つの章に分けて論じる。

第1章では、機械学習を用いたイオン化効率予測モデルの開発に先立ち、異なる LC/MS パラメーターの組み合わせで作成した条件を用いて10種の化合物でデータ取得を実施した。各化合物のデータからイオン化の特徴について解析を行い、LC/MS パラメーターを用いたピーク面積の予測モデルの開発について検討を行った。

第2章では、LC/MS パラメーターに加え、化合物特性を組み込んだ予測モデルの開発を行った。化学構造を分子記述子という物理化学的特性を表す数値に変換し、そこからイオン化効率に影響がある分子記述子の選択、複数の機械学習アルゴリズムから最良のアルゴリズムのスクリーニン

グ、選択したアルゴリズムの最適化を実施し、機械学習モデルを構築した。そして最後に、検証データを用いたモデルの予測精度の評価を行った。

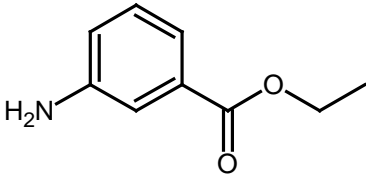
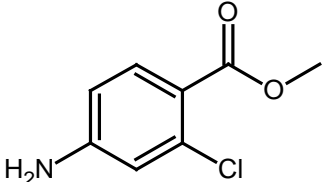
第 3 章では、本手法の有用性を評価するために、学習データにない実際の遺伝毒性物質のイオン化効率を予測した。更に、医薬品中 GTI の分析法の開発を想定し、遺伝毒性物質を医薬品原薬に添加し、親水性相互作用クロマトグラフィー (HILIC) で分離後のピークについて、イオン化効率の予測を検討した。

# 第1章 探索的データ分析と LC/MS パラメーターを用いたイオン化効率予測モデルの開発

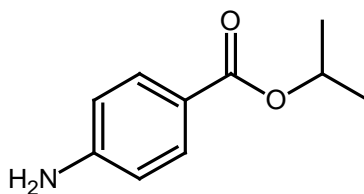
## 第1節 探索的データ分析

本研究で使用した化合物の化学構造を Table 1 - 1 に示す。これらは低分子化合物の網羅的解析を意図したものではなく、医薬品中の GTI に焦点を当て選択した化合物である。医薬品中 GTI は、原料、中間体、反応試薬、副生成物の形で混入するリスクが高く、比較的分子量が小さくシンプルな構造を有する。また、医薬品原薬の一般的な官能基、例えばアミノ基やカルボニル基も不純物構造に含まれていることが多い。これら医薬品中 GTI の化学的特徴を踏まえ、モデル化合物としてのアミノ安息香酸の類縁体を選択した。また、変異原性化合物である *N*-(3-amino-4-methoxyphenyl)acetamide [15]及び *N*-nitrosodipropylamine [16]も加えた。種々ある LC/MS パラメーターの中からベンダーの技術資料やこれまでの経験を基に移動相への添加剤（モディファイヤー）の種類、有機溶媒比率、プローブ温度、コーン電圧及びキャピラリー電圧を検討対象として選択した。Table 1 - 2 にその設定値を示す。1つの化合物に対し、これら異なる LC/MS パラメーターの組み合わせで実験計画法により作成した 135 の条件でピーク面積のデータ取得を実施した。窒素含有化合物の場合、プロトン化分子が正イオン検出における主要なイオン付加体であり、他のイオン付加体は観察されたとしてもごくわずかな量しか検出されない[9]ため一連の化合物の[M + H]<sup>+</sup>応答性を解析に用いた。

**Table 1 - 1** 本研究に用いた化合物の一般名、化学構造、CAS No.、計算された p*K*<sub>a</sub>

Name	Chemical structure	CAS No.	p <i>K</i> <sub>a</sub> <sup>1)</sup>
Ethyl 3-aminobenzoate		582-33-2	3.5
Methyl 4-amino-2-chlorobenzoate		46004-37-9	1.4

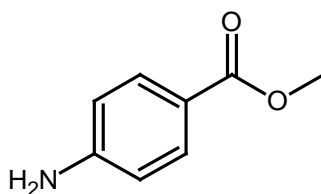
Isopropyl 4-aminobenzoate



18144-43-9

2.6

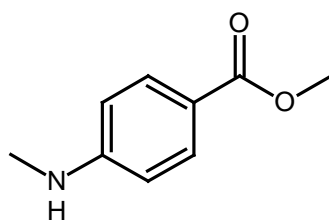
Methyl 4-aminobenzoate



619-45-4

2.5

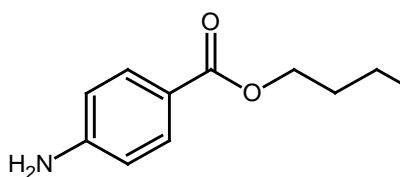
Methyl 4-(methylamino)benzoate



18358-63-9

2.3

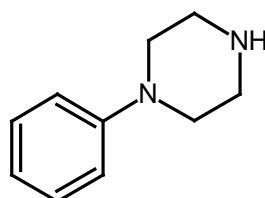
Butyl 4-aminobenzoate



94-25-7

2.4

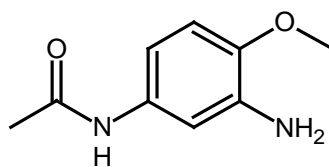
1-Phenylpiperazine



92-54-6

3.7

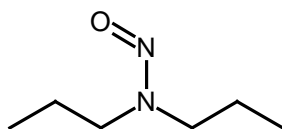
9.0

*N*-(3-amino-4-methoxyphenyl)acetamide

6375-47-9

4.1

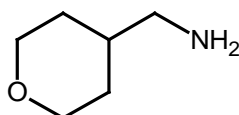
*N*-Nitrosodipropylamine



621-64-7

NA

4-Aminomethyl-  
tetrahydropyran



130290-79-8

10.2

NA: not applicable

1) Calculated  $pK_a$  values

**Table 1 - 2** 検討した LC/MS パラメーターおよびその設定値

LC/MS パラメーター	設定値		
モディファイヤーの種類	ギ酸	TFA	酢酸アンモニウム
有機溶媒比率 (v/v %)	40	60	80
プローブ温度 (°C)	300	450	600
コーン電圧 (V)	11	15	19
キャピラリー電圧 (kV)	0.8	1.0	1.2

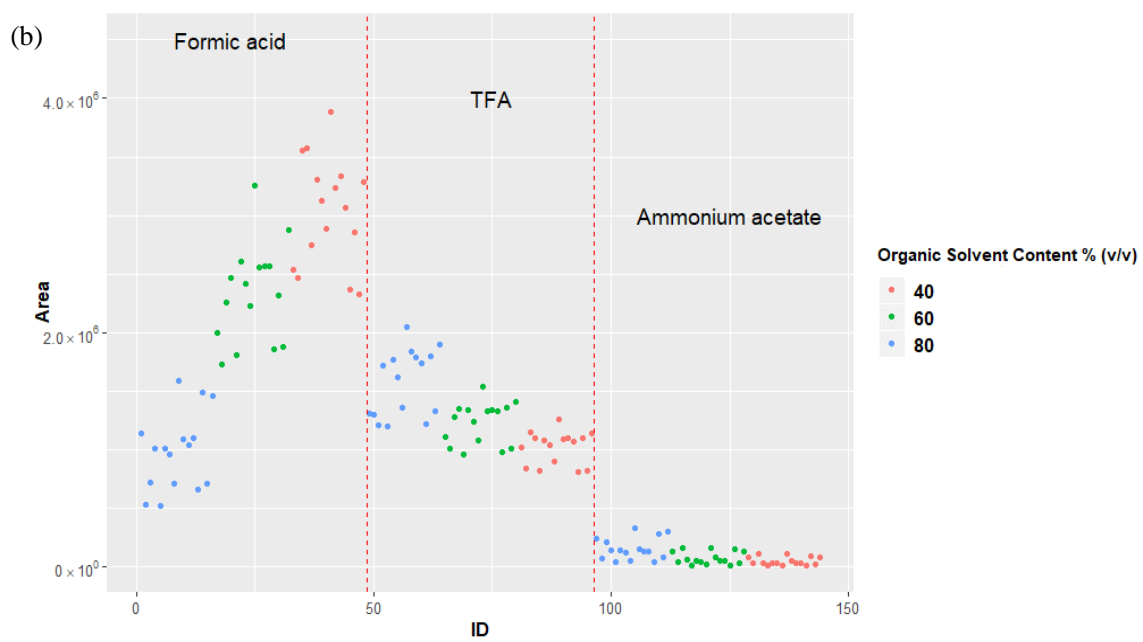
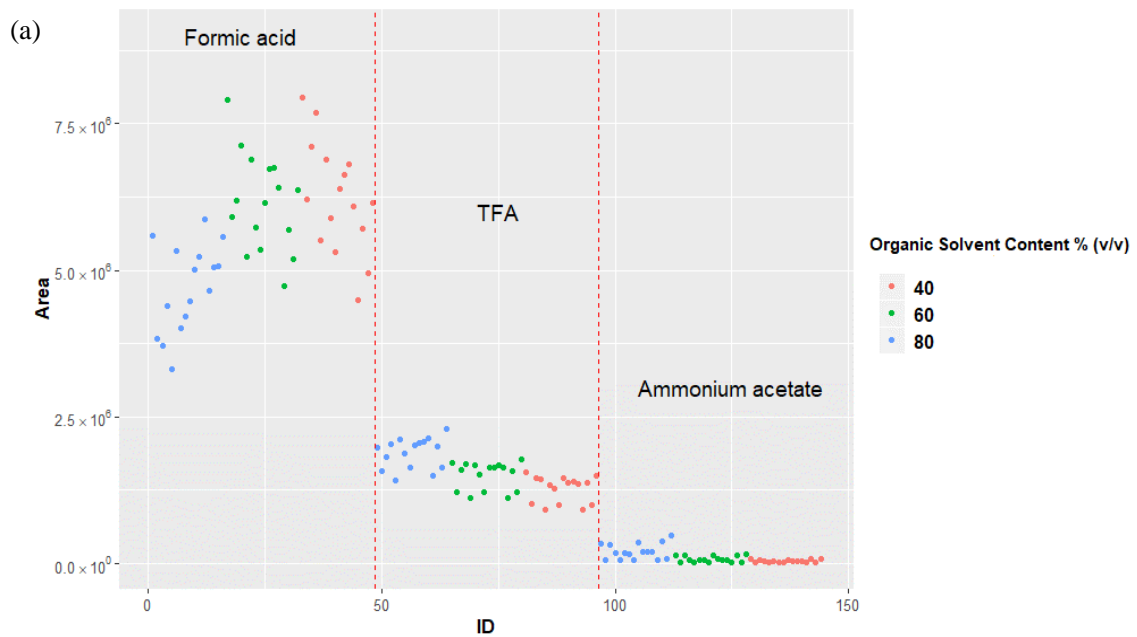
イオン化効率予測モデルの開発に先立ち、得られたデータからイオン化傾向の特徴について解析した。こうした一連の作業は探索的データ分析 (Exploratory Data Analysis: EDA) と呼ばれ、データの特徴を把握することを目的としたデータサイエンスの最初のステップとされている[17]。本章では 10 化合物のうち 3 化合物を代表として論じ、他は割愛した。まず、ethyl 3-aminobenzoate の結果についてモディファイヤーの種類及び有機溶媒比率の順に並べ、MS のピーク面積をプロットしたグラフを Fig. 1 - 1 a に示す。モディファイヤーの種類がピーク面積に最も強い影響を与え、酢酸アンモニウム、TFA、ギ酸の順に面積が大きくなった。Advanced Chemistry Development (ACD / Labs) ソフトウェア V11.02 を使用して ethyl 3-aminobenzoate の共役酸の予測  $pK_a$  は 3.5 であった。ギ酸溶液中では、ethyl 3-aminobenzoate はイオン化された状態だが、酢酸アンモニウム溶液中では分子形のままで存在すると考えられる。分析対象物がすでに液相でイオン化されている場合、高いイオン化効率を得られる[18]。一方で、TFA 移動相でも ethyl 3-aminobenzoate はイオン化されるが、ピーク面積はギ酸移動相のピーク面積よりも小さかった。TFA アニオンが正電荷を帯びた官能基とイオン対になり、ポジティブモードでのイオン化効率を抑制することはよく知られている

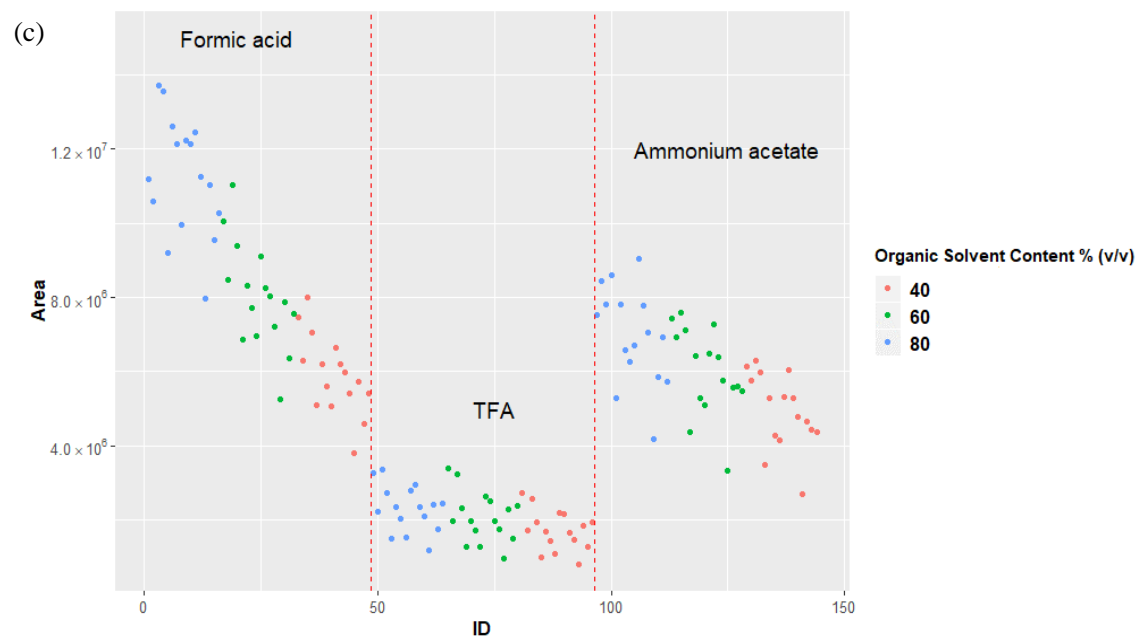
[19] [20]。また、TFA をモディファイヤーとして使用した場合、アセトニトリル含有量が多い条件でピーク面積が増加する傾向が認められた。有機溶媒含有量の増加に伴うピーク面積の増加は、液滴のより効率的な脱溶媒和によって説明でき、液滴がレイリー限界に早く到達し、分裂した液滴を生成しやすくなるため感度が向上すると考えられる[21]。一方、ギ酸をモディファイヤーとして使用し、アセトニトリル含有量が高い条件で ethyl 3-aminobenzoate のイオン化は抑制された。アセトニトリル含有量が高くなると水系移動相に添加されたギ酸濃度が下がるため、モディファイヤーの効果と有機含有量の効果はトレードオフの関係にある。LC/MS の有機溶媒移動相としてはアセトニトリルとメタノールが一般的に使用される。本研究では、HILIC モードに適用することを見据え、固定相と分析対象物間のより強い親水性相互作用が期待できるアセトニトリルを選択した[22]。メタノールはプロトン性溶媒であるため、イオン化時のプロトン供与効率が高まりイオン化を助長する効果がある[23]。一方で、溶媒の粘度が上がリイオン化を抑制することも知られており[24]、イオン化効率はアセトニトリルと異なる傾向となる可能性がある。

次に methyl 4-aminobenzoate の例を示す (Fig. 1 - 1 b)。予測  $pK_a$  は 2.5 であり、分析対象物はギ酸溶液中でイオン化され、酢酸アンモニウム溶液中ではイオン化されていない。最大のピーク面積は、低い有機溶媒含有量を使用したギ酸条件で観察された。一方、ギ酸を使用した場合の有機溶媒含有量の影響は、TFA を使用した場合よりも大きく、これらのパラメーターがイオン化効率に対し交互作用があることを示している。また、高い有機溶媒含有量を使用した TFA 条件のピーク面積は、高い有機溶媒含有量を使用したギ酸条件のピーク面積と同等かそれ以上であった。前述のように、ESI における TFA によるイオン化抑制はよく知られているが、必ずしも全ての化合物及び全ての条件下で適用できるわけではないことを示している。Methyl 4-aminobenzoate の化学構造は ethyl 3-aminobenzoate (Fig. 1 - 1 a) と類似しているにもかかわらず、上述のようにイオン化の傾向が異なることは興味深い結果であった。

1-Phenylpiperazine は、さらに異なるイオン化の傾向を示した (Fig. 1 - 1 c)。TFA よりも酢酸アンモニウムを使用した条件でピーク面積が大きかった。1-Phenylpiperazine は、予測  $pK_a$  が 3.7 および 9.0 であるため、酢酸アンモニウム中でもイオン化されている。酢酸アンモニウムと TFA のいずれの移動相もイオン化を促進したが、TFA は前述のイオンペア生成効果によりイオン化が抑制されたため、このような結果になったと推測される。ギ酸を使用すると、低有機溶媒含有量 (40%) よりも高有機溶媒含有量 (80%) を使用した方が大きなピーク面積が観察された。この傾向は、ethyl 3-aminobenzoate や methyl 4-aminobenzoate のイオン化傾向とは異なる。

これらの検討から、イオン化効率に対する LC および MS パラメーターの影響は非常に複雑であり、化合物特性に依存することが推測され、ここまで考察してきた有機溶媒によるイオン化効率の向上や TFA による抑制など単純な理論に基づいて予測することは困難と考えられる。

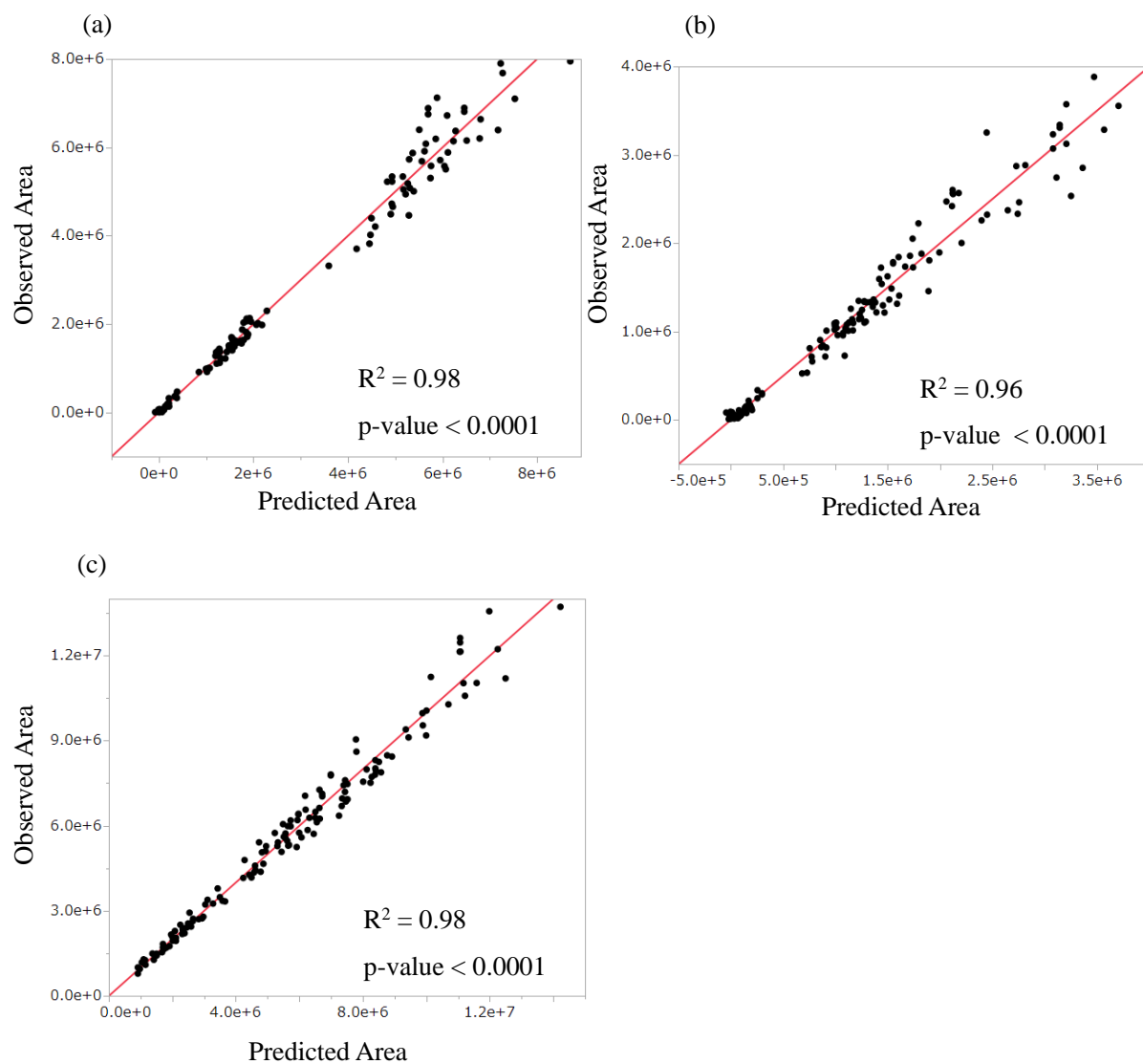




**Fig. 1 - 1** 異なる LC/MS パラメーターの組み合わせで作成した 135 の条件で得たピーク面積に対するモディファイヤーの種類及び有機溶媒比率の影響 (a) ethyl 3-aminobenzoate、(b) methyl 4-aminobenzoate、(c) 1-phenylpiperazine

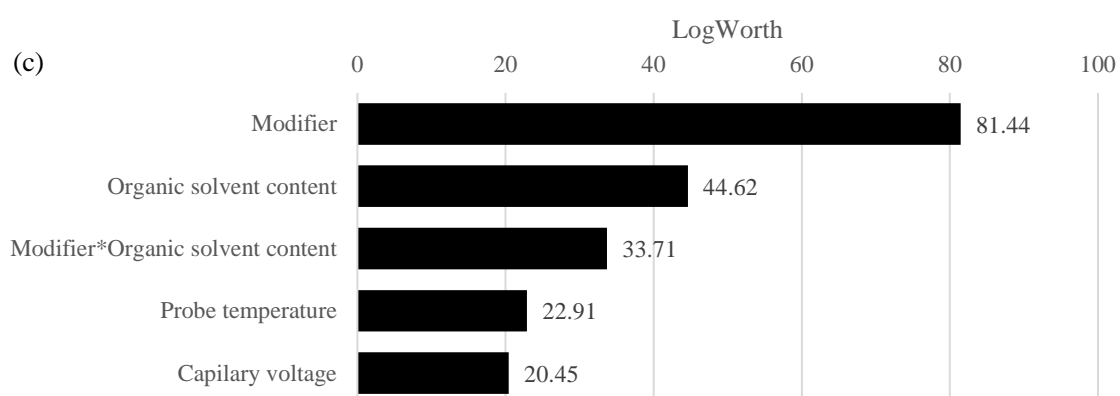
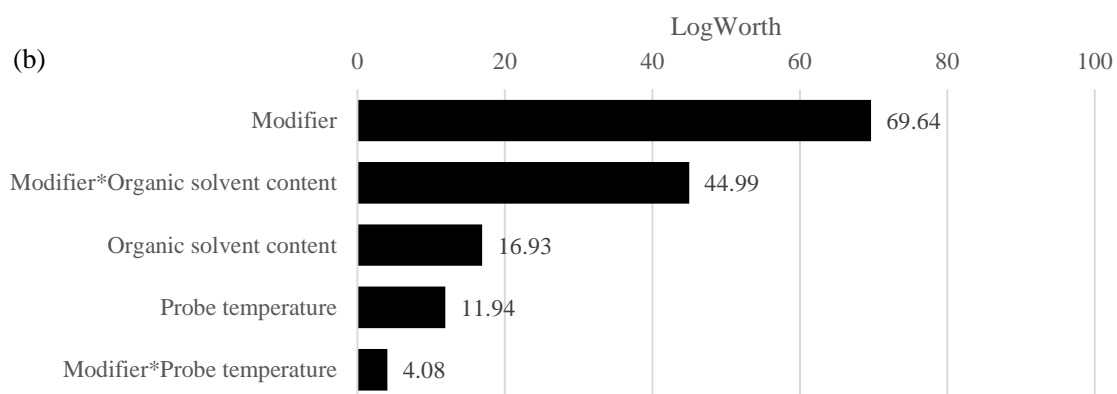
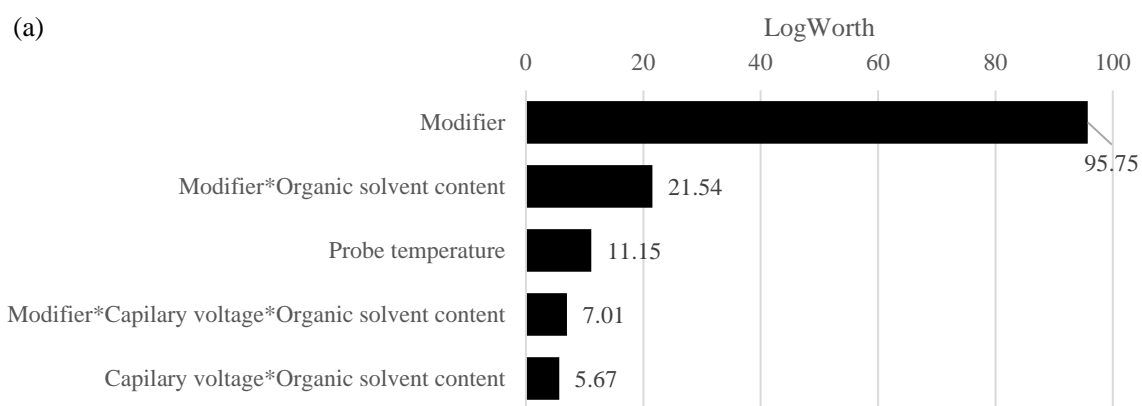
## 第2節 LC/MS パラメーターによるイオン化効率の重回帰分析

前節の検討結果から、イオン化傾向はモディファイヤーの種類や有機溶媒比率などの LC/MS パラメーターの影響を大きく受けることが示された。本節では LC/MS パラメーターからイオン化効率を予測できるか検討した。前節と同じく 10 化合物から代表として 3 つの化合物のデータを用い、各 LC/MS パラメーターを説明変数、ピーク面積を目的変数として重回帰分析を行い、ピーク面積の予測モデルを構築した。この回帰分析から得られた予測ピーク面積と実験的に観察されたピーク面積をプロットしたグラフが Fig. 1 - 2 である。重回帰分析の結果では、各化合物で良好な決定係数 ( $R^2 > 0.96$ ) と p 値 ( $< 0.0001$ ) が示され、LC/MS パラメーターでピーク面積が予測可能であることが示された。特に 1-phenylpiperazine のイオン化効率の特徴は、他の 2 つの化合物と大きく異なることが前節の結果から明らかになったが、本節の重回帰分析は良好な予測結果 ( $R^2 = 0.98$ ) を示し、イオン化効率の特徴に依らず LC/MS パラメーターで予測可能であることが明らかになった。



**Fig. 1 - 2** 重回帰分析による予測ピーク面積と実験で得られたピーク面積の相関 (a) ethyl 3-aminobenzoate、(b) methyl 4-aminobenzoate、(c) 1-phenylpiperazine, 赤線; 回帰直線

次に、どの LC/MS パラメーターがイオン化効率に大きな影響を与えているか統計的な検証を実施した。上位 5 つのパラメーターを Fig. 1 - 3 に示す。パラメーターの影響が大きい場合、関連する p 値は小さくなる。視認性を高めるため  $\text{LogWorth} = (-\log_{10}(\text{p-value}))$  スケールに変換すると、有意性の高いパラメーターの LogWorth 値が大きく、有意でないパラメーターの LogWorth 値は低くなる。同じく 10 化合物から代表として 3 つの化合物のデータを用い解析を実施したところ共通してモディファイヤー、有機溶媒含有量及びプローブ温度がイオン化効率に重要な影響を与えることが明らかになった。これらは、前節で論じた EDA における考察と齟齬ない結果であった。モディファイヤーと有機溶媒含有量の相互作用効果の LogWorth 値が大きいことは、前述したトレードオフの関係を示している。ethyl 3-aminobenzoate では、モディファイヤーの LogWorth 値が約 96 に対し、他のパラメーターの LogWorth 値は相対的に小さい。モディファイヤーの種類がイオン化効率に影響を与える主要因であることが分かる (Fig. 1 - 3 a)。一方、methyl 4-aminobenzoate では、モディファイヤーの LogWorth 値は約 70、モディファイヤーと有機溶媒含有量の相互作用は約 45 であり、イオン化効率はモディファイヤーの種類だけでなく有機溶媒含有量の影響を相対的に大きく受けることが示された (Fig. 1 - 3 b)。この交互作用の影響は、前節の Fig. 1 - 1 b のグラフとも一致する。これらの結果から、イオン化効率に重要な影響を与える LC/MS パラメーターはある程度共通しているが、その影響度は化合物によって異なることが示唆された。



**Fig. 1 - 3 重回帰モデルにおける各 LC/MS パラメーターの LogWorth 値 (a) ethyl 3-aminobenzoate、(b) methyl 4-aminobenzoate、(c) 1-phenylpiperazine**

### 第3節 RIE の導入と LC/MS パラメーターによる化合物横断的イオン化効率の予測

前節では各化合物において重回帰分析でイオン化効率の予測に成功した。本節では化合物横断的に LC/MS パラメーターを用いてイオン化効率が予測できるか検証した。本研究の目的は LC/MS 分析におけるイオン化効率を予測することであるが、LC/MS の感度は機種のみならず装置の状態にも依存するため極めて予測困難である[25]。そこで、相対的なイオン化効率を表現する relative ionization efficiency (RIE)を導入する。RIE の式を以下に示す。

$$RIE = \frac{A_{ex} (experimental\ condition)}{A_{std} (standard\ condition)} \quad (1)$$

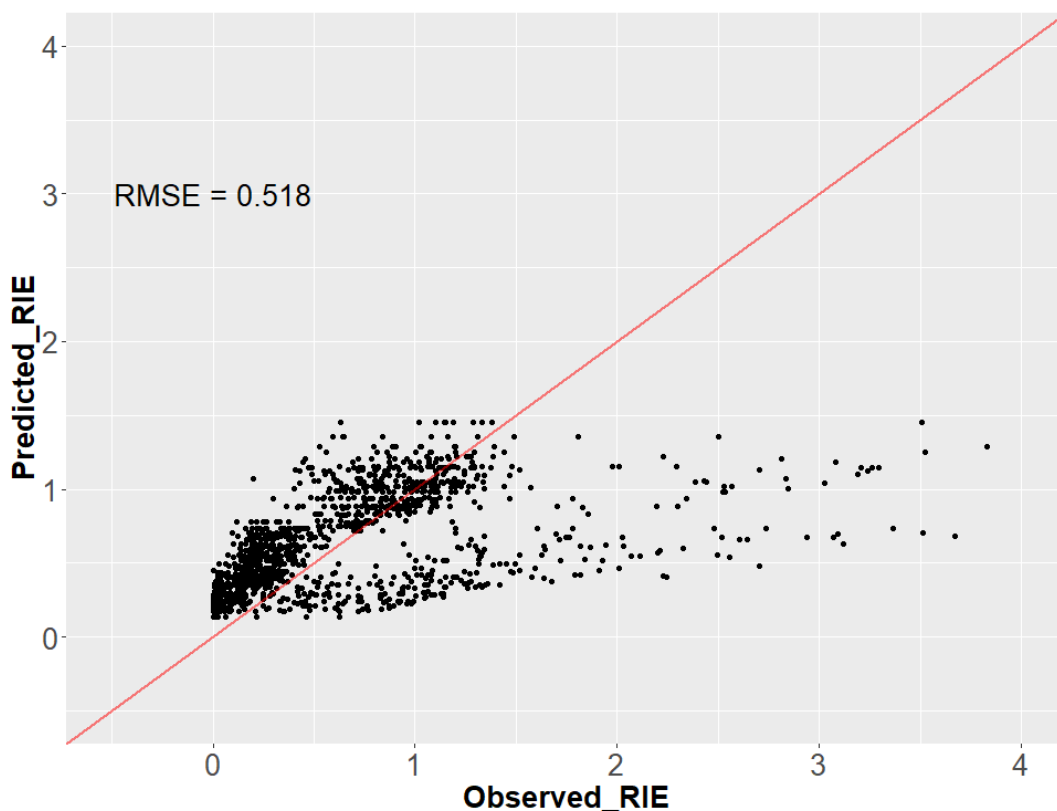
分子にある  $A_{ex} (experimental\ condition)$  は各測定条件において得られたピーク面積を、分母にある  $A_{std} (standard\ condition)$  は標準条件において得られたピーク面積を表す。本検討における標準条件とは、プローブ温度、コーン電圧、キャピラリー電圧は3水準のうち中央値で設定した。前節の検討からモディファイヤー種類と有機溶媒比率はイオン化傾向に大きな影響を与えることが分かっている。標準条件で得られるピーク面積値が小さいとバラつきが大きくなるため、高いピーク面積が期待されるモディファイヤー種類はギ酸、有機溶媒比率は80%で設定した。RIEは標準条件に対してどれだけイオン化効率が向上若しくは低下したかを表現する。RIEを導入することで、LC/MS の検出感度の日間変動や、サンプル調製濃度のバラつきが補正でき、複数化合物の比較が可能となる。methyl 4-aminobenzoate の異なる6つの測定条件における RIE の日間再現性を確認した結果を Table 1-3 に示す。感度不足となった酢酸アンモニウム条件（参考：Fig. 1-1 b）を除き、RSD は 4.7%~11.9% であった。医薬品品質試験のバリデーションでは、最適化された試験条件を用いて RSD が 10% 未満であれば十分な再現精度を有すると考える[26][27]。LC/MS の高感度分析法開発において最初の中心条件を選択する上で許容できる再現性を有し、RIE が機能していることを確認できた。

**Table 1 - 3** 6つの異なる条件を用いた methyl 4-aminobenzoate における RIE の日間再現性

Condition			RIE				
Condition No.	Modifier type	Organic solvent content (%)	Day 1	Day 2	Day 3	SD	RSD (%)
1	Formic acid	60	1.17	1.11	1.22	0.06	4.7
2	Formic acid	40	1.29	1.16	1.19	0.07	5.6
3	TFA	60	0.45	0.45	0.55	0.06	11.9
4	TFA	40	0.36	0.35	0.43	0.04	11.5
5	Ammonium acetate	80	0.03	0.03	0.04	0.01	17.3
6	Ammonium acetate	40	0.01	0.01	0.01	0.00	0.0

※標準条件：モディファイヤー種類：ギ酸、プローブ温度：450 °C、コーン電圧：15 V、キャピラリー電圧：1.0 kV、有機溶媒比率：80%

RIE の導入により化合物横断的な検証が可能になったため、全 10 化合物のデータを用いた重回帰分析を実施した。異なる LC/MS パラメーターの組み合わせで作成した 135 の条件を用いて 10 化合物それぞれで取得したピーク面積を RIE に変換した。この RIE を目的変数、LC/MS パラメーターを説明変数として重回帰分析を行った。横軸に実験的に観察された RIE を、縦軸に得られた予測 RIE をとったプロットを Fig. 1 - 4 に示す。赤線は傾き 1、切片 0 の直線で、点が赤線に近いほどモデルの予測精度が高いことを表す。



**Fig. 1 - 4** 10 化合物を用いた RIE の実験値と予測値の相関、赤線；傾き 1、切片 0 の直線

評価指標には一般的に回帰モデルの評価に用いられている平均平方二乗誤差 (RMSE)を採用した。RMSE の式を以下に示す。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{obs} - y_i^{prd})^2} \quad (2)$$

$y_i^{obs}$  は実験的に観察されたピーク面積を、 $y_i^{prd}$  は回帰モデルにより予測されたピーク面積を、 $n$  はデータの数を指す。RMSE はその回帰モデルの誤差の平均を表現し、次元は測定値と同等になるため直感的にその程度を理解できるのが特徴である。

実験的に観察された RIE と重回帰モデルにより予測された RIE から得られた RMSE は 0.518 であった。この結果は予測値が平均的に約 0.5 程度の誤差を含むことを意味している。RIE の導入により各化合物間の相対比較を可能にしたものの、LC/MS パラメーターを用いただけの予測では大きな誤差が認められた。この結果から、LC/MS パラメーターがイオン化効率に与える影響は化合物により異なり、その化合物特性を予測モデルに組み込むことが必要と考えられた。

## 第4節 小括

本章では初めにイオン化効率における特徴の把握と、LC/MS パラメーターによる化合物横断的イオン化効率の予測を目的に検証を行った。

本検討は低分子化合物の網羅的解析を意図したものではなく、医薬品中の GTI に焦点を当てた 10 種の化合物を用いた。医薬品中 GTI は、原料、中間体、反応試薬、副生成物の形で混入するリスクが高く、比較的分子量が小さくシンプルな構造を有する。また、医薬品原薬の一般的な官能基、例えばアミノ基やカルボニル基も不純物構造に含まれている場合が多い。これら医薬品中 GTI の化学的特徴を踏まえ選択した 10 種の化合物を用いて、異なる LC/MS パラメーターの組み合わせで作成した 135 の条件でデータを取得した。そして、各化合物内において LC/MS パラメーターを説明変数、ピーク面積を目的変数として重回帰分析を行った。本章では 10 化合物のうち代表として 3 化合物のデータを示したが、良好な予測結果を示し LC/MS パラメーターでピーク面積が予測可能であった。イオン化効率に重要な影響を与える変数としてはモディファイヤー、有機溶媒含有量及びプローブ温度が共通して挙がってくる一方で、その影響度は化合物によって異なることが明らかになった。

また、10 種の化合物データを用いて、各化合物間の相対比較ができるように標準条件で補正された RIE を目的変数とし、LC/MS パラメーターを説明変数として重回帰分析を行った。その結果、RIE の導入により各化合物間の相対比較を可能にしたものの、LC/MS パラメーターを用いただけの予測では大きな誤差が認められた。この結果から、LC/MS パラメーターがイオン化効率に与える影響は化合物により異なり、その化合物特性を予測モデルに組み込むことが必要と考えられた。

## 第2章 機械学習を用いた LC/MS パラメーター及び化合物の物理化学的特性によるイオン化効率の予測モデルの確立

### 第1節 分子記述子への変換と遺伝的アルゴリズムを用いた変数選択

前章において 1 化合物内で LC/MS パラメーターを用いた重回帰分析を行ったところ良好な予測精度を有したが、LC/MS パラメーターがイオン化効率に与える影響は化合物特性により異なるため、化合物横断的な予測は困難であった。本章では化合物特性を組み込んだ予測モデルの開発について検討した。

化合物特性を数学的モデルに組み込むため、化合物の構造から分子記述子に変換し数値化を行った。分子記述子とは分子の物理化学的特性を表す指標のことで構成記述子（原子数、芳香族結合水素数、環数など）、電子的記述子（分極、水素ドナー・アクセプターなど）、構造記述子（原子間距離）など様々な種類がある[28]。分子記述子への変換には、2D 及び 3D の 400 を超える分子記述子が計算可能な統合計算科学ソフトウェアの MOE を使用した。

計算された多数の分子記述子には、イオン化効率とは無関係な記述子も多く含まれている。これら全ての分子記述子を使用して機械学習を行うと、訓練データへの過学習（オーバーフィッティング）が起これ新規データに対する汎化性能が落ちたり、計算が膨大になり時間を要したりなどの欠点がある[29]。そこで、遺伝的アルゴリズム（Genetic algorithm: GA）を使用した変数選択を採用した。

GA は生物の進化を模倣した最適化手法であり、このような変数が多数の問題においても、効率的に精度の高い近似解を導くことができる[30]。GA による変数選択の一般的な流れを Fig. 2 - 1 に示す。①まず最適化を行う変数の列に対し、ランダムに「0」と「1」を附番する。「0」と「1」は対応する説明変数をモデリングに用いるか否かを示している。この 1 つの列を染色体と呼び、N 個の染色体を生成し初期集団を形成する。②各染色体それぞれで「1」に附番された変数のみを用いて回帰モデルを作成し予測精度（適合度）を求める。③その値に応じて親となる染色体が選択され、適合度の低い染色体は淘汰される。④選ばれた染色体は交差や突然変異といった遺伝的操作が施され次世代の染色体を生成する。⑤さらに、次世代の染色体は先ほどと同様に適合度の計算、淘汰、選択、遺伝的操作を経て、再び次の世代の染色体を生成する。この一連の計算を繰り返すことによって、最終的に適合度の高い染色体、つまり予測精度を最大化するような説明変数の組を選び出すことができる。GA の中では上述したように適合度を求める必要がある。本研究ではその方法として、大きなデータセットを短時間で処理できる部分的最小二乗法 (Partial least squares: PLS) という線形重回帰分析を用いて各染色体の適合度を求めた。

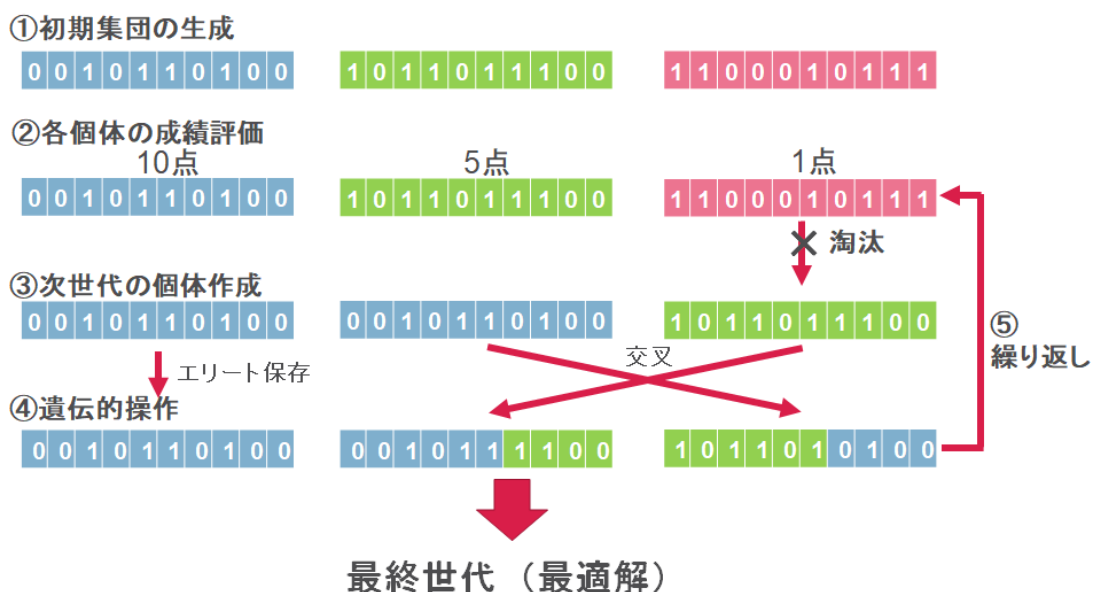


Fig. 2 - 1 GA による変数選択の一般的な流れ

GA によって選択された分子記述子の要約を Table 2 - 1 に示す。選択された記述子の概略を説明すると、A\_aro は化合物の疎水性と平面性に寄与する芳香族原子の数を表し、ASA\_P は極性官能基周辺の水と接触可能な表面積、FASA は全体の表面積に対する正に分極している原子周辺の表面積が占める割合を表す。E\_nb、E\_rele は、位置エネルギーに関する記述子で化合物の 3D コンフォメーションに関連している。mr、SMR\_VSA6 はモル屈折率に関わる記述子で、分子の屈折率は構成する原子団の双極子を反映していると考えられる。Radius は分子を距離行列で表現した際に分子全体の半径から定義される。SlogP\_VSA3 は、分子の logP(o/w)値に寄与する原子のファンデルワールス表面積を表す。Vsurf シリーズは、疎水性および親水性に関連する表面積、体積、および形状記述子を示す。そして、Weight は分子量を表す。

本検討において分子記述子には、極性、溶媒との接触面積や親和性、分子の物理的および化学的な大きさ、分極などに関連した分子記述子を選択された。Hermans 等 [31]は、pK<sub>a</sub>、疎水性、表面張力、および分子体積はすべてイオン化効率に関連していると報告している。そのため、本検討においてこれら物理化学的特性に関連した記述子を選択されたことは妥当と考えている。また、これら多様な記述子が選ばれたことは、イオン化効率のメカニズムが複雑であることも示している。

**Table 2 - 1 GA で選択された分子記述子**

Name	Description	Category
a_aro	Number of aromatic atoms	Atom count and bond count
ASA_P	Water accessible surface area of all polar atoms.	Conformation Dependent Charge Descriptors
E_nb	Value of the potential energy with all bonded terms disabled.	Potential Energy Descriptors
E_rele	Electrostatic interaction energy	Potential Energy Descriptors
FASA	Fractional surface area	Conformation Dependent Charge Descriptors
mr	Molecular refractivity	Physical Properties
radius	Smallest vertex eccentricity in the distance matrix.	Adjacency and Distance Matrix Descriptors
SlogP_VSA3	Bin 3 SlogP (0.00, 0.10)	Subdivided Surface Areas
SMR_VSA6	Molar Refractivity	Subdivided Surface Areas
vsurf_EDmin1	Lowest hydrophobic energy	vsurf_EDmin1, vsurf_EDmin2 distance Surface Area, Volume, and Shape Descriptors
vsurf_W8	Hydrophilic volume	Surface Area, Volume, and Shape Descriptors
Weight	Molecular weight	Physical Properties

## 第2節 LC/MS パラメーター及び分子記述子によるイオン化効率予測モデルの開発

GA によって選択された分子記述子と LC/MS パラメーターの組み合わせで構成されたデータセットを用いて、イオン化効率を予測する機械学習モデルを構築した。機械学習アルゴリズムには多くの種類があり、すべての問題に対し万能なものは存在しない。それぞれ長所と短所があり、その選択にはデータの構造とサイズに大きく依存するため複数のアルゴリズムで予測を行い、最適なものを選択することが有効である。本研究では、予測モデルを開発するために、線形および非線形の 8 つの回帰モデルでスクリーニングした。各機械学習アルゴリズムで 5 分割交差検定を 3 回繰り返して RMSE を計算させた結果を Fig. 2 - 2 に示す。サポートベクターマシン (SVM) が、最も良い RMSE を出し優れた予測性能を示した。SVM は回帰と分類の両方の目的で使用される教師あり学習方法[32]で、モデルの複雑さを表す正則化項と、予測値と実測値のズレがある値を超えたときに誤差として蓄積する損失関数を組み合わせることで汎化性能を高めた計算モデルである。すでに QSPR の領域で利用されているアルゴリズムで、分析科学の分野では、逆相液体クロマトグラフィーでのアミノ酸の保持時間予測[33]や、LC/MS での同位体ピークパターンの研究[34]などで実績がある。しかし、本研究のように LC/MS イオン化効率の予測に使用した例はない。

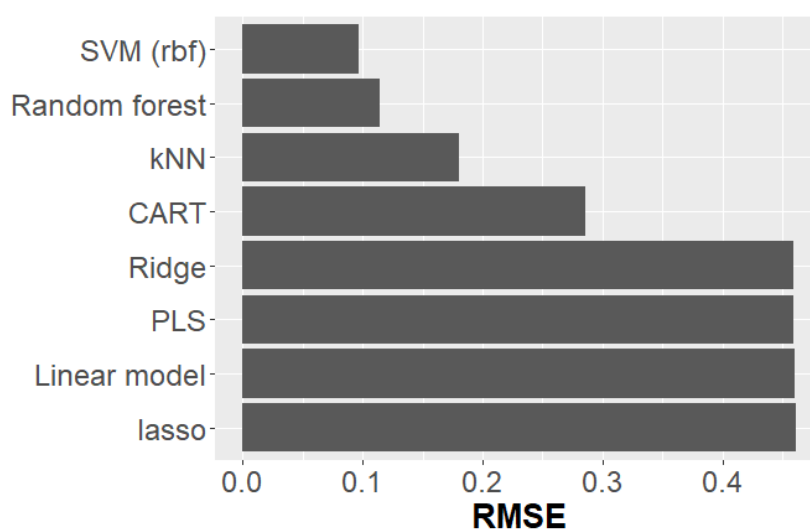


Fig. 2 - 2 機械学習アルゴリズムのスクリーニング結果

SVM を用いて、機械学習モデルの最適化を実施した。第 1 章で記載した通り、10 化合物を用いて異なる LC/MS パラメーターの組み合わせで取得した合計 1428 のデータをランダム化し、8 対 2 の比率で学習データセットと検証データセットの 2 つのグループに分けた。まず、1142 のデータからなる学習データセットを用いて SVM アルゴリズムの最適化し、機械学習モデルを構築した。次に、286 のデータからなる検証データセットを使用して実際に RIE の予測を行い、実測値と予測値から RMSE の計算を行った。その結果、予測された RIE と観測された RIE の RMSE は 0.123、単回帰分析を実施した際の決定係数 ( $R^2$ ) は 0.97、 $p$  値は  $<0.0001$  と優れた予測精度を示した (Fig. 2-3)。前章第 3 節で実施した LC/MS パラメーターのみで RIE を予測した結果 (Fig. 1-4) と比較して、RMSE は 0.518 から大幅に低下し、改善が認められた。これは、化合物の物理化学的特性がイオン化効率に大きな影響を及ぼし、かつ、選択された分子記述子のセットが高い予測精度の実現に不可欠であることを明示している。

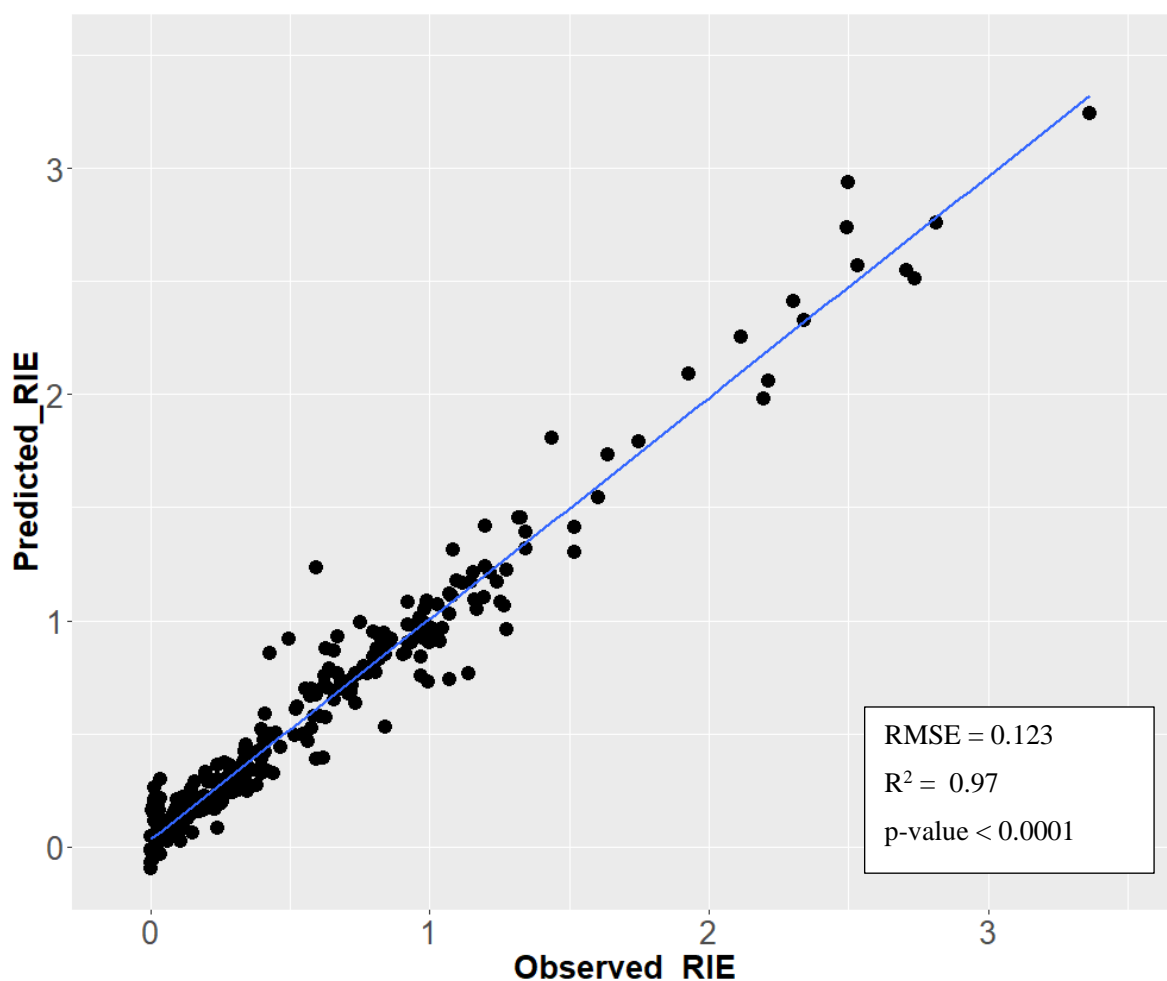


Fig. 2-3 検証データを用いた SVM の予測 RIE と実験で得られた RIE の相関、青線：回帰直線

### 第3節 小括

本章では、LC/MS パラメーター及び化合物の分子記述子によるイオン化効率の予測モデルを開発した。化合物の物理化学的特性を分子記述子に変換し数値化した後に、イオン化効率とは無関係な記述子を削除するべく GA により変数選択を実施した。選ばれた変数を用いて複数の機械学習アルゴリズムから最良のアルゴリズムをスクリーニングした結果、SVM を選択した。10 化合物で取得した全 1428 のデータを用いて学習を行い、別に分けておいたデータセットで検証したところ、予測 RIE と観測された RIE の RMSE は 0.123、 $R^2$  は 0.97 と優れた予測精度を示した。

緒言でも述べたが、化合物のイオン化効率に関する研究報告は多くあるものの、ESI のイオン化メカニズムは複雑であり、LC/MS パラメーターと化合物特性を同時に考慮したイオン化効率予測に関する研究報告はこれまで無かった。今回、医薬品中の GTI に焦点を当てた限定的な適用範囲だが、非線形の回帰モデルを用いた機械学習の検討により、LC/MS パラメーター及び化合物の物理化学的特性を用いた高精度なイオン化効率の予測が可能となることが示された。

### 第3章 イオン化効率予測モデルの医薬品中遺伝毒性不純物への適用

#### 第1節 GTI のイオン化効率予測

前章において、LC/MS パラメーター及び化合物の分子記述子を用いて SVM でイオン化効率の予測モデルを開発し、検証を行ったところ優れた予測精度を有することが確認できた。本章では実際の医薬品中 GTI の分析法開発を想定し、遺伝毒性物質を医薬品原薬に添加し、HPLC で分離後のピークについて、本予測モデルを適用しイオン化効率の予測が可能か評価した。

モデル遺伝毒性化合物には、染色体損傷を引き起こすことが一般的に知られている *N,N*-dimethylaniline を使用した[35]。医薬品への添加実験に先立ち、開発した予測モデルが学習データセットにない新規化合物に対し予測が可能かを検討した。第 1 章と同じく、Table 1 - 2 で示した 5 つの LC/MS パラメーターのそれぞれ 3 水準の組み合わせで作成した 135 の測定条件で *N,N*-dimethylaniline のピーク面積のデータを取得した。また別に、MOE を用いて *N,N*-dimethylaniline の化学構造を分子記述子に変換し、前章の予測モデル開発で記載した通り、選択された分子記述子と LC/MS パラメーターの組み合わせで構成された SVM の予測モデルに代入し、135 の測定条件における RIE の予測値を得た。これら予測値と実験値を比較した結果を Fig. 3 - 1 に示す。

*N,N*-Dimethylaniline の計算された  $pK_a$  は 5.1 であり、分析対象物はギ酸及び TFA 移動相の溶液中でイオン形をとる一方で、酢酸アンモニウム移動相では分子形で存在すると考えられる。第 1 章第 1 節で考察したように分析対象物がすでに液相でイオン化されている場合、高いイオン化効率を得られるが、分子形の場合はイオン化が抑制される。故に、酢酸アンモニウム移動相で得られた RIE が全て 0.1 以下だったと考えられる。一方で、ギ酸及び TFA の条件では RIE が 0.24~1.37 まですばやく分布した。LC/MS のパラメーターの影響によって大きくイオン化効率が変わる結果であったが、全測定に対する予測精度は  $RMSE = 0.207$  であった。これは予測した RIE が 0.2 程度の誤差を平均的に含むことを意味している。第 2 章第 2 節でモデルを検証した際の  $RMSE = 0.123$  と比較すると予測精度は落ちているが、LC/MS の高感度分析法開発において最初を中心条件を選択する上で許容可能な精度を有していると考えられ、開発した予測モデルが学習データにない新たな化合物に対しても予測が可能であることを示した。

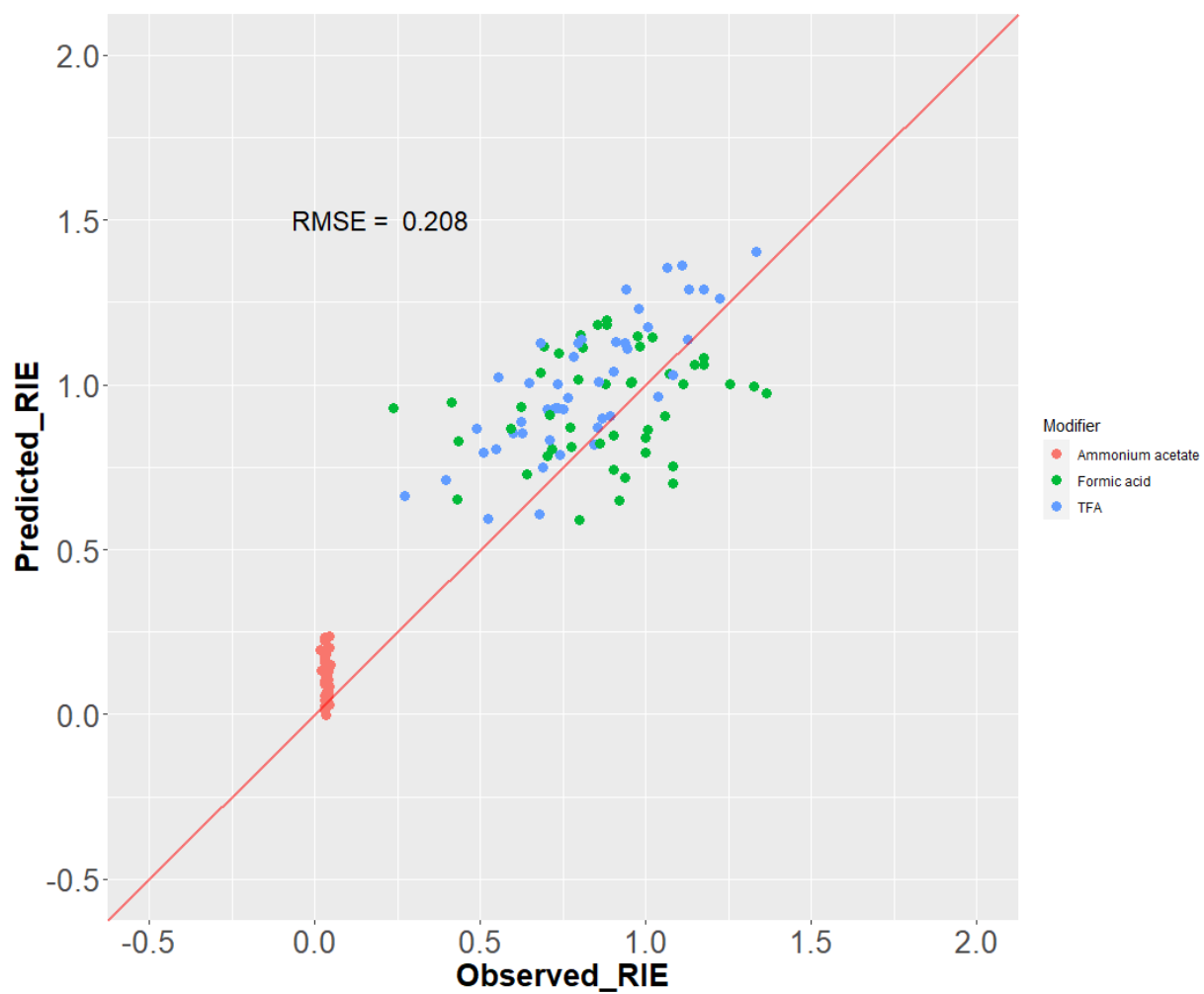


Fig. 3 - 1 *N,N*-Dimethylaniline の予測した RIE と実験で得られた RIE の相関、赤線；傾き 1, 切片 0 の直線

## 第2節 HILIC モードを使用した医薬品原薬中遺伝毒性不純物の分離

これまでイオン化効率の予測には、測定対象単品を試料として flow injection analysis (FIA) によりデータを取得してきた。実際の医薬品原薬中の GTI の測定の際には、試料溶液中には主成分である原薬、製造過程の不純物など様々な夾雑物が含まれる。これら夾雑物と測定対象が共溶出した条件では、測定対象のイオン化に影響を与え、頑健な結果を得ることが出来ないため、HPLC カラムで GTI を分離して測定を実施する必要がある。本節ではモデル医薬品原薬に不純物と GTI を想定した化合物を添加し HPLC で分離後のピークについて、本予測モデルを適用しイオン化効率の予測が可能か検討した。

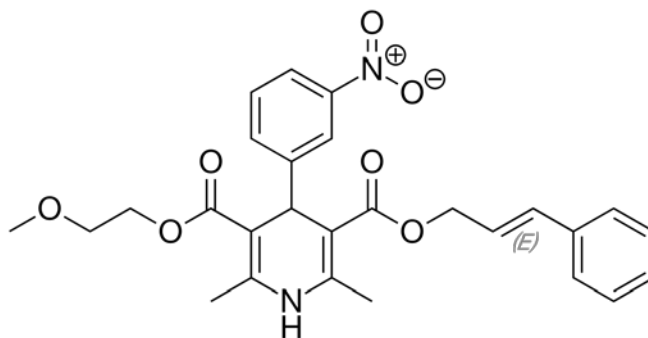
モデル医薬品原薬として cilnidipine (Fig. 3 - 2) を用いることとした。Cilnidipine はカルシウム拮抗薬の 1 つで高血圧症治療薬として市販されている。そこに、不純物を想定した 1-phenylpiperazine と GTI モデル化合物である *N,N*-dimethylaniline を添加した。第 1 章 第 1 節で考察したように有機溶媒比率が高い方が液滴のより効率的な脱溶媒和により感度が向上するケースは多い。一方で、医薬品において不純物分析に一般的に使用されるのは逆相クロマトグラフィーである。逆相クロマトグラフィーで高い有機溶媒比率を使用した場合、溶出力が強く、目的物質が保持されずに夾雑物から分離することが困難になる。そこで、本研究では親水性相互作用クロマトグラフィー(Hydrophilic Interaction Chromatography; HILIC)を使用することとした。

HILIC では主にカラム固定相と移動相で形成された水和層に測定対象が親水性分配して保持される[36]。逆相モードでは十分に保持されない高極性化合物に特に有効な分離手段である[37]。また、HILIC では高い有機溶媒比率を含む移動相を用いて分離できるため、LC/MS での検出感度向上が期待できる[38]。

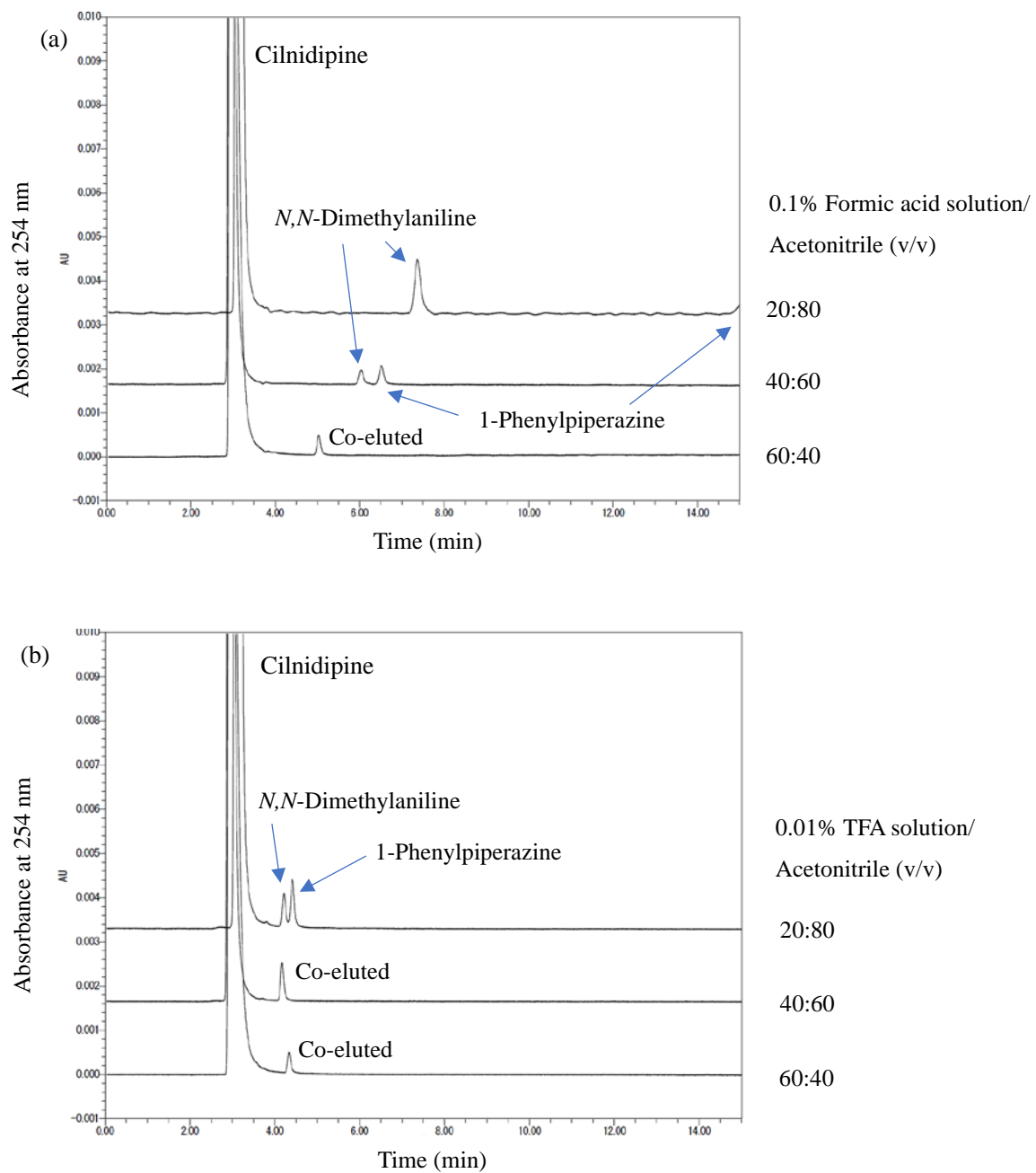
Cilnidipine 溶液に 1-phenylpiperazine 及び *N,N*-dimethylaniline を添加したサンプルを ZIC-HILIC (粒子径 3.5  $\mu\text{m}$ , 4.6 mm I.D $\times$ 150 mm, Merck Millipore) で分離し UV 検出器 (254 nm) で測定した結果を Fig. 3 - 3 に示す。前節の結果から、酢酸アンモニウム移動相では感度向上が期待できないため、ギ酸及び TFA 酸性条件で分離を確認した。医薬品の主成分になる cilnidipine は疎水性が高いため非保持時間 ( $t_0$ ) 付近に溶出し、添加した 2 つの不純物から分離することが出来た。同時に測定した LC/MS の結果をもとにピークを同定すると、ギ酸条件においては、有機溶媒比率 40% で 2 つの添加した不純物が共溶出している。有機溶媒比率を上げていくとカラム固定相に形成された水和層と親水性不純物が強く相互作用することでより保持されるとともに、2 つの不純物の親水性分配の差から分離される傾向が認められた (Fig. 3 - 3 a)。一方で、TFA 条件では有機溶媒比率を上げても保持の改善は認められなかった (Fig. 3 - 3 b)。これは 1-phenylpiperazine 及び *N,N*-dimethylaniline のイオン化したアミノ基と TFA がイオンペアを形成し、水和層との親水性相互作用が働かなかったことが主因と考えられる [39]。これらの結果から、*N,N*-dimethylaniline はギ酸条件、

有機溶媒比率 60%—80%の範囲で夾雑物の影響を受けることなく LC/MS の測定が可能であることが明らかになった。

この移動相条件の範囲において、本研究により開発したイオン化効率の予測モデルを適用し、高感度検出条件を探索した。



**Fig. 3 - 2** Cilnidipine の化学構造



**Fig. 3 - 3** Cilnidipine に添加した不純物の HILIC モードを使用した分離 (a) ギ酸条件 (b) TFA 条件

### 第3節 HILIC モードで分離した医薬品原薬中遺伝毒性不純物のイオン化効率予測

前節で明らかになった分離可能な移動相条件の範囲において、本研究により開発したイオン化効率の予測モデルを適用した。ギ酸酸性条件において有機溶媒比率は 60%~80%、プローブ温度は 300°C~600°C、キャピラリー電圧は 0.8~1.2 kV を検討範囲とし、コーン電圧が与える影響は小さいことが第 1 章 第 2 節の結果から分かったので中央値の 15V で固定した。開発した予測モデルを用い、上記の範囲で *N,N*-dimethylaniline のイオン化効率を予測した。一例としてキャピラリー電圧を 1.0 kV で固定し、有機溶媒比率、プローブ温度と予測 RIE の関係を 3D グラフにした結果を Fig. 3-4 に示す。本予測モデルは学習データの範囲内であれば、任意のパラメーターの値における予測 RIE を求めることができる。プローブ温度 500°C、有機溶媒比率 60% 付近のイオン化効率がここでは良いことが分かる。キャピラリー電圧を変動させ同様の検討を実施したが、グラフに示した 1.0 kV 付近が最も良好であった。このように、医薬品原薬に含まれる様々な夾雑物から GTI を分離する HPLC 条件の範囲が決まり、その許容される範囲において開発した手法により最もイオン化効率が高くなる条件を探索できることが示唆された。

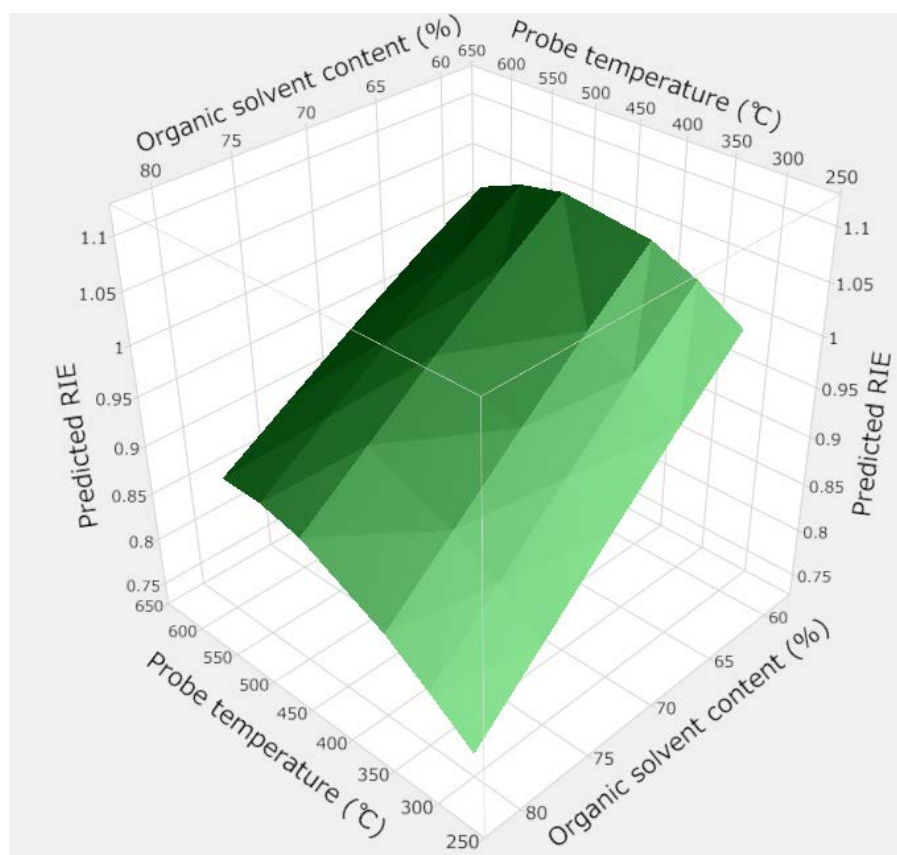


Fig. 3-4 *N,N*-Dimethylaniline のイオン化効率を予測した 3D グラフ

次に、ランダムに生成した測定条件において、cilnidipine 溶液に 1-phenylpiperazine 及び *N,N*-dimethylaniline を添加したサンプルを前節と同様に ZIC-HILIC で分離し MS 測定した。ICH M7 ガイドラインでは、GTI のリスクが無視できる許容摂取量は 1.5 µg/person/day と定められている。1 日の投与量が原薬換算で 10 mg の場合、1.5 µg/person/day は原薬の 150 ppm に相当する。今回、1 mg/mL の cilnidipine に対し、80 ppm の *N,N*-dimethylaniline を添加し検証を実施した。GTI の観測された RIE と予測 RIE の比較結果を Table 3 - 1 に示す。有機溶媒比率が 70%、60% の条件で特に大きな乖離が認められた。本章 第 1 節で FIA においては良好な予測精度を有していたが、HPLC カラムで分離後のピークに対して大きな予測誤差が認められた原因は、カラムからのブリードによる影響[40]が考えられる。ブリードとは、カラム固定相からの分解生成物の溶出により生成されるバックグラウンドシグナルのことである。LC/MS などの高感度分析の場合、ブリードが大きいとバックグラウンドノイズが高くなって感度が低下することがある。一方で、別に実施した methyl 4-aminobenzoate における FIA での RIE と、ZIC-HILIC で分離後の RIE の比較を Table 3 - 2 に示すが、同等な結果が得られており、影響の受け方は化合物の分子量や他の特性にも依存すると考えられる。

**Table 3 - 1** Cilnidipine に添加した *N,N*-Dimethylaniline の HILIC/MS を用いた実測 RIE と予測 RIE の比較

Observed RIE	Predicted RIE	Organic content (%)	Probe temperature (°C)	Cone voltage (V)	Capillary voltage (kV)
0.88	0.68	80	300	15	0.8
1.00	0.86	80	500	15	1
0.94	0.87	80	600	15	1
0.76	0.86	70	400	15	0.8
0.69	0.88	70	500	15	0.8
0.55	0.92	70	600	15	1.2
0.45	1.05	60	400	15	1
0.46	1.06	60	500	15	1
0.36	0.97	60	300	15	1.2

**Table 3 - 2 Methyl 4-aminobenzoate を用いた FIA による実測 RIE と ZIC-HILIC で保持後の実測 RIE の比較**

Observed RIE		Modifier type	Organic content (%)	Probe temperature (°C)	Cone voltage (V)	Capillary voltage (kV)
FIA	HPLC					
0.85	0.80	Formic acid	60	300	11	0.8
0.53	0.49	Formic acid	40	450	19	1.0
0.56	0.69	TFA	80	300	15	1.0
0.38	0.43	TFA	60	450	15	1.2
0.02	0.01	Ammonium acetate	60	600	15	1.0
0.02	0.01	Ammonium acetate	40	300	19	1.2

## 第4節 小括

本章では、実際の医薬品原薬中 GTI の分析法開発を想定し、遺伝毒性物質を医薬品原薬に添加し、HPLC で分離後のピークについて、本予測モデルを適用しイオン化効率の予測が可能か評価した。

GTI のモデル化合物には *N,N*-dimethylaniline を用い、予測モデル開発時と同じく FIA で 5 つの LC/MS パラメーターのそれぞれ 3 水準の組み合わせで作成した 135 の測定条件で得た RIE と、開発した SVM の予測モデルから得た RIE を比較したところ RMSE = 0.207 となり、開発した予測モデルが学習データにない新たな化合物に対しても予測が可能であることを示した。

次に、モデル医薬品原薬として cilnidipine を用い、不純物を想定した 1-phenylpiperazine と GTI モデル化合物 *N,N*-dimethylaniline を添加して HILIC による分離検討を行ったところ、ギ酸酸性条件の有機溶媒比率 60%~80% の範囲で夾雑物の影響を受けることなく LC/MS の測定可能であることが明らかになった。そして、この移動相条件の範囲において、開発した予測モデルを適用し任意の条件における RIE の予測値を得る方法を確認した。しかしながら、実際に cilnidipine に 1-phenylpiperazine と *N,N*-dimethylaniline を添加したサンプルを ZIC-HILIC で分離し MS 測定したところ、実測の RIE と予測の RIE において有機溶媒比率が 70%、60% の条件で特に大きな乖離が認められた。その原因としてカラムからのブリードによる影響が考えられる。以上、HPLC カラムで分離後のピークに対しては大きな予測誤差が認められたが、FIA においては良好な予測精度を有し、分離可能な移動相の範囲において、任意の条件における予測 RIE を得ることが出来た。

## 総括

医薬品中の GTI はわずかな量でも DNA 損傷を引き起こす可能性があるため、微量定量が必要となり、高い特異性と検出感度を有する LC/MS は強力なツールとして期待される。LC/MS の ESI における化合物のイオン化効率、LC/MS の条件パラメーター、化合物特性などの要因によって影響を受ける。しかしながら、LC/MS パラメーターと化合物特性を同時に考慮したイオン化効率予測に関する研究報告はこれまで無かった。以上を踏まえ、本研究では医薬品開発を効率化することを最終目標とした簡便かつ迅速な LC/MS イオン化効率予測法の開発を目的とした

第 1 章では医薬品中の不純物を模した 10 種の化合物を用いて、5 つの LC/MS パラメーターのそれぞれ 3 水準の組み合わせで作成した 135 の条件でそれぞれデータ取得を実施した。化合物内で LC/MS パラメーターを説明変数、ピーク面積を目的変数として重回帰分析を行い、ピーク面積の予測モデルを取得したところ良好な予測結果を示し、LC/MS パラメーターでピーク面積が予測可能であることが明らかになった。一方で、各化合物を相対比較できるように標準条件で補正された RIE を目的変数とし、LC/MS パラメーターを説明変数として 10 種の化合物を用いた重回帰分析を行ったところ予測誤差は大きく、イオン化効率を予測するには、化合物特性を回帰モデルに組みこむことで精度が向上すると考えられた。

第 2 章では LC/MS パラメーターに加え、化合物特性を組み込んだ予測モデルを開発するため、化学構造を分子記述子で数値化し、GA により変数選択した後に、SVM を用いた機械学習を行った。学習データとは別に分けておいた検証データを用いてバリデーションを実施したところ、RMSE は 0.123、 $R^2$  は 0.97 と優れた予測精度を示した。

第 3 章では本手法の有用性を確認するために、学習データにない遺伝毒性物質 *N,N*-dimethylaniline のイオン化効率の予測を実施した。確立した機械学習モデルに *N,N*-dimethylaniline の分子記述子を代入し、種々の条件における RIE の予測値と実測値を比較したところ、RMSE = 0.207 程度の予測精度が得られた。LC/MS の高感度分析法開発において、最適化を始める中心条件として許容可能な精度を有しており、開発した予測モデルが学習データにない新たな化合物に対しても予測可能であることが示された。また、モデル医薬品原薬 cilnidipine に不純物を想定した 1-phenylpiperazine と GTI モデル化合物である *N,N*-dimethylaniline を添加したサンプルを用い、HPLC で分離後のピークについて、本予測モデルを適用した。開発した手法により HPLC 分離で許容される範囲において最もイオン化効率が高くなる条件を探索できることが示された。しかしながら、有機溶媒比率が低い条件で実測の RIE と予測の RIE において大きな乖離が認められた。

本研究では医薬品に多い塩基性化合物を中心に検討したため、中性および酸性化合物は追加の検討が必要である。また、今回の検討には含めなかった他の LC/MS パラメーター（有機溶媒の種

類、モディファイヤー濃度など) についても同様の検討により汎用性が高い予測モデルになる可能性がある。

本研究を基盤とした更なる研究の進展により、機械学習を用いた医薬品中 GTI の簡便かつ迅速な分析法開発が可能になり、医薬品開発の効率化が期待される。

## 実験の部

### 試薬

試薬名	グレード	供給者
acetonitrile	HPLC	関東化学株式会社
ammonium acetate	特級	関東化学株式会社
<i>N</i> -(3-amino-4-methoxyphenyl)acetamide	>98.0%	東京化成工業株式会社
4-aminomethyl-tetrahydropyran	>97%	富士フィルム和光純薬株式会社
butyl 4-aminobenzoate	>99.0%	東京化成工業株式会社
cilnidipine	>98.0%	東京化成工業株式会社
<i>N</i> -nitrosodipropylamine	>98.0%	東京化成工業株式会社
<i>N,N</i> -dimethylaniline	特級	富士フィルム和光純薬株式会社
ethyl 3-aminobenzoate	>97.0%	東京化成工業株式会社
formic acid	LC/MS 用	和光純薬工業株式会社
isopropyl 4-aminobenzoate	>98.0%	東京化成工業株式会社
methyl 4-aminobenzoate	>98.0%	東京化成工業株式会社
methyl 4-amino-2-chlorobenzoate	>95%	Matrix Scientific
methyl 4-(methylanino)benzoate	>98%	Alfa Aesar
1-phenylpiperazine	>98.0%	東京化成工業株式会社
trifluoroacetic acid	特級	関東化学株式会社

水は Milli-Q Integral 15 超純水製造装置 (Merck Millipore, Tokyo, Japan) により製造した超純水を使用した。

### 装置

装置名	型式	供給者
pH 計	HM-30R	東亜ディーケーケー株式会社
UHPLC	ACQUITY UPLC H-class	Waters Corporation
MS	ACQUITY QDa	Waters Corporation
電子天秤	ME235S	Sartorius

## ソフトウェア

名称	バージョン	供給者
Empower3	7.41.00.00	Waters Corporation
Microsoft Excel 2010	Office 365 バージョン 2002	Microsoft Corporation
JMP	14.2.0	SAS Institute Inc.
MOE	2019.0102	Chemical Computing Group
R	3.6.3	-

## R パッケージ

名称	パッケージ情報
caret	Classification and Regression Training. R package version 6.0-85, 2020. <a href="https://cran.r-project.org/package=caret">https://cran.r-project.org/package=caret</a> .
dplyr	Hadley Wickham, Romain Francois, Lionel Henry and Kirill Muller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. <a href="https://CRAN.R-project.org/package=dplyr">https://CRAN.R-project.org/package=dplyr</a>
gaselect	Genetic Algorithm (GA) for Variable Selection from High-Dimensional Data. R package version 1.0.7., 2019. <a href="https://cran.r-project.org/package=gaselect">https://cran.r-project.org/package=gaselect</a> .
ggplot2	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
kernlab	Kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20, 2004. <a href="http://www.jstatsoft.org/v11/i09/">http://www.jstatsoft.org/v11/i09/</a> .

## 第 1 章 第 1 節の実験操作

### 測定試料

試料：Table 1 - 1 に示した 10 種の化合物

試料溶解液：0.1% ギ酸水溶液とアセトニトリルを 40:60 (v/v) で混和した溶液、0.01% TFA 水溶液とアセトニトリルを 40:60 (v/v) で混和した溶液、5 mM 酢酸アンモニウム緩衝液とアセトニトリルを 40:60 (v/v) で混和した溶液の 3 種の試料溶解液を用意し、水系移動相の種類に合わせて使用した。

試料溶液：試料溶液濃度は 0.2 µg/mL。

### 移動相

移動相 A：0.1% (v/v) ギ酸水溶液

移動相 B：0.01% (v/v) TFA 水溶液

移動相 C : 5 mM 酢酸アンモニウム緩衝液 (pH 7.0)

移動相 D : アセトニトリル

### UHPLC 条件

流速 :	0.5 mL/min
モード :	アイソクラテック
移動相混合比 (アセトニトリル/水系移動相) :	80%, 60%, 40%
注入量 :	2 $\mu$ L
分析時間 :	3 分

### MS 条件

イオンソース :	ESI
イオン化モード :	ポジティブイオンモード
測定モード :	SIR
	$m/z$ = 各化合物の $[M + H]^+$
プローブ温度 (°C) :	300, 450, 600
コーン電圧 (V) :	11, 15, 19
キャピラリー電圧 (kV) :	0.8, 1.0, 1.2
サンプリングレート :	10 points/sec

### 測定内容

Table 1 - 2 に示した 5 つの LC/MS パラメーターのそれぞれ 3 水準において、JMP を用いた実験計画法で 135 の条件を作成した。異なる LC/MS パラメーターの組み合わせからなる 135 条件を用いて、各化合物を FIA で測定し、ピーク面積のデータを取得した。

### データ解析

エクセルを用いて化合物ごとに得られたピーク面積を緩衝液の種類及び有機溶媒比率で並べ直し、散布図でグラフ化を行った。

## 第 1 章 第 2 節の実験操作

### データ

第 1 章 第 1 節で測定したデータを用いた。

## データ解析

JMP を用いて各 LC/MS パラメーターを説明変数、ピーク面積を目的変数として重回帰分析を行った。3 次まで考慮した多項式で回帰し、実測値と予測値のグラフ及び LogWorth 値を含む統計量を出力した。

### 第 1 章 第 3 節の実験操作

- ・ RIE の日間差に関する検討

#### 測定試料

試料：Methyl 4-aminobenzoate

試料溶解液：0.1%ギ酸水溶液とアセトニトリルを 40:60 (v/v)で混和した溶液、0.01% TFA 水溶液とアセトニトリルを 40:60 (v/v)で混和した溶液、5 mM 酢酸アンモニウム緩衝液とアセトニトリルを 40:60 (v/v)で混和した溶液の 3 種の試料溶解液を用意し、水系移動相の種類に合わせて使用した。

試料溶液：試料溶液濃度は 0.2 µg/mL。

#### 測定条件

第 1 章 第 1 節の条件に準じた。

#### 測定内容

methyl 4-aminobenzoate を用いて、測定日を変えた RIE を求め、日間再現性を確認した。

- ・ 重回帰分析による化合物横断的イオン化効率の予測

## データ

第 1 章 第 1 節で測定したデータを用いた。

## データ解析

各 LC/MS パラメーターを説明変数、RIE を目的変数として R を用いた重回帰分析を行った。3 次まで考慮した多項式で回帰し、実測値と予測値のグラフを出力した。また、計算式 (2) に従い、RMSE を算出した。

以下に R のコードを示す。

```
#ライブラリーの読み込み
```

```
library(ggplot2)
```

```
library(dplyr)
```

```
#データ読み込み
```

```
data <- read.csv("dataset_R.csv")
```

```

data <- data %>%
  slice(1:1435)

#欠損値のある行を削除
data <- na.omit(data)

#重回帰分析
model_lm <- lm(Area_STD ~. ^3, data = data)
summary(model_lm)
pred_lm <- predict(model_lm, newdata = data)
RMSE_lm <- sqrt(sum((pred_lm - data$Area_STD)^2)/nrow(data))

#グラフ化
tibble(actual_RIE = data$Area_STD, predicted_RIE = pred_lm) %>%
  ggplot(aes(x = actual_RIE, y = predicted_RIE)) +
  xlim(-0.5, 4) +
  ylim(-0.5, 4) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red", size = 1, alpha = 0.5)+
  annotate("text", x = 0, y = 3, label = paste("RMSE =",round(RMSE_lm, 3)), size = 7)

```

## 第2章 第1節の実験操作

- ・ 分子記述子への変換

### 操作

Table 1 - 1 に示した 10 種の化合物の化学構造を MOE にスケッチし、分子動力学を利用して配座解析する LowModeMD[41]を使用して安定配座を得た。そこから 2D 及び 3D の記述子を出力した。

- ・ 遺伝的アルゴリズム

### データ

第1章 第1節で測定したデータを用いた。

### データ解析

R を用いて LC/MS パラメーター・分子記述子から RIE に対する重要変数を選択するため、GA による解析を行った。1 世代ごとの染色体の数 (populationSize) を 100、世代数 (numGenerations) を 1500、選ばれる説明変数の最小/最大値を 13/15、適合度が高く世代を超えて引き継がれる染色体の数 (elitism) を 10 と設定した。適合度を求める方法は PLS (Partial Least Squares) 法を採用した。評価は繰り返し回数 5 のダブルクロスバリデーションで実施した。

以下に R のコードを示す。

```
#パッケージの読み込み
```

```
library(gaselect)
```

```
library(tidyr)
```

```
library(dplyr)
```

```
#データの読み込み
```

```
dataset <- read.csv("dataset_R.csv")
```

```
#標準化
```

```
data_norm <- dataset
```

```
for (i in 2:ncol(data_norm)){
```

```
  data_norm[, i] <- (data_norm[, i] - min(data_norm[, i])) / (max(data_norm[, i]) - min(data_norm[, i]))
```

```
}
```

```
# 外部バリデーションのデータを抜く
```

```
datatest <- data_norm %>%
```

```
  slice(1436:nrow(data_norm))
```

```
data_norm <- data_norm %>%
```

```
  slice(1:1435)
```

```
#面積値 NA の行を削除
```

```
datatest <- filter(datatest, Area_STD != "NA")
```

```
data_norm <- filter(data_norm, Area_STD != "NA")
```

```
#無作為化
```

```
data_norm <- cbind(data_norm, index = sample(1:nrow(data_norm)))
```

```
data_norm <- data_norm[order(data_norm[ncol(data_norm)]), ]
```

```
data_norm <- data_norm[, -ncol(data_norm)]
```

```
#train data から分散が 0 の列を削除
```

```
data_nZero <- select(data_norm, -nearZeroVar(data_norm, names = TRUE))
```

```
#GA-PLS
```

```
ctrl <- genAlgControl(populationSize = 100, numGenerations = 1500, minVariables = 13, maxVariables =  
15, elitism = 10, mutationProbability = 0.01, crossover = c("random"), verbosity = 1)
```

```
evaluatorRDCV <- evaluatorPLS(numReplications = 5, innerSegments = 10, outerSegments = 3,
                               numThreads = 1)
set.seed(12345)
X <- as.matrix(data_nZero[, -1])
y <- drop(data_nZero[, 1]);
result <- genAlg(y, X, control = ctrl, evaluator = evaluatorRDCV, seed = 123)
```

## 第2章 第2節の実験操作

- ・ 機械学習アルゴリズムのスクリーニング

### データ

第1章 第1節で測定したデータを用いた。

### データ解析

GA の解析で選択された分子記述子と LC/MS のパラメーターを合わせて説明変数とし、目的変数 RIE を予測する最良のアルゴリズム選択のため、R を用いた機械学習アルゴリズムのスクリーニングを実施した。線形回帰と非線形回帰の 8 つのアルゴリズムで比較を行った。線形回帰からは線形モデル (linear model)、リッジ回帰 (ridge)、ラッソ回帰 (lasso)、部分最小 2 乗回帰 (PLS) を選択した。非線形回帰からは k 最近傍法 (kNN)、決定木 (CART)、ランダムフォレスト (random forest)、サポートベクターマシン (SVM) を選択した。評価は 5 分割クロスバリデーションで得られた RMSE で行った。

以下に R のコードを示す。

# パッケージの読み込み

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(caret)
library(randomForest)
library(kernlab)
```

# データの読み込み

```
dataset <- read.csv("dataset_R.csv")
```

# 標準化

```
data_norm <- dataset
for (i in 2:ncol(data_norm)){
  data_norm[, i] <- (data_norm[, i] - min(data_norm[, i])) / (max(data_norm[, i]) - min(data_norm[, i]))}
```

```

# 外部バリデーションのデータを抜く
datatest <- data_norm %>%
  slice(1436:nrow(data_norm))
data_norm <- data_norm %>%
  slice(1:1435)

#NA の行を削除
datatest <- filter(datatest, Area_STD != "NA")
data_norm <- filter(data_norm, Area_STD != "NA")

#無作為化
data_norm <- cbind(data_norm, index = sample(1:nrow(data_norm)))
data_norm <- data_norm[order(data_norm[ncol(data_norm)]), ]
data_norm <- data_norm[, -ncol(data_norm)]

#train data から分散が 0 の列を削除
data_nZero <- select(data_norm, -nearZeroVar(data_norm,names = TRUE))

#GA で選択されたパラメータ、LC/MS パラメータでリスト作成
X1<-c('E_rele','FASA..1','radius','mr', 'SlogP_VSA6', 'Weight', 'vsurf_EDmin1',
'vsurf_W8','SMR_VSA6','ASA_P', 'a_aro','E_nb')
columnList <- data.frame(X1) #make columnList
columnList.vector<-as.vector(columnList[,1])
parameter.vector<-as.vector(colnames(dataset[,2:8]))
columnList.bind <- c(columnList.vector, parameter.vector)
data.selected <- select(.data = data_norm, columnList.bind)
data.selected <- cbind(data_norm[, 1],data.selected)
names(data.selected)[ c(1)] <- c(names(data_norm)[ c(1)])

#Caret によるスクリーニング
library(caret)
models <- tibble(model = c("Linear model", "Random forest", "SVM (rbf)", "lasso", "PLS", "Ridge",
"kNN", "CART"),
  method =c("lm", "rf", "svmRadial", "lasso", "pls", "ridge", "knn", "rpart2"),
  RMSE = c(NA))
trControl <- trainControl(method = "repeatedcv", number = 5, repeats = 3)
set.seed(1234)

```

```
for (i in 1:nrow(models)) {
  model_temp <- train(Area_STD ~., data = data.selected, method = models$method[i],
    tuneLength = 10, trControl = trControl)
  models$RMSE[i] <- min(model_temp[["results"]])$RMSE)
}
```

#グラフ化

```
ggplot(data = models, aes(x = reorder(model, -RMSE), y = RMSE)) +
  geom_bar(stat = "identity") +
  coord_flip()
```

- ・ SVM を用いた機械学習モデルの最適化

### データ

第1章第1節で測定したデータを用いた。

### データ解析

スクリーニングで優れた予測精度を示した **SVM** について最適化と検証を実施した。カーネル関数には、データに関する事前知識が十分でない場合に用いられる汎用的なガウシアンカーネルを用いた[42]。最適化したハイパーパラメーターは、カーネル関数の広がり制御するシグマの値と、誤差を許容する幅を示すコストパラメーターの値である。10 化合物で取得したデータをランダム化し、8 対 2 の比率で学習データセットと検証データセットの 2 つのグループに分けた。まず、学習データセットを用いて **SVM** アルゴリズムの最適化を行い、機械学習モデルを構築した。次に、ホールドアウトした検証データセットを使用して **RIE** の予測を行い、実測値と予測値から **RMSE** の計算を行った。

以下に R のコードを示す。

# パッケージの読み込み

```
library(ggplot2)
```

```
library(tidyr)
```

```
library(dplyr)
```

```
library(kernlab)
```

# データの読み込み

```
dataset <- read.csv("dataset_R.csv")
```

```

##標準化
data_norm <- dataset
for (i in 2:ncol(data_norm)){
  data_norm[, i] <- (data_norm[, i] - min(data_norm[, i])) / (max(data_norm[, i]) - min(data_norm[, i]))
}

# 外部バリデーションのデータ抜く
datatest <- data_norm %>%
  slice(1436:nrow(data_norm))
data_norm <- data_norm %>%
  slice(1:1435)

#NA の行を削除
datatest <- filter(datatest, Area_STD != "NA")
data_norm <- filter(data_norm, Area_STD != "NA")

#無作為化
data_norm <- cbind(data_norm, index = sample(1:nrow(data_norm)))
data_norm <- data_norm[order(data_norm[ncol(data_norm)]), ]
data_norm <- data_norm[, -ncol(data_norm)]

#train data から分散が 0 の列を削除
data_nZero <- select(data_norm, -nearZeroVar(data_norm, names = TRUE))

#GA で選択されたパラメーターと LC/MS パラメーターでリスト作成
X1<-c('E_rele','FASA..1','radius','mr', 'SlogP_VSA6', 'Weight', 'vsurf_EDmin1', 'vsurf_W8', 'SMR_VSA6',
'ASA_P', 'a_aro', 'E_nb')
columnList <- data.frame(X1) #make columnList
columnList.vector<-as.vector(columnList[,1])
parameter.vector<-as.vector(colnames(dataset[,2:8]))
columnList.bind <- c(columnList.vector, parameter.vector)
data.selected <- select(.data = data_norm, columnList.bind)
data.selected <- cbind(data_norm[, 1],data.selected)
names(data.selected)[ c(1)] <- c(names(data_norm)[ c(1)])
columnList

```

```

#SVR
d <- floor(nrow(data.selected) * 0.8)
formula.model <- as.formula(paste(colnames(data.selected)[1], "~."))
model <- ksvm(formula.model, data = data.selected[1:d, ], type="eps-svr", kernel="rbfdot", kpar =
list(sigma=0.08), cross=5, C=5)
test <- data.frame(actual_RIE = data.selected[(d + 1):nrow(data.selected), 1], predicted_RIE =
predict(model, data.selected[(d + 1):nrow(data.selected), 2:ncol(data.selected)]))
RMSE <- round(sqrt(sum((test$actual_RIE - test$predicted_RIE)^2)/(nrow(data.selected) - d)), 5)

#グラフ化
res <- summary(lm(formula = predicted_RIE ~ ., data = test))
r1 = round(res$r.squared, digits = 3)
ggplot(data = test, aes(x = actual_RIE, y = predicted_RIE)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  xlim(-0.1, 4) +
  ylim(-0.1, 4) +
  annotate("text", x=0.2, y=2.5, label = as.character(paste("RMSE = ", round(RMSE, 3)), size = 6)) +
  annotate("text", x = 0.2, y = 2, label = as.character(paste("R2 = ", round(r1, 2))), size = 4)

```

### 第3章 第1節の実験操作

- ・ *N,N*-Dimethylaniline の RIE 実測

#### 測定試料

試料： *N,N*-Dimethylaniline

試料溶解液：0.1% ギ酸水溶液とアセトニトリルを 40:60 (v/v) で混和した溶液、0.01% TFA 水溶液とアセトニトリルを 40:60 (v/v) で混和した溶液、5 mM 酢酸アンモニウム緩衝液とアセトニトリルを 40:60 (v/v) で混和した溶液の 3 種の試料溶解液を用意し、水系移動相の種類に合わせて使用した。

試料溶液：試料溶液濃度は 0.08 µg/mL。

#### 測定条件

第1章 第1節の条件に準じた。

#### 測定内容

第1章 第1節と同様に、異なる LC/MS パラメーターの組み合わせで作成した 135 の条件で FIA の測定を実施し、ピーク面積のデータを取得した。

## データ解析

計算式 (1) を用いて実測値の RIE を算出した。

- ・ *N,N*-Dimethylaniline の RIE 予測

## データ解析

MOE を用いて *N,N*-dimethylaniline の化学構造を第 2 章 第 1 節の記載した操作で分子記述子に変換した。その中から、第 2 章 第 1 節で GA により選ばれた 12 の記述子 'E\_rele', 'FASA..1', 'radius', 'mr', 'SlogP\_VSA6', 'Weight', 'vsurf\_EDmin1', 'vsurf\_W8', 'SMR\_VSA6', 'ASA\_P', 'a\_aro', 'E\_nb' の値を SVM の予測モデルに代入し、135 の LC/MS パラメーターの組み合わせにおける RIE の予測値を算出した。実測から得られた RIE と予測値を用いて RMSE を算出した。

以下に R のコードを示す。

```
# パッケージの読み込み
library(ggplot2)
library(tidyr)
library(dplyr)
library(kernlab)

# データの読み込み
dataset <- read.csv("dataset_R.csv")

# 標準化
data_norm <- dataset
for (i in 2:ncol(data_norm)){
  data_norm[, i] <- (data_norm[, i] - min(data_norm[, i])) / (max(data_norm[, i]) - min(data_norm[, i]))
}

# 外部バリデーションのデータ抜く
datatest <- data_norm %>%
  slice(1436:nrow(data_norm))
data_norm <- data_norm %>%
  slice(1:1435)

# NA の行を削除
datatest <- filter(datatest, Area_STD != "NA")
data_norm <- filter(data_norm, Area_STD != "NA")
```

```

#無作為化
data_norm <- cbind(data_norm, index = sample(1:nrow(data_norm)))
data_norm <- data_norm[order(data_norm[ncol(data_norm)]), ]
data_norm <- data_norm[, -ncol(data_norm)]

#train data から分散が 0 の列を削除
data_nZero <- select(data_norm, -nearZeroVar(data_norm,names = TRUE))

#選択されたパラメータ+LC/MS パラメータでリスト作成
X1<-c('E_rele', 'FASA..1', 'radius', 'mr', 'SlogP_VSA6', 'Weight', 'vsurf_EDmin1', 'vsurf_W8', 'SMR_VSA6',
'ASA_P', 'a_aro', 'E_nb')
columnList <- data.frame(X1) #make columnList
columnList.vector<-as.vector(columnList[,1])
parameter.vector<-as.vector(colnames(dataset[,2:8]))
columnList.bind <- c(columnList.vector, parameter.vector)
data.selected <- select(.data = data_norm, columnList.bind)
data.selected <- cbind(data_norm[, 1],data.selected)
names(data.selected)[ c(1)] <- c(names(data_norm)[ c(1)])

#検証データセットから変数選択
datatest.select <- select(.data = datatest, columnList.bind)
datatest.select <- cbind(datatest[, 1],datatest.select)
names(datatest.select)[ c(1)] <- c(names(datatest)[ c(1)])

#外部バリデーション
formula.model <- as.formula(paste(colnames(data.selected)[1], "~."))
model2 <- ksvm(formula.model, data = data.selected, type="eps-svr",kernel="rbfdot",kpar =
list(sigma=0.08),cross=5, C=5)
test2 <- data.frame(actual = datatest.select[,1 ], predicted = predict(model2, datatest.select[,
2:ncol(datatest.select)]))
RMSE <- round(sqrt(sum((test2$actual - test2$predicted)^2)/(nrow(datatest.select))), 5)
ggplot(data = test2, aes(x = actual, y = predicted)) +
  geom_point() +
  xlim(-0.5, 2) +
  ylim(-0.5, 2) +
  geom_abline(colour = "red", alpha = 0.5, size = 1, linetype = 2) +

```

```
annotate("text", x=0.2, y=0.9, label = as.character(paste("RMSE = ", RMSE)))
```

## 第3章 第2節の実験操作

### 測定試料

試料：cilnidipine、1-phenylpiperazine、*N,N*-dimethylaniline

試料溶解液：0.1%ギ酸とアセトニトリルを2:8で混和した溶液を試料溶解液とした。

試料溶液：cilnidipineは1 mg/mL、1-phenylpiperazine及び*N,N*-dimethylanilineは4 µg/mLになるように試料溶液を調製した。

### 移動相

移動相 A：0.1% (v/v) ギ酸水溶液

移動相 B：0.01% (v/v) TFA 水溶液

移動相 D：アセトニトリル

### UHPLC 条件

分析カラム：ZIC-HILIC (粒子径 3.5 µm, 4.6 mm I.D×150 mm, Merck Millipore)

流速：0.5 mL/min

モード：アイソクラテック

移動相混合比（アセトニトリル/水系移動相）：80%, 60%, 40%

注入量：2 µL

分析時間：15 分

UV 検出器：254 nm

サンプリングレート：20 points/sec

## 第3章 第3節の実験操作

- ・ 分離可能な移動相条件範囲におけるイオン化効率予測

### データ解析

第3章 第1節で作成した*N,N*-dimethylanilineの予測モデルを用いて、以下の範囲における RIE の予測を行った。得られた予測 RIE の結果を JMP に取り込み、曲面グラフを描いた。

モディファイヤーの 種類	コーン電圧 (V)	有機溶媒比率 (v/v %)	プローブ温度 (°C)	キャピラリー 電圧 (kV)
ギ酸	15	60	300	0.8
		65	350	0.9
		70	400	1.0
		75	500	1.1
		80	550	1.2
		-	600	-

- ・ 分離可能な移動相条件範囲におけるイオン化効率実測

#### 測定試料

試料：cilnidipine、1-phenylpiperazine、*N,N*-dimethylaniline

試料溶解液：0.1%ギ酸とアセトニトリルを2:8で混和した溶液を試料溶解液とした。

試料溶液：cilnidipine は1 mg/mL、1-phenylpiperazine は0.2 µg/mL、*N,N*-dimethylaniline は0.08 µg/mL になるように試料溶液を調製した。

#### 移動相

移動相 A：0.1% (v/v) ギ酸水溶液

移動相 D：アセトニトリル

#### LC/MS 条件

以下の9つの条件でデータを取得した。

有機溶媒比率 (v/v %)	プローブ温度 (°C)	コーン電圧 (V)	キャピラリー 電圧 (kV)
80	300	15	0.8
80	500	15	1
80	600	15	1
70	400	15	0.8
70	500	15	0.8
70	600	15	1.2
60	400	15	1
60	500	15	1
60	300	15	1.2

### 共通条件

分析カラム:	ZIC-HILIC (粒子径 3.5 $\mu\text{m}$ , 4.6 mm I.D $\times$ 150 mm, Merck Millipore)
流速:	0.5 mL/min
モード:	アイソクラティック
注入量:	2 $\mu\text{L}$
分析時間:	15 分
イオンソース:	ESI
イオン化モード:	ポジティブイオンモード
測定モード:	SIR
	$m/z = 122.1$
サンプリングレート:	10 points/sec

- ・ Methyl 4-aminobenzoate を用いた FIA と HPLC 分離後の RIE 比較

### 測定試料

試料: Methyl 4-aminobenzoate

試料溶解液: 0.1% ギ酸とアセトニトリルを 4:6 で混和した溶液を試料溶解液とした。

試料溶液: Methyl 4-aminobenzoate は 0.2  $\mu\text{g/mL}$  になるように試料溶液を調製した。

### 移動相

移動相 A: 0.1% (v/v) ギ酸水溶液

移動相 B: 0.01% (v/v) TFA 水溶液

移動相 C: 5 mM 酢酸アンモニウム緩衝液

移動相 D: アセトニトリル

## LC/MS 条件

以下の 6 つの条件でデータを取得した。

モディファイヤーの 種類	有機溶媒比率 (v/v %)	プローブ温度 (°C)	コーン電圧 (V)	キャピラリー 電圧 (kV)
ギ酸	60	300	11	0.8
ギ酸	40	450	19	1.0
TFA	80	300	15	1.0
TFA	60	450	15	1.2
酢酸アンモニウム	60	600	15	1.0
酢酸アンモニウム	40	300	19	1.2

## 共通条件

分析カラム:	ZIC-HILIC (粒子径 3.5 µm, 4.6 mm I.D×150 mm, Merck Millipore) 又は FIA
流速:	0.5 mL/min
モード:	アイソクラテック
注入量:	2 µL
分析時間:	FIA: 3 分 HPLC: 15 分
イオンソース:	ESI
イオン化モード:	ポジティブイオンモード
測定モード:	SIR $m/z = 152.1$
サンプリングレート:	10 points/sec

## 謝辞

本研究を行うにあたり、懇切なるご指導を賜りました静岡県立大学 薬学研究院 生体機能分子分析学講座 轟木 堅一郎 教授に心より感謝致します。また、本研究を遂行するにあたってご指導ご鞭撻を賜りました同講座 豊岡 利正 名誉教授、科学の視点に立った具体的なアドバイスを賜りました同講座 水野 初 准教授、本論文をまとめるにあたりにご指導を賜りました同講座 杉山 栄二 助教に厚く御礼申し上げます。

本論文の作成に際し、ご多忙の中、懇切丁寧なご校閲とご指導を賜りました静岡県立大学 薬学研究院 薬剤学講座 尾上 誠良 教授、同研究院 医薬生命科学講座 浅井 知浩 教授、同研究院 創剤工学講座 近藤 啓 教授に深く感謝の意を表します。

本研究に取り組む又とない機会を与えてくださり、遂行するに当たって多大なる御支援を頂きましたアステラス製薬株式会社 製薬技術本部 物性研究所 元永 圭 所長、同研究所 瀧本 直人 室長に深く感謝致します。また、本研究を遂行するにあたりデータサイエンスの知識をご教授いただいた同研究所 阿形 泰義 氏に心より感謝致します。

大学学部学生、修士学生時代に研究者精神を教えてくださった東北大学 大学院薬学研究科教授 大江 知行 博士、同研究科 李 宣和 博士に深謝いたします。

最後に、これまで私を温かく応援してくれた家族に心より感謝致します。

## 引用文献

- [1] T. McGovern, D. Jacobson-Kram, Regulation of genotoxic and carcinogenic impurities in drug substances and products, *TrAC - Trends Anal. Chem.* 25 (2006) 790–795. <https://doi.org/10.1016/j.trac.2006.06.004>.
- [2] M. De Moura, B. Van Houten, Mechanisms of DNA Damage, Repair, and Mutagenesis, *Environ. Mol. Mutagen.* 405 (2010) 391–405. <https://doi.org/10.1002/em>.
- [3] Guideline on the limits of genotoxic impurities, European Medicines Agency, 2004.
- [4] Genotoxic and carcinogenic impurities in drug substances and products: recommended approaches, Food and Drug Administration, 2008.
- [5] Assessment and Control of DNA Reactive (mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2015.
- [6] D.Q. Liu, M. Sun, A.S. Kord, Recent advances in trace analysis of pharmaceutical genotoxic impurities, *J. Pharm. Biomed. Anal.* 51 (2010) 999–1014. <https://doi.org/10.1016/j.jpba.2009.11.009>.
- [7] M.A. Raji, K.A. Schug, Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design, *Int. J. Mass Spectrom.* 279 (2009) 100–106. <https://doi.org/10.1016/j.ijms.2008.10.013>.
- [8] J. Liigand, A. Laaniste, A. Krüge, pH Effects on Electrospray Ionization Efficiency, *J. Am. Soc. Mass Spectrom.* 28 (2017) 461–469. <https://doi.org/10.1007/s13361-016-1563-1>.
- [9] A. Kiontke, A. Oliveira-Birkmeier, A. Opitz, C. Birkemeyer, Electrospray ionization efficiency is dependent on different molecular descriptors with respect to solvent pH and instrumental configuration, *PLoS One* 11 (2016) 1–16. <https://doi.org/10.1371/journal.pone.0167502>.
- [10] S. Caetano, T. Decaestecker, R. Put, M. Daszykowski, J. Van Bocxlaer, Y. Vander Heyden, Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors, *Anal. Chim. Acta* 550 (2005) 92–106. <https://doi.org/10.1016/j.aca.2005.06.069>.
- [11] F. Qiu, D.L. Norwood, Identification of pharmaceutical impurities, *J. Liq. Chromatogr. Relat. Technol.* 30 (2007) 877–935. <https://doi.org/10.1080/10826070701191151>.
- [12] M. D'Hondt, B. Gevaert, E. Wynendaele, B. De Spiegeleer, Implementation of a single quad MS detector in routine QC analysis of peptide drugs, *J. Pharm. Anal.* 6 (2016) 24–31. <https://doi.org/10.1016/j.jpba.2015.09.002>.
- [13] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research - Part 2, *Pharm. Sci. Technol. Today* 3 (2000) 50–57. <http://www.scopus.com/inward/record.url?eid=2-s2.0-0034142319&partnerID=40&md5=62d998ad6c004833dfcd8abb700538a8>.
- [14] S. Chinta, R. Rengaswamy, Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to Predict Drug Solubility in Binary Solvent Systems, *Ind. Eng. Chem. Res.* 58 (2019) 3082–3092. <https://doi.org/10.1021/acs.iecr.8b04584>.

- [15] Experimental Data of 3'-Amino-4'-methoxyacetanilide without Metabolic Activation, Ministry of Health, Labour and Welfare. <https://anzeninfo.mhlw.go.jp/user/anzen/kag/pdf/C/C6375-47-9.pdf> (Accessed 12 November 2020).
- [16] T. Watanabe, K. Tobe, Y. Nakachi, Y. Kondoh, M. Nakajima, S. Hamada, C. Namiki, T. Suzuki, S. Maeda, A. Tadakuma, M. Sakurai, Y. Arai, A. Hyogo, M. Hoshino, T. Tashiro, H. Ito, H. Inazumi, Y. Sakaki, H. Tashiro, C. Furihata, Differential Gene Expression Induced by Two Genotoxic N-nitroso Carcinogens, Phenobarbital and Ethanol in Mouse Liver Examined with Oligonucleotide Microarray and Quantitative Real-time PCR, *Genes Environ.* 29 (2007) 115–127. <https://doi.org/10.3123/jemsge.29.115>.
- [17] J.T. Behrens, Principles and procedures of exploratory data analysis., *Psychol. Methods* 2 (1997) 131–160. <https://doi.org/10.1037/1082-989X.2.2.131>.
- [18] R. Kostianinen, T.J. Kauppila, Effect of eluent on the ionization process in liquid chromatography-mass spectrometry, *J. Chromatogr. A* 1216 (2009) 685–699. <https://doi.org/10.1016/j.chroma.2008.08.095>.
- [19] M. Nshanian, R. Lakshmanan, H. Chen, R.R.O. Loo, J.A. Loo, Enhancing sensitivity of liquid chromatography–mass spectrometry of peptides and proteins using supercharging agents, *Int. J. Mass Spectrom.* 427 (2018) 157–164. <https://doi.org/10.1016/j.ijms.2017.12.006>.
- [20] A. Apffel, S. Fischer, G. Goldberg, P.C. Goodley, F.E. Kuhlmann, Enhanced sensitivity for peptide mapping with electrospray liquid chromatography-mass spectrometry in the presence of signal suppression due to trifluoroacetic acid-containing mobile phases, *J. Chromatogr. A* 712 (1995) 177–190. [https://doi.org/10.1016/0021-9673\(95\)00175-M](https://doi.org/10.1016/0021-9673(95)00175-M).
- [21] S. Zhou, M. Hamburger, Effects of Solvent Composition on Molecular Ion Response in Electrospray Mass Spectrometry : Investigation of the Ionization Processes 9 (1995) 1516–1521.
- [22] X. Liu, C.A. Pohl, HILIC behavior of a reversed-phase/cationexchange/ anion-exchange trimode column, *J. Sep. Sci.* 33 (2010) 779–786. <https://doi.org/10.1002/jssc.200900645>.
- [23] K. Granelli, C. Branzell, Rapid multi-residue screening of antibiotics in muscle and kidney by liquid chromatography-electrospray ionization-tandem mass spectrometry, *Anal. Chim. Acta* 586 (2007) 289–295. <https://doi.org/10.1016/j.aca.2006.12.014>.
- [24] R. Dams, T. Benijts, W. Günther, W. Lambert, A. De Leenheer, Influence of the eluent composition on the ionization efficiency for morphine of pneumatically assisted electrospray, atmospheric-pressure chemical ionization and sonic spray, *Rapid Commun. Mass Spectrom.* 16 (2002) 1072–1077. <https://doi.org/10.1002/rcm.683>.
- [25] A. Krueve, K. Kaupmees, J. Liigand, I. Leito, Negative electrospray ionization via deprotonation: Predicting the ionization efficiency, *Anal. Chem.* 86 (2014) 4822–4830. <https://doi.org/10.1021/ac404066v>.
- [26] S. Karmarkar, X. Yang, R. Garber, A. Szajkovichs, M. Koberda, Quality by design (QbD) based development and validation of an HPLC method for amiodarone hydrochloride and its impurities in the drug substance, *J. Pharm. Biomed. Anal.* 100 (2014) 167–174. <https://doi.org/10.1016/j.jpba.2014.07.002>.
- [27] A.P. Kumar, V.R.L. Ganesh, D.V.S. Rao, C. Anil, B.V. Rao, V.S. Hariharakrishnan, A. Suneetha, B.S. Sundar, A validated reversed phase HPLC method for the determination of process-related impurities in almotriptan malate API, *J. Pharm. Biomed. Anal.* 46 (2008) 792–798. <https://doi.org/10.1016/j.jpba.2007.11.029>.

- [28] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, *Chem. Rev.* 96 (1996) 1027–1043. <https://doi.org/10.1021/cr950202r>.
- [29] H. Baba, J.I. Takahara, F. Yamashita, M. Hashida, Modeling and Prediction of Solvent Effect on Human Skin Permeability using Support Vector Regression and Random Forest, *Pharm. Res.* 32 (2015) 3604–3617. <https://doi.org/10.1007/s11095-015-1720-4>.
- [30] S. Kawamura, M. Arakawa, K. Funatsu, Development of genetic algorithm-based wavelength regional selection technique, *J. Comput. Aided Chem.* 7 (2006) 10–17. <https://doi.org/10.2751/jcac.7.10>.
- [31] J. Hermans, S. Ongay, V. Markov, R. Bischoff, Physicochemical Parameters Affecting the Electrospray Ionization Efficiency of Amino Acids after Acylation, *Anal. Chem.* 89 (2017) 9159–9166. <https://doi.org/10.1021/acs.analchem.7b01899>.
- [32] W.S. Noble, What is a support vector machine?, *Nat. Biotechnol.* 24 (2006) 1565–1567. <https://doi.org/10.1038/nbt1206-1565>.
- [33] J. Samuelsson, F.F. Eiriksson, D. Åsberg, M. Thorsteinsdóttir, T. Fornstedt, Determining gradient conditions for peptide purification in RPLC with machine-learning-based retention time predictions, *J. Chromatogr. A* 1598 (2019) 92–100. <https://doi.org/10.1016/j.chroma.2019.03.043>.
- [34] J. Cui, Q. Chen, X. Dong, K. Shang, X. Qi, H. Cui, A matching algorithm with isotope distribution pattern in LC-MS based on support vector machine (SVM) learning model, *RSC Adv.* 9 (2019) 27874–27882. <https://doi.org/10.1039/c9ra03789f>.
- [35] M. Taningher, R. Pasquini, S. Bonatti, Genotoxicity analysis of N,N-dimethylaniline and N,N-dimethyl-p-toluidine, *Environ. Mol. Mutagen.* 21 (1993) 349–356. <https://doi.org/10.1002/em.2850210406>.
- [36] B. Buszewski, S. Noga, Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique, *Anal. Bioanal. Chem.* 402 (2012) 231–247. <https://doi.org/10.1007/s00216-011-5308-5>.
- [37] S. Arase, S. Kimura, T. Ikegami, Method optimization of hydrophilic interaction chromatography separation of nucleotides using design of experiment approaches I: Comparison of several zwitterionic columns, *J. Pharm. Biomed. Anal.* 158 (2018) 307–316. <https://doi.org/10.1016/j.jpba.2018.05.014>.
- [38] H.P. Nguyen, K.A. Schug, The advantages of ESI-MS detection in conjunction with HILIC mode separations: Fundamentals and applications, *J. Sep. Sci.* 31 (2008) 1465–1480. <https://doi.org/10.1002/jssc.200700630>.
- [39] D. V. McCalley, Effect of mobile phase additives on solute retention at low aqueous pH in hydrophilic interaction liquid chromatography, *J. Chromatogr. A* 1483 (2017) 71–79. <https://doi.org/10.1016/j.chroma.2016.12.035>.
- [40] A. Periat, I. Kohler, A. Bugey, S. Bieri, F. Versace, C. Staub, D. Guilleme, Hydrophilic interaction chromatography versus reversed phase liquid chromatography coupled to mass spectrometry: Effect of electrospray ionization source geometry on sensitivity, *J. Chromatogr. A* 1356 (2014) 211–220. <https://doi.org/10.1016/j.chroma.2014.06.066>.
- [41] P. Labute, LowModeMD - Implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops, *J. Chem. Inf. Model* 50 (2010) 792–800. <https://doi.org/10.1021/ci900508k>.

- [42] J. Xu, Y.Y. Tang, B. Zou, Z. Xu, L. Li, Y. Lu, Generalization performance of Gaussian kernels SVMC based on Markov sampling, *Neural Networks* 53 (2014) 40–51.  
<https://doi.org/10.1016/j.neunet.2014.01.013>.