Wydział Chemii

Antoni Bilicki

Modelowanie profilu aminokwasowego w pyłkach pszczelich na bazie widm FTIR ATR z zastosowaniem techniki automatycznej optymalizacji zakresów widmowych

Praca wykonana pod opieką:
dr hab. Sylwestra Mazurka
w Zespole Chemometrii i Spektroskopii Stosowanej

2022

Spis treści

Wykaz skrótów i oznaczeń

API Interfejs programowalny aplikacji

ATR Spektroskopia osłabionego całkowitego odbicia

CE Elektroforeza kapilarna

CLS Metoda klasycznych najmniejszych kwadratów

CV Walidacja krzyżowa

EM Promieniowanie elektromagnetyczne

GC-MS Chromatografia gazowa sprzężona ze spektrometrią mas

HPLC Wysokosprawna chromatografia cieczowa

ICP-AES Atomowa spektroskopia emisyjna indukcyjnie sparowanej plazmy

ILS Regresja odwrotnych najmniejszych kwadratów

iPLS Interwałowa metoda selekcji danych PLS

IR Spektroskopia w zakresie podczerwieni

L-B Prawo Lamberta-Beera

LC-MS Chromatografia cieczowa sprzężona ze spektrometrią mas

LOO Walidacja krzyżowa pozostawienia jednej próbki

MLR Wielokrotna regresja liniowa

MSC Multiplikatywna korekcja rozproszenia

NIPALS Nieliniowy iteratywny algorytm PLS

NIR Spektroskopia w zakresie bliskiej podczerwieni

NMR Jądrowy rezonans magnetyczny

PCA Analiza głównych składowych

PCR Regresja głównych składowych

PLS Regresja częściowych najmniejszych kwadratów

PLSTB Oprogramowanie PLS Toolbox

PRESS Przewidywana reszta sum kwadratów błędów

RMSECV Pierwiastek średniego błędu standardowego walidacji krzyżowej

RMSEP Pierwiastek średniego błędu kwadratowego przewidywania

RSEP Względny standardowy błąd przewidywania

RSEPV Względny standardowy błąd walidacji dla próbek walidacyjnych

SIMPLS Algorytm PLS poddany modyfikacjom inspirowanym statystycznie

TLC Chromatografia cienkowarstwowa

TQ Oprogramowanie Turbo Quant Analyst

UPLC Ultrasprawna chromatografia cieczowa

UV-VIS Spektroskopia w ultrafiolecie i zakresie widzialnym

VIP Waga projektowanej zmiennej

Rozdział 1

Cel Pracy

Celem poniższej pracy jest opracowanie nowej metody selekcji danych, mogącej znaleźć zastosowanie w szybkiej analizie ilościowej naturalnych układów wieloskładnikowych na bazie danych spektroskopowych. Przedmiotem badań jest oznaczenie profilu aminokwasowego pyłków pszczelich, których widma zarejestrowano przy pomocy techniki FTIR-ATR. Zaproponowana metoda zostanie oparta o modelowanie PLS z zastosowaniem własnego algorytmu heurystycznego poprawiającego proces selekcji danych.

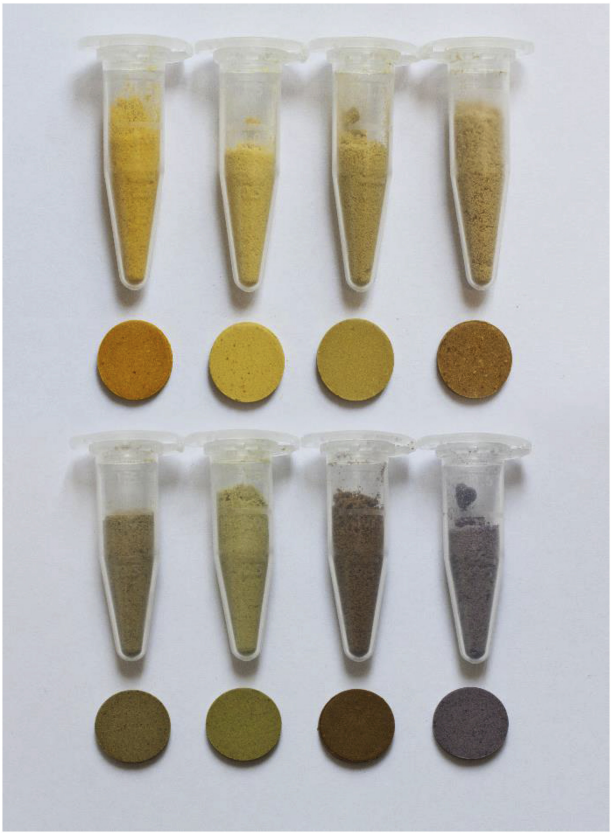
Rozdział 2

Wstęp teoretyczny

2.1 Wprowadzenie w tematykę

2.1.1 Pyłek kwiatowy

Pyłek kwiatowy jest naturalnym produktem zbieranym przez pszczoły miodne z kwiatów roślin kwitnących w okolicy ula. Zawiera on związki niezbędne do produkcji mleczka pszczelego, którym karmione są larwy w ulu przez pierwsze dni swojego życia oraz jest niezbędnym elementem diety dorosłych pszczół w połączeniu z miodem i nektarem [5]. Pyłek pszczeli jest zbierany przez robotnice na odnóżach w postaci granulatu, które są przechwytywane przez pszczelarzy przy wejściu do ula w tak zwanych "pułapkach". Granulat ten jest zbitym, kruchym ciałem stałym o lekko woskowej konsystencji. Jak przedstawiono na rysunku 2.1, po wysuszeniu i rozbiciu przybiera on formę proszku. Każde ziarnko granulatu posiada swój specyficzny kolor, morfologię, skład i smak zależny od gatunków roślin, z których zostało pozyskane przez pszczołę.



Rysunek 2.1: Przykładowe próbki pyłku kwiatowego użyte do wcześniejszych badań [4]

Pyłek zawsze zawiera zestaw substancji niezbędnych do prawidłowego wzrostu larw - cukrów, aminokwasów i soli mineralnych [6]. Jak pokazano w poprzedniej pracy [7], skład chemiczny pyłków jest bardzo zróżnicowany. Poniżej wymieniono zakres stężeniowy wybranych składników obecnych w analizowanych próbkach w toku poprzednich badań:

- tłuszcze 7-9% masowych,
- cukry 25-45% masowych,
- białka 15-26% masowych,
- polifenole 4-14% masowych,
- wilgotność 6-12% masowych.

Tradycyjnie oznaczanie składu wykonuje się z zastosowaniem szeregu metod opartych na ekstrakcji, koncentracji i analizie poszczególnych składników chemicznych. Do badania ogólnej zawartości białek stosuje się testy Bradforda [8] lub ekstrakcję z użyciem dedykowanej aparatury do analizy ilościowej. Wyznaczenie zawartości tłuszczu wykonuje się poprzez ekstrakcję Soxhleta łączoną z chromatografią gazową, a do analizy minerałów stosuje się ICP-AES po wcześniejszej mineralizacji [9].

Ustalenie składu złożonych produktów naturalnych jest więc skomplikowaną i kosztowną procedurą, która dodatkowo jest narażona na wiele potencjalnych błędów pomiarowych związanych zarówno z czynnikiem ludzkim, jak i użytą aparaturą. Cechy te czynią z pyłku pszczelego idealną mieszaninę modelową do zaproponowania nowej procedury analizy układów naturalnych, która pozwoli rozwiązać wyżej wymienione problemy.

W tej pracy położono nacisk na badanie składu aminokwasowego bez wcześniejszej preparatyki próbek pyłku. Aminokwasy zawarte w badanych próbkach będą dalej związane w strukturach II i III rzędowych wewnątrz białek - pozwoli to na uniknięcie konieczności przeprowadzenia hydrolizy. Takie podejście pozwala na krok w stronę jeszcze większej dokładności procedur analitycznych dla próbek naturalnych; usuwa problem strat podczas ekstrakcji, które często są trudne do ustalenia, a także ewentualnych strat powstałych podczas hydrolizy białek do aminokwasów. Zaproponowana metoda dodatkowo pozwoli na oznaczenie aminokwasów, niezależnie od tego w jaki sposób zostały biologicznie zmodyfikowane, przy założeniu, że zostały one wykryte przez analizy referencyjne.

2.1.2 Chemometria

Chemometria jest gałęzią chemii odpowiedzialną za analizę, wizualizację i interpretację informacji z danych pomiarowych. Wraz ze stworzeniem automatycznych metod analitycznych na przełomie XX i XXI wieku pojawiła się potrzeba opracowania metodologii przetwarzania znacznej ilości danych otrzymywanych z nowych aparatur. Utworzono w tym

celu nową dziedzinę chemii, skupiającą się na opracowaniu nowych technik i algorytmów zwiększających precyzję pomiarów analitycznych. Stosując chemoinformatykę i statystykę chemometria oferuje nowe możliwości uzyskiwania informacji z wcześniej niedostępnych zmiennych i fluktuacji pomiarowych, przyspieszając tym samym rozwój analityki i automatyzacji procesów w przemyśle [10].

Procedura chemometryczna budowy modelu opiera się na trzech podstawowych etapach:

- preprocessingu,
- kalibracji,
- weryfikacji.

Otrzymany model chemometryczny pozwala na analizę jakościową lub oznaczanie wybranych parametrów próbki. Poniżej zostanie omówiony wstępnie każdy z tych etapów.

Preprocessing

Preprocessing jest etapem, w którym przygotowuje się dane do dalszej analizy. Jego celem jest dobór odpowiednich technik obróbki widm, ich korekty i normalizacji. Bardzo często od preprocessingu zależy jakość otrzymanego modelu kalibracyjnego.

Sam preprocessing można podzielić na trzy grupy czynności:
1) Selekcja wstępna

- **Podział próbek:** Dla danych pomiarowych przeprowadza się PCA (Analiza głównych składowych), co pozwala na identyfikację i odrzucenie próbek odstających oraz wstępną analizę trendów występujących w badanej serii danych. Następnie próbki dzieli się na zestaw kalibracyjny i walidacyjny, które zostaną użyte w dalszych etapach modelowania.
- **Wybór zakresów:** Na tym etapie można odrzucić zakresy spektralne o znikomym wkładzie w jakość otrzymanego modelu, np. region absorpcji filtrów, naczyń pomiarowych czy kryształu ATR (Spektroskopia osłabionego całkowitego odbicia).

2) Operacje na widmie

- **Korekta linii bazowej:** pozwala na usunięcie udziałów widmowych kuwety, użytego rozpuszczalnika lub kryształu ATR, na którym znajduje się próbka. Korekta ta jest szczególnie użyteczna w przypadku analiz danych NIR i Ramana.

- **Korekta MSC:** algorytm umożliwiający zniwelowanie wpływu rozproszenia światła na cząsteczkach substancji stałej lub emulsji [11].

- **Wygładzanie widma:** pozwala na kompensację skoków intensywności, wynikających z niskiej rozdzielczości oraz redukcję szumu. Najczęściej stosowaną metodą wygładzania jest procedura Savitzky-Golaya, opierająca się na kroczącej aproksymacji wielomianowej kolejnych odcinków widma [12].

- **Obliczanie pochodnych:** różniczkowanie widm pozwala na usunięcie zmienności związanej z przebiegiem linii bazowych, jest podstawową techniką zwiększania zdolności rozdzielczej danych.

3) Korekty danych

- **Centrowanie widma:** niezbędny etap do przeprowadzenia procedury PLS (Regresja częściowych najmniejszych kwadratów), który polega na odjęciu widma średniego od widm poszczególnych próbek, co pozwala na usunięcie redundancji.

- **Auto-skalowanie:** procedura normalizująca, która wyrównuje wkład zmiennych różniących się intensywnością. Odbywa się poprzez przeprowadzenie centrowania, a następnie podzielenie każdej kolumny danych przez wartość jej odchylenia standardowego.

Kalibracja

Kalibracja to kluczowy etap budowy modelu, którego konsekwencją jest ustalenie relacji matematycznej między wprowadzonymi zmiennymi zależnymi a niezależnymi. Działanie większości algorytmów predykcyjnych używanych w chemometrii opiera się na podobnej procedurze - sprowadzają się one do znalezienia określonego typu zależności, zwykle liniowej, poprzez zastosowanie metod regresji. Różnice między poszczególnymi algorytmami kalibracji wynikają z ich implementacji regresji, typu danych wejściowych oraz dodatkowych funkcji, takich jak grupowanie na podstawie wspólnych cech.

W modelowaniu ilościowym PLS w wyniku kalibracji otrzymuje się wstępny model charakteryzowany przez podstawowe parametry jakościowe. Określają one stopień dopasowania pomiędzy danymi wejściowymi a tymi przewidzianymi przez model oraz informacje o zasobie zmienności reprezentowanej przez każdy użyty w modelowaniu faktor.

Weryfikacja

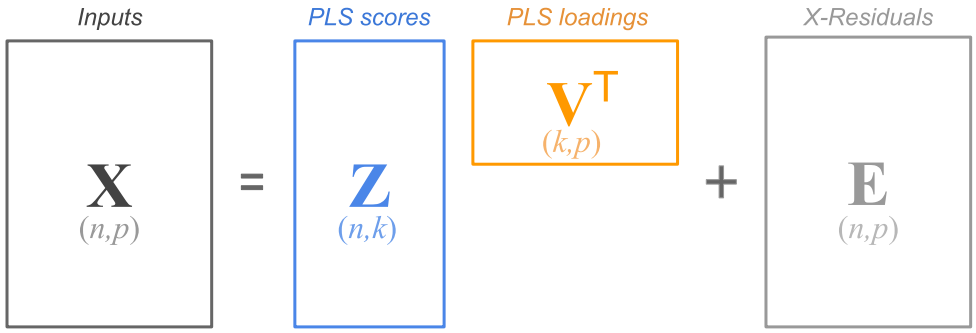
Ostatnim etapem budowy modelu jest jego walidacja, podczas której weryfikuje się jakość oznaczeń próbek o nieznanym składzie. Do modelu wprowadza się próbki walidacyjne,

nie uwzględnione w procesie kalibracji, dokonuje prognozy ich wartości i porównuje wyniki z analizą referencyjną. Pozwala to na ustalenie rzeczywistej jakości prognostycznej modelu, w tym skuteczności przewidywania i dokładności z jaką ją przeprowadzono. Po pomyślnym przejściu weryfikacji model jest gotowy do zastosowania w przewidywaniu parametrów nowych próbek.

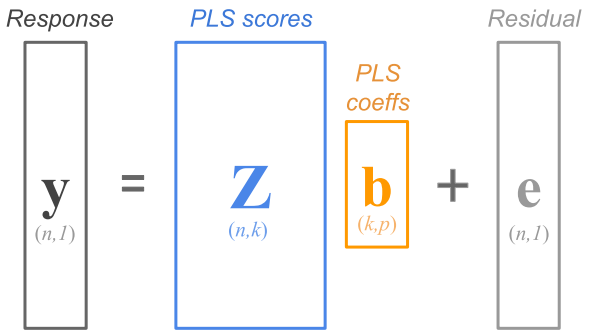
Regresja Cząstkowych Najmniejszych Kwadratów (PLS)

PLS to statystyczne podejście do modelowania relacji pomiędzy wynikami pomiarów, zawierającymi widoczne i ukryte zmienne, a badaną odpowiedzią systemu. Metoda ta należy do technik tzw. "modelowania miękkiego". Oznacza to, że charakteryzuje się większym stopniem elastyczności względem innych algorytmów ilościowych, co pozwala jej operować w sytuacjach gdzie nie są spełnione wszystkie warunki niezbędne do użycia innych typów analiz [13]. Jedną z głównych zalet PLS jest możliwość wykonywania analiz wielowymiarowych, tj. takich gdzie ustala się wiele parametrów badanej próbki na podstawie jednego pomiaru.

W niniejszej pracy PLS(SIMPLS) użyto jako główne narzędzie do modelowania stężeń aminokwasów. W przeciwieństwie do CLS (Metoda klasycznych najmniejszych kwadratów), które wymaga dokładnej znajomości widm składników zawartych w próbce i ich interakcji, i ILS (Regresja odwrotnych najmniejszych kwadratów), które pozwala na modelowanie próbek jedynie w momencie gdy ilość próbek przekracza ilość punktów pomiarowych, PLS maksymalizuje kowariancje między obserwowanymi zmiennymi a odpowiedzią systemu. To podejście sprawia, że metoda ta doskonale nadaje się do modelowania układów, w których występuje współliniowa zależność między widmami jej składników. Te cechy sprawiły, że modele PLS są obecnie często wykorzystywane w przemyśle do automatycznych testów pomiarów jakości produktów spożywczych, farmaceutyków, a także wyrobów przemysłu chemicznego.

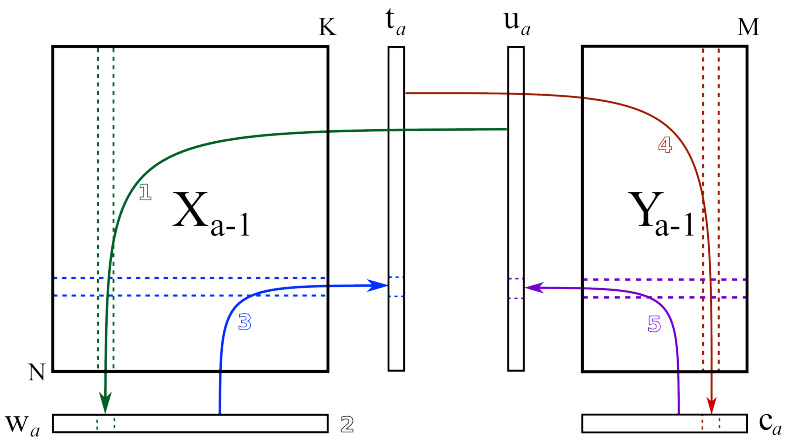


Rysunek 2.2: Diagram rozkładu macierzy widma(X) na macierze współrzędnych(Z), ładunków(V) i pozostałości (E), ustalony relacją $X = Z * V^T + E$ [14]



Rysunek 2.3: Diagram rozkładu macierzy odpowiedzi(Y) na macierze współrzędnych(Z), współczynników(B) i pozostałości (E), ustalony relacją $X = Z * V^T + E$ [14]

Działanie PLS łączy metody PCR (Regresja głównych składowych) i MLR (Wielokrotna regresja liniowa). Tradycyjnie, metoda PCR stara się utworzyć faktory które jak najlepiej opisują całość zmienności zawartej w modelowanych danych. W PLS podejście to zostaje zmodyfikowane tak, aby w faktorach została zawarta jedynie zmienność użyteczna w przewidywaniu zmiennej niezależnej. W przeciwieństwie do PCR, obliczanie zmiennych niezależnych w modelowaniu PLS dla nowych danych nie jest jedynie projekcją ładunków na macierz danych wejściowych, ale jest także uwzględniony wpływ obliczonych wag poszczególnych zmiennych, co pozwala na redukcję wpływu szumu i danych nieinformatywnych [15]. Kreacja modelu PLS odbywa się poprzez użycie jednego z dwóch algorytmów - NIPALS (Nieliniowy iteratywny algorytm PLS), przedstawionego na rysunku 2.4, lub SIMPLS (Algorytm PLS poddany modyfikacjom inspirowanym statystycznie). W tej pracy użyto SIMPLS ze względu na dużo krótszy czas potrzebny na budowę modelu, co przyspieszyło optymalizację i pozwoliło na uzyskanie modeli o lepszej jakości. Jak pokazano na rysunku 2.2 i 2.3, działanie PLS opiera się na rozkładzie macierzowym danych i ustaleniu ich wzajemnej relacji.



Rysunek 2.4: Diagram przedstawiający działanie algorytmu NIPALS

1. Przeprowadzenie regresji kolumn X na wektor u, zapisanie współczynników regresji w wektorze w.
2. Normalizacja wektora wag w.
3. Regresja każdego rzędu X na wektor wag. Zapisanie współczynników regresji w wektorze t.
4. Regresja kolumn Y na wektor współrzędnych t, zapisanie współczynników regresji w wektorze c.
5. Regresja rzędów Y na wektor wag c, zapisanie współczynników krzywych regresji w wektorze u.
6. Powtórzenie poprzednich etapów do zaniku zmian w wektorze u.

W wyniku powyższej procedury otrzymuje się model predykcyjny dla jednego składnika a [14]

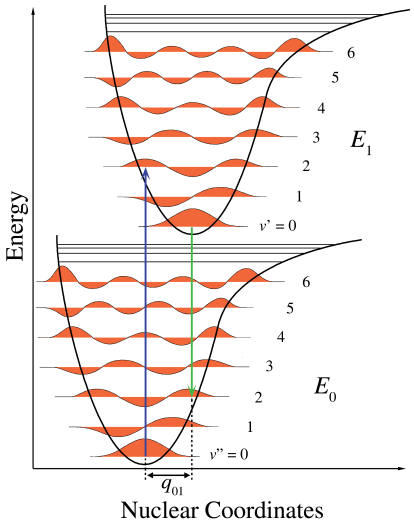
Klasycznie wyróżnia się także podział na PLS1 oraz PLS2. PLS1 odnosi się do algorytmu, w którym modeluje się zależność danych X między jedną zmienną obserwowaną Y, podczas gdy PLS2 jest w stanie przeprowadzać analizę wielowymiarową, modelując zależność między wieloma odpowiedziami Y_1,Y_2,Y_3,...,Y_n, poprzez uwzględnienie ich wzajemnych wpływów i brak pełnej ortogonalności.

2.1.3 Spektroskopia IR [1]

Spektroskopia w średniej podczerwieni jest metodą analityczną, w której bada się oddziaływanie materii z promieniowaniem z zakresu 2,5μm - 25 μm. W przeciwieństwie do spektroskopii w zakresie widzialnym, która wywołuje wzbudzenie elektronów, promieniowanie w zakresie podczerwonym absorbowane jest przez wiązania chemiczne wewnątrz cząsteczek. Absorpcja promieniowania IR (Spektroskopia w zakresie podczerwieni) powoduje wzrost energii kinetycznej atomów w cząsteczkach, na skutek tego spektroskopia IR jest nazywana też spektroskopią oscylacyjną (do tej grupy metod zaliczają się też spektroskopia Ramana i rozproszenia neutronów). Cecha ta sprawia, że spektroskopia IR jest zaliczana do jednej z najbardziej uniwersalnych metod analitycznych, zdolnych do badania niemal każdej grupy substancji.

Zasada działania

Zgodnie z zasadami fizyki kwantowej, cząsteczka ma dyskretne stany energetyczne dla ruchu elektronów, wibracji atomów i ich rotacji. Aby nastąpiła absorpcja promieniowania, energia fotonu musi być zbliżona różnicy obecnego i dostępnego stanu energetycznego występującego w cząsteczce. Dodatkowo muszą zostać spełnione reguły wyboru. Dzielą one przejścia na dozwolone i wzbronione, gdzie przejścia dozwolone charakteryzują się dużo wyższą intensywnością z faktu większego prawdopodobieństwa ich zajścia. Szansa zajścia przejścia elektronowego może być też opisana regułą Franka-Condona, przedstawioną na rysunku 2.5.



Rysunek 2.5: Diagram Franka-Condona obrazujący prawdopodobieństwo przejść elektronowych. Prawdopodobieństwo przejścia rośnie, gdy odpowiada mu stan o niższej energii kinetycznej w momencie przejścia

W trakcie pomiarów spektroskopowych następuje ustalenie, jakie długości fali (i jak silnie) są absorbowane przez cząsteczkę i przedstawia się wyniki na widmie. Długości fali korespondują z odpowiednimi przejściami energetycznymi zachodzącymi w cząsteczce, co z kolei daje informację o jej budowie lub wręcz pozwala na jej identyfikacje na podstawie sygnałów występujących w określonych obszarach widma. Dodatkowo, zależność maksimum absorpcji od stężenia jest ściśle opisana przez prawo Lamberta-Beera,

$$T(\vartheta) = 10^{-\varepsilon c l}, \tag{2.1}$$

gdzie:

$T(\vartheta)$: transmitancja,

ε: molowy współczynnik absorpcji,

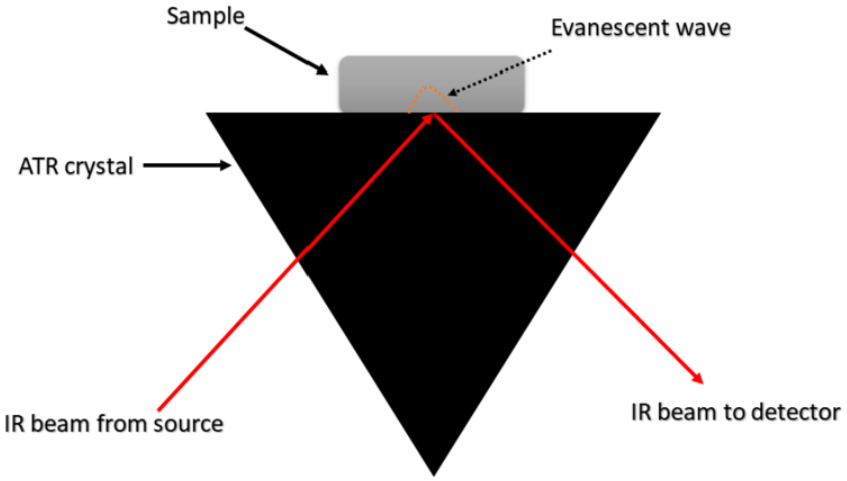
c: stężenie molowe,

l: długość drogi optycznej w roztworze,

co umożliwia ilościową analizę składu badanej próbki. Należy jednak pamiętać, że transmitancja nie jest proporcjonalna do stężenia i należy ją przeliczyć na absorbancje lub użyć drugiej formy prawa L-B (Prawo Lamberta-Beera), $A(\vartheta) = \varepsilon c l$ oraz że różne przejścia charakteryzują się z natury inną, typową dla nich intensywnością. Wynika z tego, że prawo L-B ma zastosowanie jedynie przy obserwacji jednego typu przejścia dla jednego typu cząsteczki, chyba że widmo zostanie najpierw poddane odpowiedniej normalizacji.

2.1.4 Spektroskopia ATR

Spektroskopia całkowitego osłabionego odbicia jest jedną z najpopularniejszych technik pomiarowych w podczerwieni. Do badań stosuje się podstawkę zawierającą kryształu diamentu, germanu lub krzemu. Pozwala to na ułatwienie całej procedury pomiarowej - nie ma tutaj potrzeby mieszania próbki z bromkiem potasu i przygotowania pastylki - wystarczy jedynie umieścić badaną próbkę na krysztale i przeprowadzić pomiar. Takie podejście nie tylko zwiększa powtarzalność, ale także pozwala na ponowne użycie próbki i przyspiesza pomiary. Zalety te sprawiły, że ATR jest obecnie dominującą techniką pomiarową IR stosowaną w badaniach naukowych i przemysłowych [1]. Zasada działania tej metody została przedstawiona na rysunku 2.6.



Rysunek 2.6: Diagram przedstawiający zasadę funkcjonowania jedno-odbiciowej przystawki ATR. Na próbkę umieszczoną na krysztale pada wiązka promieniowania, generując falę zanikającą, która przenika w głąb próbki. Jeżeli nie następuje absorpcja, fala jest następnie odbijana w całości do detektora. W innym przypadku następuje zmniejszenie transmitancji o wartość zaabsorbowaną w próbce [16]

2.1.5 Profil aminokwasowy

Aminokwasy to jedna z podstawowych grup związków, o które oparte jest całe życie na Ziemi. Produkt ich polimeryzacji, białka, stanowią obecnie jeden z głównych przedmiotów badań zespołów biomedycznych [17]. Ustalenie zawartości poszczególnych białek w próbce pozwala na przewidywanie kluczowych parametrów dotyczących organizmów, z których pochodzą - wieku, jednostek chorobowych, a także pochodzenia oraz jakości, w przypadku produktów spożywczych. Badanie profilu aminokwasowego pozwala na ustalenie części z tych parametrów bez potrzeby bezpośredniego sekwencjonowania aminokwasów, które jest niezwykle czasochłonne i kosztowne.

Przykładowo - wysoka zawartość glicyny informuje o potencjalnie dużej zawartości helis w badanych białkach; może to wskazywać na obecność białek kolagenu, transbłonowych

lub wiążących DNA.

Aminokwasy tradycyjnie dzieli się na egzo- i endogenne, z których te pierwsze nie mogą być naturalnie syntetyzowane w organizmie człowieka. Z tego powodu czasem aminokwasy egzogenne nazywa się także aminokwasami kluczowymi (essential).

Wraz z rozwojem biotechnologii i genetyki wzrasta potrzeba szybkiego i taniego ustalania profilu aminokwasowego badanych próbek. W tabeli 2.1 przedstawiono porównanie najczęściej używanych metod analizy profilu aminokwasowego.

Properties / Method	Simple	convenient	Cheap	Effective	Sensitivity	Fast time	cost-effective	Wide range	Repeatability	Qualitative	Quantitative	Diadvantages
TLC	+	+	+	+			+			+		No quantitative Poor identification ability
HPLC				+	+		++	+	+	+	+	Extra column effect
LC-MS		+		++	++	+		+		+	+	Poor differentiation of isomers and stereochemistry ion source pollution
GC-MS	+			++	++					+	+	Samples can vaporize and ionize many isomers (especially positional isomerism) cannot be distinguished
CE				+	++	+		+	+	+	+	Weak separation ability high pH requirement
NMR								+		+	+	Expensive
Amino acid analyzer	+			+				+		+	+	Expensive
Electrochemical sensor	+	+				+				+	+	

+ More
++ High

Tablica 2.1: Tabela zawierająca obecnie stosowane metody ustalania profilu aminokwasowego [18]. "+"w tabeli oznacza przewagę metody w danej kategorii względem innych

TLC

Dzięki swojej niskiej cenie, szybkości i prostocie [19] TLC (Chromatografia cienkowarstwowa) jest najczęściej używaną metodą ustalania zawartości aminokwasów. Metoda ta opiera się na hydrolizie białek i peptydów, a następnie naniesieniu ich na płytki pokryte związkiem retencyjnym i porównanie wyników ze wzorcem. Pozwala ona szybko ustalić przybliżony skład aminokwasów w próbce, jednak daje ona słabe rezultaty ilościowe i nie rozróżnia aminokwasów o podobnych współczynnikach retencji.

HPLC

Obecnie najszybciej rozwijająca się z wymienionych metod, HPLC (Wysokosprawna chromatografia cieczowa), jest skuteczną metodą zarówno analizy ilościowej jak i jakościowej

profilu aminokwasowego. Jej głównymi wadami są koszt aparatury (oraz stały koszt operacyjny związany z zakupem eluentów), kolumn i konieczność hydrolizy białek, oferuje jednak dużą szybkość pomiaru przy relatywnej prostocie obsługi i analizy wyników.

LC-MS

Klasycznie metoda HPLC używa do detekcji aparatów opartych o spektroskopię UV-VIS [20–22]. W razie konieczności zwiększenia precyzji pomiaru mogą one być zastąpione przez spektrometry mas [23, 24]. Użycie bardziej czułego detektora pozwala także na zastosowanie kolumn o mniejszej średnicy pod zwiększonym ciśnieniem UPLC (Ultrasprawna chromatografia cieczowa). Znacząco zwiększa to koszt całej aparatury, oferuje jednak wyjątkową czułość, której dorównuje tylko kolejna z wymienionych metod [25].

GC-MS

Jako najczulsza z wymienionych metod, GC-MS (Chromatografia gazowa sprzężona ze spektrometrią mas) jest powszechnie używana do wykonywania testów na próbkach krwi podczas testów antydopingowych czy w wykrywaniu jednostek chorobowych. Dodatkowo znajduje zastosowanie w wykrywaniu zanieczyszczeń w lekach oraz kryminalistyce sądowej. Główną przeszkodą w zastosowaniu jej na szerszą skalę jest konieczność użycia dużego stopnia derywatyzacji, zwykle metylacji, aby umożliwić przejście próbek do fazy gazowej. Konsekwencją tego jest duża podatność na błąd ludzki, stosunkowo słaba powtarzalność (brak możliwości zagwarantowania pełnego zajścia derywatyzacji) i duży koszt analiz. Dodatkowo nie wszystkie aminokwasy są podatne na derywatyzację, co sprawia że metoda ta jest zwykle używana jedynie do ustalania 6-16 aminokwasów [26, 27].

CE

Elektroforeza kapilarna jest metodą podobną do chromatografii, jednak w przeciwieństwie do czynnika powodującego retencje, stosuje się tutaj zewnętrznie przyłożone pole elektromagnetyczne, które wymusza ruch aminokwasów przez kapilarę. W zależności od czynników takich jak masa, naładowanie i powinowactwo do powierzchni kapilary aminokwasy zostają rozdzielone według swoich współczynników retencji, a następnie zbadane za pomocą analizatorów odpowiadających technice HPLC.

NMR

Jądrowy rezonans magnetyczny jest prawdopodobnie jedną z najbardziej złożonych technik analizy aminokwasów, jak i najbardziej kosztownych. Charakteryzuje się jednak potencjałem, którego nie osiągają inne metody - jest zdolny do badania i ustalania rze-

czywistej struktury badanych związków. Pozwala to nie tylko na precyzyjne badanie biomodyfikowanych aminokwasów, ale także na ustalanie struktur całych peptydów i potencjalnie mniejszych białek [28, 29]. Dodatkową zaletą jest brak konieczności separacji aminokwasów w przypadku prostszych mieszanin, co przyspiesza analizę i zmniejsza ryzyko błędu [30].

Analizator aminokwasów

Analizator aminokwasów to kompaktowe urządzenie dedykowane do badań na aminokwasach. Jest to zautomatyzowana wersja zmodyfikowanej aparatury do HPLC, często z detektorem UV i kolumnami z modyfikowaną powierzchnią [31]. Jego ukierunkowane zastosowanie pozwala na uproszczenie obsługi i zmniejszenie kosztów użytkowania względem zwykłego podejścia chromatograficznego.

Sensor bioelektrochemiczny

Sensory to dedykowane czujniki zdolne do wykrywania ściśle określonych grup aminokwasów, operujące na zasadzie substrat - receptor. Pozwalają na osiągnięcie nadprecedensowej czułości kosztem dużej ceny i wąskiego zakresu analizowanych substancji [32].

Spektroskopia w podczerwieni

Spektroskopia IR ma szereg zalet względem poprzednich metod, jednak jej najważniejszą cechą jest możliwość badania próbek w stanie surowym. Pozwala to nie tylko na zaoszczędzenie materiału badawczego, ale także ułatwia procedurę pomiarową, skracając czas oraz zmniejszając koszt analiz. Wariację techniki z użyciem podstawki ATR zastosowano do ustalania składu aminokwasowego różnych typów próbek (serum, zwłok, tuńczyka) [33–35].

Większą popularnością cieszy się jednak analiza NIR (Spektroskopia w zakresie bliskiej podczerwieni), którą z powodzeniem zastosowano do ustalania składu aminokwasowego soi, pszenicy, traw, serów, nasion, orzeszków ziemnych [36–46].

Innymi metodami spektroskopowymi używanymi do ustalania profilu aminokwasowego są spektroskopia Ramana, zastosowana do badania profilu aminokwasowego linii komórkowych produkujących przeciwciała [47] oraz spektroskopia terahertzowa, użyta do analizy profilu aminokwasowego płatków śniadaniowych [48].

2.1.6 Badania pyłku pszczelego

Wcześniejsze badania o podobnej tematyce przeprowadzano jedynie w bardzo ograniczonym zakresie. Technika ATR była wielokrotnie używana do ustalania ogólnej zawartości

białka w próbkach pyłku pszczelego [49–51], obok oznaczeń innych składników pyłku, jak wilgotność, zawartość polifenoli i cukrów. Nie znaleziono jednak prac, w których dokonano pełnej analizy profilu aminokwasowego z użyciem technik spektroskopii oscylacyjnej.

W przypadku badań profilu aminokwasowego pyłków najczęściej stosowaną techniką analityczną jest chromatografia MS-HPLC [52, 53] lub TLC.

2.2 Metody selekcji danych stosowane w modelowaniu PLS

Tradycyjne podejście do PLS opiera się na użyciu pełnego zakresu danych, w tym przypadku widma. Taka metodologia nie jest doskonała - obecność szumów i zakresów nieinformatywnych może skutkować obniżeniem jakości i zdolności prognostycznej otrzymanych modeli. Obecnie wszystkie popularne implementacje PLS charakteryzują się dodatkowymi narzędziami selekcji danych, czyli automatycznego procesu wybrania najbardziej informatywnych zakresów zmiennych zależnych, które następnie zostają użyte do stworzenia modelu. W tym dziale zostaną krótko omówione najbardziej popularne metody selekcji danych używane w modelowaniu danych spektroskopowych, które w dalszych częściach pracy porównano do zaproponowanej, nowej implementacji.

2.2.1 Ręczna selekcja danych

Najprostsza z używanych metod selekcji, opiera się na wiedzy i intuicji twórcy modelu. Z widma oscylacyjnego zostają usunięte te zakresy, które według badacza nie będą zawierały wartościowych informacji do tworzenia modelu opisującego dany parametr. Przykładowo, mogą zostać usunięte obszary zawierające pasma substancji niepożądanych (np. wody), zakresy absorpcji filtra, kuwety lub kryształu oraz potencjalnych zanieczyszczeń. Podstawową wadą takiego podejścia jest wymaganie posiadania wiedzy dotyczącej próbki i aparatury, a także większa ilość czasu potrzebna na prawidłowe zidentyfikowanie pasm pochodzących od w.w. czynników.

2.2.2 Sieci neuronowe [2]

Sieci neuronowe to algorytmy mimikujące działanie ludzkiego mózgu. Składają się z dużej liczby podjednostek (neuronów), które są połączone siecią relacji opartą o podejmowanie decyzji na bazie wag. Złożoność otrzymanej relacji skutkuje jednak brakiem możliwości wglądu i interpretacji procesu jej tworzenia. Oznacza to, że poza metodami zewnętrznej walidacji jest niezwykle trudno ustalić przyczyny decydujące o selekcji wybranych zakresów. Istnieje szereg wariacji tej metody, poniżej opisano dwie najpopularniejsze:

Wstecznie propagowana sieć neuronowa

Ten typ sieci neuronowej charakteryzuje się zastosowaniem dodatkowego algorytmu rekursywnego, który zawraca otrzymane wyniki do warstwy wejściowej, iteracyjnie zwiększając precyzję otrzymanych rezultatów.

Proces ten składa się z trzech etapów:

1. Determinacja liczby neuronów
   Liczba neuronów w warstwie ukrytej powinna być ściśle skorelowana z liczbą wymiarów reprezentowanych przez dane wejściowe. W tym celu przeprowadza się PCA, a następnie dobiera liczbę neuronów, która pozwoli dobrze opisać około 80% zmienności zawartej w danych.

2. Obliczenie wag neuronów w warstwie ukrytej
   Poprzez zastosowanie adekwatnego algorytmu uczącego przypisuje się wagi odpowiedziom poszczególnych neuronów w sieci, tworząc strukturę dobrze opisującą zależność między parametrami wejściowymi a odpowiedzią (jakością otrzymanego modelu).

3. Wybór danych wejściowych
   Na tym etapie bada się udział każdej cechy (np. liczby falowej) na jakość otrzymanego modelu i usuwa się te o niskim lub negatywnym wpływie. Rezultatem jest otrzymanie zakresów liczb falowych o dobrej korelacji z modelowanymi parametrami.

2.2.3 Algorytmy genetyczne

Jest to klasa heurystycznych algorytmów wzorowanych na teorii selekcji naturalnej. Opierają się na tworzeniu zmutowanej populacji, jej krzyżowaniu i selekcji najlepszych osobników (w tym przypadku zakresów spektralnych). Proces działania algorytmu genetycznego można sprowadzić do następujących kroków:

1. Utworzenie chromosomów
   Widmo dzieli się na odcinki o ustalonej szerokości, które utworzą pierwsze chromosomy. Tak otrzymany zbiór nazywa się "generacją 0".

2. Ewaluacja chromosomów
   Na podstawie danych zawartych w chromosomach tworzy się modele i zapisuje odpowiadające im parametry jakości modelu, otrzymane przez walidację i walidację krzyżową.

3. Selekcja
   Z generacji wybiera się określony % chromosomów o najlepszych parametrach.

4. Mutacja

Od każdego chromosomu odcina się fragment o określonej długości (specyfikowanej lub losowej), a następnie fragmenty te miesza się i ponownie łączy z chromosomami. W wyniku takiego procesu otrzymuje się kolejną generację. W zależności od użytych ustawień liczba otrzymanych chromosomów może być większa niż przed tym procesem, co rekompensuje proces selekcji.

5. Repetycja

Powtórzenie etapów 2-4, skutkujące stopniową iteracją i powstawaniem nowych modeli. Proces powtarza się przez ustaloną ilość iteracji, po czym wybiera się model o najlepszych możliwych parametrach znalezionych w ciągu całego procesu.

2.2.4 Metoda selekcji interwałów

Strategia ta opiera się na podzieleniu widma na odcinki o ściśle ustalonej szerokości, a następnie zbadanie parametrów odpowiadających jakości otrzymanych odcinków. Po wybraniu odcinka o najlepszej jakości przeprowadza się ponowny podział i bada jakość modeli powstałych w wyniku połączenia poprzedniego odcinka z nowo powstałymi.

Miarę wpływu otrzymanych odcinków na jakość modelu bada się poprzez monitorowanie wartości parametru RMSECV (Pierwiastek średniego błędu standardowego walidacji krzyżowej), opisanego wzorem:

$$RMSECV = \sqrt{\frac{\sum (Y_i^{rzeczywisty} - Y_i^{crosswalidacji})^2}{N}}, \tag{2.2}$$

gdzie:

$Y_i^{rzeczywisty}$ oznacza referencyjny pomiar i,

$Y_i^{crosswalidacji}$ oznacza wartość walidacji krzyżowej obliczonej dla pomiaru i,

N jest liczbą pomiarów.

Parametr ten reprezentuje parametr walidacji krzyżowej otrzymanych modeli. Metoda ta może być poddana dalszym modyfikacjom, takim jak użycie interwału kroczącego - w tym trybie nowo otrzymane odcinki mogą nakładać się na siebie, co pozwala na precyzyjniejsze znalezienie krawędzi informatywnych zakresów. Główną wadą tej metody jest niezwykle duży czas potrzebny na jej przeprowadzenie - do przeprowadzenia wyczerpującego przeszukania widma niezbędne jest zbudowanie od kilkuset do kilku tysięcy modeli, a proces ten trzeba powtórzyć dla każdego kolejnego dodawanego zakresu. Dodatkowo, kluczowe jest tutaj wybranie zakresu skanowania o odpowiedniej szerokości - zbyt mały użyty zakres będzie powodował silną korelację z szumem zawartym w danych, a zbyt duży

będzie skutkował brakiem precyzji dopasowania użytych zakresów do pasm informatywnych w widmie.

2.2.5 VIP Scores [3]

Czynniki VIP (Waga projektowanej zmiennej) to estymatory wagi wkładu poszczególnych liczb falowych w wynik predykcji modelu. Można je obliczyć wzorem:

$$VIP_j = \sqrt{\frac{\sum_F^{f=1} w_{jf}^2 * SSY_f * J}{SSY_{total} * F}}, \tag{2.3}$$

gdzie:

w_{jf} oznacza wagę zmiennej j i składowej f,

SSY_f jest sumą kwadratów wyjaśnionej wariancji dla składowej f_J opisującego X zmiennych,

SSY_{total} jest sumą kwadratów wyjaśnionych przez zmienną zależną,

F jest łączną ilością składowych.

Wyższa wartość parametru VIP przypisanego do sygnału dla danej liczby falowej oznacza jej wyższy udział w tworzeniu zdolności prognostycznej modelu; VIP wyższy od 1 uważa się za znaczący, zmienne charakteryzowane przez wartości niższe zostają zwykle odrzucone.

Rozdział 3

Część doświadczalna

Badania przeprowadzono na danych spektroskopowych pozyskanych techniką IR-ATR dla 63 próbek pyłków pszczelich, z zastosowaniem pomiarów referencyjnych wykonanych z użyciem adekwatnych procedur analitycznych.

3.1 Materiał badawczy

Dwadzieścia siedem próbek pyłku zakupiono z polskich pszczelarni. Każda próbka zawierała mieszaninę pyłków pochodzącą z różnych roślin rosnących w okolicach pasiek. Dodatkowe dwadzieścia dwie próbki otrzymano poprzez mieszanie zakupionych pyłków w różnych proporcjach, a czternaście próbek otrzymano poprzez łączenie ziaren pyłków o podobnej morfologii.

3.1.1 Pomiary referencyjne

Analizy zawartości azotu i oznaczenia ilościowe aminokwasów wykonano w laboratoriach we Wrocławskim Uniwersytecie Przyrodniczym.

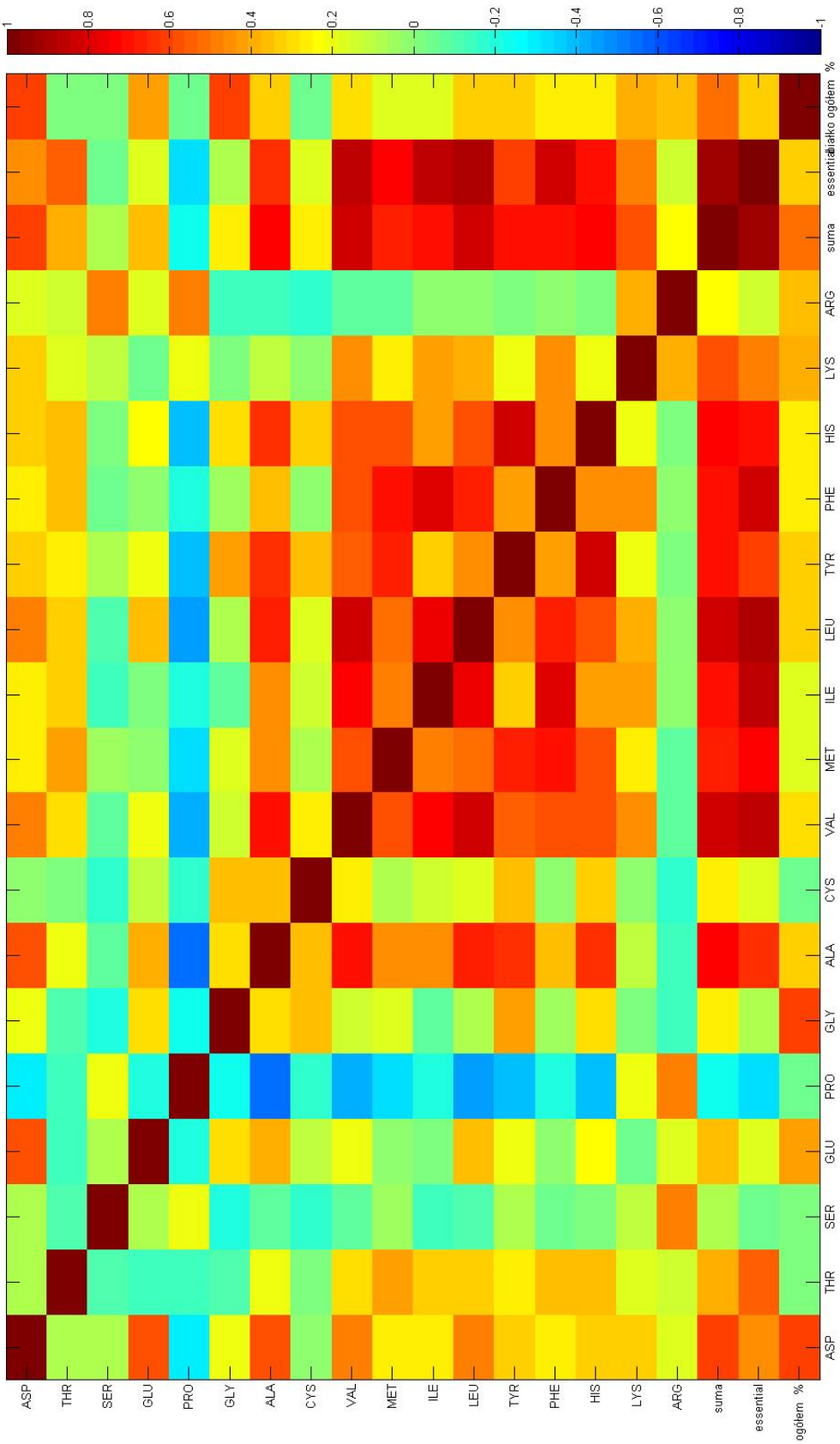
Zawartość azotu zmierzono zgodnie z procedurą AOAC 981.10, opartą na metodzie Kjeldahl'a, a następnie przeliczono na zawartość białek używając czynnika konwersji o wartości 6.25.

Badanie profilu aminokwasowego wykonano techniką LC-PDA-Qtof-ESI-MS (identyfikacja) i UPLC-PDA (analiza ilościowa). Analizę przeprowadzono poprzez pobranie 3 μL ekstraktu każdej próbki i przepuszczenie przez kolumnę AccQ Tag Ultra BEH (2.1 $\times$ 100 mm, 1.7 μm; Waters Corp.; Ireland) w temp. 50 °C i gradiencie o przepływie 0.5 mL/min dla czasu 15 min. Faza mobilna składała się z mieszaniny acetonitrilu:kwasu mrówkowego:octanu amonu (10:6:84, v/v/v) i rozpuszczalnika B (acetonitrilu:kwasu mrówkowego, 99.9:0.1, v/v) gdzie użyto 99% A do 0.3 min, 97% A do 3.2 min i 40% A do 11.0 min.

Spektrometr mas został ustawiony na badanie zakresu m/z od 100 do 700, napięcie kapilary na 2500V, napięcie stożka na 30V, temperaturę desolwatacji na 350 °C, a przepływ azotu na 535 l/h. Identyfikację aminokwasów wykonano poprzez pomiar czasów retencji w trybie jonów ujemnych. Czasy retencji i widma porównano z czystymi standardami. Analizy ilościowej dokonano poprzez wykonanie krzywej kalibracyjnej ze znanych stężeń w zakresie od 20 do 100 mg/l, których widma wykonano przy 260 nm.

3.2 Tabela korelacji

W tabeli 3.1 przedstawiono zależności między stężeniami referencyjnie oznaczonych substancji, tj. białka ogółem, aminokwasów egzogennych i poszczególnych aminokwasów. Najwyższe wartości współczynnika korelacji zaobserwowano między sumą aminokwasów a aminokwasami egzogennymi, co jest naturalne z uwagi na fakt, że aminokwasy kluczowe stanowią około połowy substancji zawartych w sumie aminokwasów. Ujemna korelacja proliny z pozostałymi aminokwasami jest zjawiskiem typowym dla białek i wynika z faktu, że prolina występuje jedynie w bardzo specyficznych strukturach III-rzędowych. Arginina jako jedyna wykazuje bardzo niską korelację z resztą oznaczonych aminokwasów. Z uwagi na występowanie silnych zależności stężeniowych między niektórymi składnikami próbek można oczekiwać dla nich podobnych parametrów jakościowych modeli kalibracyjnych.



Rysunek 3.1: Tabela korelacji między stężeniami związków oznaczonych w próbkach.

3.2.1 Pomiary spektroskopowe

Pomiary widm FTIR (ATR) w zakresie MIR zostały wykonane z użyciem jednoodbiciowej przystawki Golden Gate (Specac, Slough, UK) przy użyciu spektrometru Nicolet iS50(ThermoFisher). Do pomiarów użyto dzielnika wiązki wykonanego z KBr i detektora DTGS. Interferogram uśredniono dla 128 skanów, a następnie przekształcony w widmo w zakresie 400–4000 cm^{-1} przy rozdzielczości 4 cm^{-1} z zastosowaniem transformaty Fouriera. Pomiary wykonano dla porcji kilku mg każdej próbki. Finalne widma zostały otrzymane przez uśrednienie wyników dla trzech powtórzeń.

3.3 Omówienie użytych programów

W toku badań użyto trzech programów: Matlab, TQ Analyst oraz OMNIC. Zostaną im poświęcone osobne sekcje, gdzie krótko zostanie omówione ich zastosowanie w kontekście omawianej pracy.

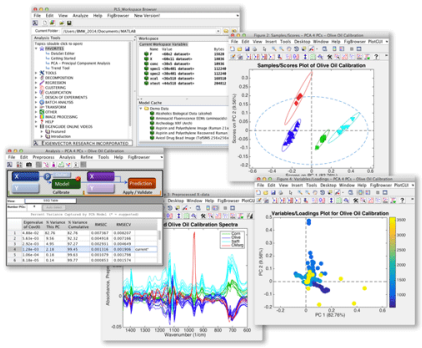
3.3.1 Matlab

Matlab został wybrany na główne środowisko pracy z uwagi na możliwość zastosowania pakietu 'PLS TOOLBOX' firmy Eigenvector, użyto wersji R2010a.

Matlab jest zamkniętym środowiskiem programistycznym nastawionym na wykonywanie operacji macierzowych. Jako język skryptowy charakteryzuje on się też stosunkowo prostą składnią, dodatkowo jest on zintegrowany z interfejsem graficznym który stanowi jednocześnie środowisko pracy.

PLS TOOLBOX

Jest to zespół gotowych funkcji chemometrycznych z interfejsem graficznym, który pozwala na wgranie, obróbkę danych pomiarowych i tworzenie z nich różnego typu modeli chemometrycznych. Mimo dość nieintuicyjnego interfejsu graficznego (rysunek 3.2) jest to jedno z najbardziej rozbudowanych rozwiązań do tworzenia modeli prognostycznych do celów chemicznych, które dodatkowo jest stale rozwijane.



Rysunek 3.2: PLS Toolbox pozwala na obrazowanie i tworzenie modeli analitycznych

W kontekście tej pracy najważniejszą cechą tego narzędzia jest rozbudowane API (Interfejs programowalny aplikacji), pozwalające na wykonywanie poszczególnych funkcji toolboxa poprzez pisanie

własnych programów. Umożliwia to na automatyzację wybranych etapów tworzenia modeli, co było jednym z głównych elementów tej pracy.

Dodatkowymi zaletami PLSTB (Oprogramowanie PLS Toolbox) względem kolejnego z omówionych programów jest możliwość przeprowadzenia analizy iPLS (Interwałowa metoda selekcji danych PLS) oraz zbadania udziału VIP scores'ów w powstałym modelu. W poniższej pracy użyto wersji 6.2 Toolboxa.

3.3.2 TQ Analyst

Jest to program firmy ThermoFisher, przeznaczony docelowo do obróbki danych pochodzących z przyrządów pomiarowych tego producenta. Mimo mniejszej ilości funkcji niż omówione wcześniej narzędzie PLSTB ma on znacznie lepiej rozbudowany i przejrzysty interfejs graficzny, co ułatwia ręczne dobieranie zakresów widmowych do budowy modeli, odrzucanie i dodawanie próbek podczas analizy oraz szybką ocenę PRESS (Przewidywana reszta sum kwadratów błędów) i innych parametrów jakości modelu.

Mimo szeregu zalet, TQ (Oprogramowanie Turbo Quant Analyst) ma kluczową wadę - nie ma zewnętrznego API którym można sterować z poziomu języka programowania. Istnieje możliwość jego automatyzacji poprzez środowisko Macros Basic, jednak jest ono ograniczone w kwestii wizualizacji danych oraz dokumentacji.

3.3.3 OMNIC

Jest to kolejny program firmy ThermoFisher, przeznaczony do wizualizacji i obróbki widm spektroskopowych. W tej pracy został zastosowany jako interfejs pomiędzy TQ a PLSTB. Dane z programu TQ można eksportować wyłącznie w formacie .SPG, który jest nierozpoznawany przez PLSTB. Użyto więc programu OMNIC do rozpakowania .SPG na poszczególne widma, które zapisano pojedynczo w formacie .CSV, a następnie przesłano do środowiska Matlab.

3.4 Przygotowanie danych

Procedurę tworzenia modelu rozpoczęto od wgrania surowych widm do programu TQ i przeprowadzenia w nim korekty MSC (Multiplikatywna korekcja rozproszenia), liczenia drugiej pochodnej i centrowania danych spektroskopowych. W kolejnym kroku zbudowano model testowy, pozwalając oprogramowaniu samoczynnie podzielić dane na zestaw kalibracyjny i walidacyjny. Na bazie wstępnego modelu dokonywano początkowego odrzucenia próbek odstających, starając się nie przekroczyć 20% populacji.

3.5 Omówienie procedury budowy modeli

Ręczne tworzenie modelu w PLS Toolbox składa się z następujących kroków:

1. Wgranie bloku X - kalibracyjnych danych spektroskopowych zapisanych w postaci macierzy.
2. Wgranie bloku Y - kalibracyjnych danych referencyjnych zapisanych w postaci wektora Y.
3. Wybrania sposobu preprocessingu bloku X i Y wybierając kolejność i sposób postępowania.
4. Wgranie bloku Xv - walidacyjnych danych spektroskopowych zapisanych w postaci macierzy.
5. Wgranie bloku Yv - walidacyjnych danych referencyjnych zapisanych w postaci wektora Y.
6. Ustalenie trybu walidacji krzyżowej na LOO (Walidacja krzyżowa pozostawienia jednej próbki).
7. Zbudowanie testowego modelu, ręczne dobranie liczby użytych faktorów.
8. Sprawdzenie parametrów jakości modelu i decyzja czy zostaje on zapisany lub czy wymaga dalszej optymalizacji.

3.6 Etapy automatyzacji

Osiągnięcie założonego celu pracy wymagało udoskonalenia narzędzi używanych do tej pory w tworzeniu modeli chemometrycznych. PLS Toolbox w wersji 6.2 nie posiada dwóch funkcji, niezbędnych do tworzenia modeli o wysokim stopniu automatyzacji, mianowicie możliwości łatwego odrzucania próbek ignorowanych i dzielenia pełnej macierzy danych na próbki walidacyjne i kalibracyjne oraz wygodnego interfejsu do wyboru zakresów spektralnych użytych do budowy modelu.

Jak pokazano w poprzedniej pracy [7], selekcja zakresów spektralnych w modelowaniu parametrów dla naturalnych układów wieloskładnikowych ma decydujący wpływ na jakość otrzymanego modelu regresyjnego. Obecnie, wybór zakresu widmowego odbywa się ręcznie, poprzez żmudną pracę w oprogramowaniu TQ, często metodą prób i błędów aż do znalezienia zakresów dających akceptowalne wyniki. Proces ten często zajmuje wiele godzin i nie gwarantuje otrzymania modelu o wysokiej zdolności prognostycznej. W tej

pracy zaproponowano alternatywę do tego procesu: Algorytm heurystyczny ModelHelper, który automatycznie przeszukuje widmo i proponuje zakresy spektralne, generujące modele o możliwie wysokiej jakości.

3.6.1 Etap I: Synchronizacja oprogramowania

Przy rozpoczęciu procedury automatyzacji zdecydowano się na użycie oprogramowania TQ analyst jako zewnętrznej formy weryfikacji jakości otrzymanych modeli. Oparto się na założeniu, że program ModelHelper powinien proponować zakresy, które utworzą model o identycznych parametrach jakości zarówno w środowisku TQ jak i PLSTB. Pozwoli to potwierdzić uniwersalność i skuteczność metody w tworzeniu wysokiej jakości modeli.

Program TQ nie pozwala jednak na bezpośredni eksport modeli do środowiska PLSTB. Dane widmowe można uzyskać jako zgrupowane widma .SPG, które rozdzielono w programie OMNIC i zapisano pojedynczo jako pliki .csv. Problem stanowiło natomiast przeniesienie informacji o użytych próbkach, parametrach otrzymanego modelu i użytym preprocessingu. Te informacje można wyeksportować z TQ jedynie w formie pliku tekstowego, który poddano ręcznemu odczytowi przez dedykowany skrypt w Matlabie, który konwertował go na strukturę używaną przez PLSTB.

Po wstępnej obróbce danych, opisanej w sekcji 3.4, zdecydowano się powtórzyć procedurę w środowisku MATLAB, aby potwierdzić reproduktywność wyników. Stosując takie same ustawienia preprocessingu i dobór próbek otrzymano wyniki różniące się o 5-15% parametrami jakości (RMSEC, RMSEP, RMSECV). Udało się ustalić, że przyczyna różnic wynika z innych algorytmów liczenia drugiej pochodnej i korekty MSC. Problem ten rozwiązano poprzez wykonywanie drugiej pochodnej w programie TQ dla każdej próbki, a następnie wyeksportowaniu widm, odgrupowywaniu ich w oprogramowaniu OMNIC i imporcie do Matlaba. Dzięki temu uzyskano zgodność wyników na poziomie 1% błędu względnego dla parametrów jakości modeli.

Takie postępowanie pozwoliło na łatwą synchronizację i transport danych między środowiskiem MATLAB i TQ Analyst.

3.6.2 Etap II: Siłowa kombinacja

Ocena jakości modelu

Proces automatyzacji generuje tysiące modeli regresyjnych charakteryzowanych różnymi parametrami jakościowymi. Koniecznym krokiem jest ocena otrzymanych modeli, co pozwala na wybranie najlepszego z nich na każdym etapie optymalizacji. Poniżej przedstawiono procedurę oceny otrzymanych modeli chemometrycznych:

1. przyznaj po 1 punkcie za każde -0.01 różnicy R^2_{CV} między obecnym najlepszym a ocenianym modelem,
2. wykonaj regresję liniową dla krzywej PRESS,
3. jeśli współczynnik regresji jest większy niż -0.25, odbierz 3 punkty modelowi,
4. jeśli R^2_{CV} jest większy od 0.8, dodaj modelowi 1 punkt za każde -0.01 różnicy R^2 między obecnym najlepszym a ocenianym modelem,
5. jeśli R^2_{CV} jest większy od 0.8, dodaj modelowi 1 punkt za każde -10% różnicy RSEP między obecnym najlepszym a ocenianym modelem,
6. jeśli R^2_{CV} jest większy od 0.8, dodaj modelowi 1 punkt za każde -10% różnicy $RSEP_V$ między obecnym najlepszym a ocenianym modelem.

Taki sposób oceny mocno faworyzuje modele o wysokiej wartości parametru walidacji krzyżowej. Wynika to z założenia, że modele o niskim R_{CV} nie powinny być rozważane w kontekście przeprowadzania prognoz w rzeczywistym świecie - są one zbyt zależne od tego konkretnego zestawu próbek kalibracyjnych i najprawdopodobniej ich zdolność prognostyczna wynika z przypadkowych korelacji, które nie będą dobrze opisywać kolejnych próbek.

Jednym z udoskonaleń jest także sprawdzanie krzywej nachylenia PRESS. Dobre modele charakteryzują się przebiegiem PRESS przypominającym osuwisko. Uwzględnienie tej zależności pozwala na wykluczenie przypadków, gdzie algorytm podąża błędną ścieżką optymalizacji mimo pozornie dobrych parametrów R.

Dopiero po przekroczeniu progu 0.8 R_{CV} w modelu zostają uwzględnione inne parametry jakości, ponieważ w innym przypadku optymalizacja miałaby tendencję do faworyzowania over-fittowanych modeli o niskiej rzeczywistej zdolności prognostycznej.

Automatyzacja tworzenia modeli

Po osiągnięciu zgodności między TQ i PLSTB kolejnym krokiem było zautomatyzowanie etapów 1-8 opisanych w sekcji 3.4. Dokonano tego używając rozbudowanego API wbudowanego w PLSTB, pozwalającego na wywoływanie funkcji PLSTB z poziomu środowiska Matlab, poprzez pisanie własnych programów. Proces ten został ułatwiony dzięki rozbudowanej wikipedii oraz dokumentacji dostarczonej przez firmę Eigenvector [15]. Dodano także funkcjonalność podziału próbek na kalibracyjne, walidacyjne i ignorowane, aby nie trzeba było wykonywać tego procesu ręcznie przed każdą budową modelu. Po ukończeniu tego etapu prac tworzenie pojedynczego, prostego modelu, który zawierał pełne widmo, z podziałem na próbki kalibracyjne, walidacyjne i ignorowane zajmowało jedynie kilka sekund.

Automatyzacja doboru zakresów

W następnym etapie podjęto próbę opracowania prymitywnego algorytmu dzielącego widmo spektralne na części, z których później tworzono modele.

Zdecydowano się na użycie algorytmu bazującego na suwaku - na początku ustalano szerokość uwzględnianego zakresu w punktach pomiarowych oraz krok, o jaki będzie przesuwać się ten zakres na osi liczb falowych. Zastosowanie takiego podejścia pozwoliło na dość znaczne zwiększenie jakości tworzonych modeli. Istotną kwestią było ustalenie optymalnej szerokości okna. Problemem był również fakt, że w widmie może znajdować się wiele potencjalnie dobrych zakresów, a użycie tylko jednego z nich nie musi skutkować powstaniem optymalnego modelu.

W celu rozwiązania pierwszego problemu zaproponowano koncepcję map spektralnych - widmo rozcinano algorytmem suwaka przy niewielkiej szerokości okna (10 punktów pomiarowych, odpowiadających szerokości 5cm^{-1}). Po przejściu przez cały zakres powtarzano procedurę zwiększając szerokość okna o kolejne 10 punktów pomiarowych. Podejście to gwarantowało, że znaleziony zakres spektralny będzie optymalnej szerokości (z maksymalnym błędem 9 punktów szerokości), kosztem znacznego przedłużenia procedury skanowania. Wygenerowanie każdej z map trwało kilka minut, a ukończenie procedury wymagało stworzenia kilkudziesięciu z nich. Podejście to pozwoliło jednak na uzyskanie najlepszego możliwego modelu opartego na jednym wycinku z widma spektralnego.

Kombinatoryka

Aby rozwiązać problem potrzeby znalezienia wielu zakresów jednocześnie zdecydowano się na kombinację zakresów wycinanych z oryginalnych danych widmowych. Wszystkie n zakresów otrzymanych w wyniku metody 'suwak' było siłowo spajane, łącząc ze sobą k odcinków, gdzie k było definiowane przez użytkownika. Skutkowało to utworzeniem

$$ {}_n\mathrm{C}_k = \binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{3.1} $$

kombinacji, które następnie można było ewaluować pod kątem jakości otrzymanych modeli.

Oczywistym problemem jest tutaj kwestia czasu potrzebnego na ewaluację wyników takiego procesu kombinacji. Dla zakresu o niskiej rozdzielczości (szerokość 300, krok 50, kombinacje 3, przy 6000 punktów pomiarowych) proces taki generowałby pulę 280 tysięcy potencjalnych modeli do ewaluacji. Konieczne było więc dalsze zwiększenie szerokości odstępów między utworzonymi wycinkami widmowymi, co zmniejszało szanse na idealne 'trafienie' odcinka w informatywne pasmo spektralne. Kombinatoryka, mimo długiego

czasu wykonywania (4-16h) była jednak niezwykle skuteczna w znajdowaniu połączeń do 5 kombinowanych zakresów, dając modele o bardzo dobrej zdolności prognostycznej.

3.6.3 Etap III: Optymalizacja zakresów

Optymalizacja krawędzi

Na tym etapie podjęto kolejną próbę udoskonalenia algorytmu. Proces opierał się o przesuwanie krawędzi zakresów otrzymanych w wyniku kombinacji. Procedura wyglądała następująco:

1. Weź lewą krawędź pierwszego zakresu w widmie;
2. Przesuwaj krawędź w lewo o 300 punktów pomiarowych, tworząc tymczasowe modele;
3. Jeżeli jakikolwiek model tymczasowy jest lepszy od obecnego, zastąp obecny model modelem tymczasowym;
4. Weź lewą krawędź pierwszego zakresu w widmie;
5. Przesuwaj krawędź w prawo o 300 punktów pomiarowych, tworząc tymczasowe modele;
6. Jeżeli jakikolwiek model tymczasowy jest lepszy od obecnego, zastąp obecny model modelem tymczasowym;
7. Powtarzaj etapy 1-4 dla każdego stworzonego zakresu;
8. Weź prawą krawędź pierwszego zakresu w widmie;
9. Powtórz etapy 1-5 dla prawych krawędzi zakresów w widmie.

Procedura ta może dodatkowo udoskonalić modele otrzymywane w wyniku kombinacji, gwarantując, że każda krawędź wybranego zakresu znajdzie się na odpowiednim miejscu. Pozwoliło to na użycie większego kroku k w etapie kombinacji, co przyspieszyło wykonanie analiz.

Zasiew

Po dodaniu algorytmu optymalizacji krawędzi zakresów zauważono, że nie ma potrzeby znajdowania precyzyjnej pozycji zakresów poprzez kombinacje, jeśli procedura optymalizacji krawędzi i tak pozwoli na przesunięcie ich w optymalną pozycję. Zaproponowano więc prostszą metodę dodawania nowych zakresów spektralnych, polegającą na dodaniu

odcinka o stałej długości w miejsce, które najbardziej poprawi jakość modelu. Procedurę tę nazwano zasiewem, ze względu na podobieństwo rozrzucania nasion na grządce. Jest ona *de facto* zastosowaniem kombinacji dla jednego nowego odcinka.

Optymalizacja zakresów zasiewu

Zastosowanie zasiewu, a następnie optymalizacji krawędzi pozwoliło na szybkie otrzymywanie kolejnych generacji modeli. Problemem pozostawała kwestia bardzo długiego czasu potrzebnego na kombinatorykę, aby otrzymać model startowy na którym przeprowadzano zasiew. Postanowiono użyć podejścia zaproponowanego w metodzie iPLS - zamiast przeprowadzać na początku siłową rekombinację dodając wiele zakresów na raz, tak jak do tej pory, zacząć od pustego zakresu i dodawać kolejne jeden po drugim. Rozwiązanie to nie gwarantuje znalezienia optymalnego punktu startowego, jednak po przeprowadzeniu porównania podejść dla obu metod na kilku próbkach stwierdzono brak znaczącej różnicy w jakości otrzymanych modeli.

Ostatecznie porzucono metodę kombinatoryczną oraz tworzenie map spektralnych - obecnie algorytm używa ustalonej z góry szerokości okna i kroku do skanowania, a następnie przeprowadza optymalizację krawędzi, starając się rozszerzyć i zwęzić każdy zakres w lewo i prawo o ustaloną w konfiguracji liczbę punktów. Takie podejście może potencjalnie pominąć obiecujące pasma znajdujące się w pobliżu zakresów nieinformatywnych, ale problem ten można rozwiązać zmniejszając szerokość okna skanu w momencie gdy model nie poprawia się w kolejnych generacjach. Stworzony algorytm został nazwany 'ModelHelper' i opisany w kolejnych paragrafach.

ModelHelper

Działanie algorytmu heurystycznego wewnątrz skryptu ModelHelper można sprowadzić do następującego cyklu:

0. Start: Rozpocznij od pustej lub zawierającej już wycinki widma macierzy intensywności;

1. Zasianie: Dodanie kolejnego zakresu do modelu, charakteryzującego się najlepszymi parametrami jakości w puli wygenerowanych modeli regresyjnych;

2. Rozszerzenie: Rozszerzenie po kolei krawędzi wszystkich istniejących zakresów do znalezienia ich optymalnych krawędzi;

3. Kurczenie: Zawężenie krawędzi wszystkich istniejących zakresów do znalezienia optimum.

Dzięki nowemu podejściu do przesiewania zakresu udało się zredukować czas potrzebny na uzyskanie modelu względem prototypu z kilkunastu godzin do kilkunastu - kilkudziesięciu minut dla trudniejszych modeli, przy jednoczesnym ulepszeniu parametrów otrzymywanych modeli względem poprzedniej metody.

Jedynymi częściami całej procedury, których nie poddano automatyzacji, były selekcja próbek odbiegających oraz liczba faktorów PLS. Całość analizy przeprowadzano dla liczby faktorów i konfiguracji próbek ustalonej w modelu TQ. Automatyzacja tych parametrów wybiegałaby poza zakres pracy i wymagała użycia dodatkowych narzędzi, takich jak analiza PCA i zastosowanie algorytmu genetycznego przy doborze próbek.

3.7 Tworzenie modeli stężeń aminokwasów

Po stworzeniu skryptu ModelHelper cała procedura optymalizacji modelu dla pojedynczego aminokwasu trwała około jednej-dwóch godzin. Poniżej przedstawiono proces tworzenia modelu z zastosowaniem nowego algorytmu:

1. Wprowadzenie danych pomiarowych i referencyjnych do programu TQ Analyst;
2. Wykonanie preprocessingu, automatyczny dobór próbek kalibracyjnych;
3. Odrzucenie próbek odstających;
4. Eksport informacji o próbkach, użytych zakresach i danych referencyjncyh do formatu .TXT;
5. Eksport widm po preprocessingu do formatu .SPG;
6. Rozdzielenie widm w programie OMNIC i zapisanie w formacie .CSV;
7. Załadowanie widm z formatu .CSV i informacji z pliku .TXT do MATLAB;
8. Włączenie algorytmu ModelHelper poprzez jedno polecenie;
9. Nadzorowanie algorytmu ModelHelper, podczas gdy poprawia on modele, poprzez wydawanie poleceń znajdowania kolejnych zakresów lub optymalizacji istniejących;
10. Eksport danych z PLSTB do TQ;
11. Ponowna budowa modelu w TQ Analyst z zastosowaniem zakresów zaproponowanych przez ModelHelper;
12. Ponowne dobranie próbek walidacynych, włączenie próbek ignorowanych pasujących do modelu, usunięcie odstających;

13. Dobranie odpowiedniej liczby faktorów PLS;
14. Ponowny eksport pliku .TXT i załadowanie go razem z wcześniej zapisanymi widmami;
15. Ponowne użycie skryptu ModelHelper;
16. Powtórzenie etapów 8-12 do momentu zaniku poprawy między generacjami (średnio 4 generacje);
17. Weryfikacja parametrów jakości, zapisanie modelu.

Rozdział 4

Test wydajności ModelHelper

W tym rozdziale nastąpi porównanie opracowanego w toku badań algorytmu ModelHelper w zestawieniu z szeregiem powszechnie używanych metod selekcji danych.

W celu otrzymania reprezentatywnego porównania modele testowe zbudowano dla tych samych zestawów próbek, przy zachowaniu identycznego preprocessingu i bez wcześniejszego doboru zakresów widmowych. Wyniki dla utworzonych modeli będą różnić się od tych przedstawionych w dalszej części pracy, ponieważ na potrzeby tego porównania stworzono je z zachowaniem takiego samego podziału próbek dla wszystkich prezentowanych metod selekcji danych.

4.1 Białko ogółem

Model odniesienia

Wyjściowy model PLS został utworzony poprzez wybranie pełnego zakresu widma w programie TQ Analyst, podzielenie próbek na kalibracyjne, walidacyjne i ignorowane oraz dopasowanie liczby faktorów i zbudowanie modelu wraz z walidacją krzyżową.

Ręcznie zbudowany model

Model ręczny został opracowany przed rozpoczęciem tej pracy w programie TQ. Stanowi on dobry odnośnik do typowego, wysokiej jakości, ręcznie wykonanego modelu chemometrycznego. Szacunkowy czas utworzenia modelu: 4h.

TQ Analyst

Model ten został wykonany poprzez wybranie pełnego zakresu widmowego, a następnie użycie opcji 'suggest' w programie TQ Analyst. Program ten wykonał selekcję danych na podstawie niejawnego algorytmu. Szacunkowy czas tworzenia modelu: 30s.

IPLS

Model ten został wykonany poprzez transfer danych z programu TQ do PLSTB, a następnie uruchomienie iPLS z parametrami:

- Szerokość przedziału: 80 punktów pomiarowych
- Liczba zakresów: 3
- Liczba obliczonych faktorów: 10

Szacunkowy czas utworzenia modelu: 8h, automatycznie.

VIP Scores

Model dla metody VIP Scores został opracowany poprzez transfer danych z programu TQ do PLSTB, a następnie ekstrakcję czynników VIP z modelu utworzonego w PLSTB i zbudowanie nowego modelu dla macierzy intensywności z zakresu gdzie $VIP_i > 1$.

ModelHelper

Model został opracowany poprzez transfer danych z programu TQ do PLSTB, a następnie uruchomienie skryptu 'Model Helper'. Ulepszanie modelu polegało na wydawaniu poleceń znalezienia nowych zakresów widma oraz manipulacji obecnie istniejącymi zakresami. Szacunkowy czas utworzenia modelu: 35min, półautomatycznie.

Parametr	RMSEC	RMSECV	RMSEP	R^2	R^2_{CV}	R^2_V
Model zerowy	0.561	2.61	1.530	0.970	0.363	0.716
Model ręczny	0.663	1.33	0.895	0.958	0.833	0.897
TQ Analyst	1.200	1.51	0.798	0.863	0.784	0.917
IPLS	0.845	1.58	1.504	0.932	0.770	0.789
VIP Scores	0.920	1.61	0.856	0.919	0.765	0.905
Model Helper	0.887	1.05	0.890	0.925	0.900	0.918

Tablica 4.1: Tabela parametrów uzyskanych dla różnych metod selekcji danych dla modelu zawartości białka

4.2 Walina

Model odniesienia

Wyjściowy model PLS został utworzony przez wybranie pełnego zakresu widma w programie TQ Analyst, podział próbek na kalibracyjne i walidacyjne, dopasowanie liczby

faktorów i zbudowanie modelu wraz z walidacją krzyżową, a następnie ręczne usunięcie 4 próbek odstających.

Ręcznie zbudowany model

Model ręczny został opracowany w programie TQ. Szacunkowy czas utworzenia modelu: 1.5 h.

TQ Analyst

Model ten został wykonany poprzez wybranie pełnego zakresu widmowego, a następnie użycie opcji 'suggest' w programie TQ Analist. Program ten wykonał selekcję danych na podstawie niejawnego algorytmu. Szacunkowy czas tworzenia modelu: 30s.

iPLS

Model ten został wykonany poprzez transfer danych z programu TQ do PLSTB, a następnie uruchomienie iPLS z parametrami:

- szerokość przedziału: 80 punktów pomiarowych
- liczba zakresów: 3
- liczba obliczonych faktorów: 10

Szacunkowy czas utworzenia modelu: 13h, automatycznie.

VIP Scores

Model dla metody VIP Scores został opracowany poprzez transfer danych z programu TQ do PLSTB, a następnie ekstrakcję czynników VIP z modelu utworzonego w PLSTB i zbudowanie nowego modelu dla liczb falowych gdzie $VIP_i > 1$.

ModelHelper

Model został opracowany poprzez transfer danych z programu TQ do PLSTB, a następnie uruchomienie skryptu 'Model Helper'. Ulepszanie modelu polegało na wydawaniu poleceń znalezienia nowych zakresów widma oraz manipulacji obecnie istniejącymi zakresami.

Parametr	RMSEC	RMSECV	RMSEP	R^2	R^2_{CV}	R^2_V
Model zerowy	1.210	2.97	2.851	0.84	0.116	0.089
Model ręczny	2.03	2.73	1.840	0.53	0.214	0.454
TQ Analyst	2.780	2.99	1.460	0.13	0.033	0.635
IPLS	1.196	2.16	2.313	0.84	0.491	0.160
VIP Scores	1.714	2.37	2.321	0.67	0.402	0.015
Model Helper	0.940	1.60	3.777	0.90	0.714	0.260

Tablica 4.2: Tabela parametrów uzyskanych dla różnych metod selekcji danych dla modelu zawartości waliny

4.3 Omówienie wyników testu wydajności

Jak przedstawiono na załączonych tabelach 4.1 i 4.2, modele otrzymane w wyniku działania ModelHelper charakteryzują się najwyższym parametrem walidacji krzyżowej R_{CV}, co świadczy o ich stabilności (ang. *robustness*) i dużej zdolności do tworzenia modeli, które nie są over-fitowane względem danych kalibracyjnych. Tendencja ta wynika z silnej faworyzacji parametru R_{CV} w logice proponowanego algorytmu. Proces optymalizacji skutkuje jednak często zwiększeniem RMSEP (Pierwiastek średniego błędu kwadratowego przewidywania) - wynika to z dużego błędu dla pojedynczych walidacyjnych próbek skrajnych. W toku konstrukcji typowego modelu próbki te byłyby zawrócone do zestawu kalibracyjnego, po czym wybrano by dodatkowy zestaw walidacyjny. W przedstawionej procedurze, w celu zachowania reprezentatywności próbek, ten krok został pominięty.

Algorytm ModelHelper jest wysoce skuteczny w otrzymywaniu modeli charakteryzujących się bardzo dobrymi parametrami jakości. Jego dodatkową zaletą jest także możliwość udoskonalania modeli stworzonych innymi metodami selekcji danych - na potrzeby dalszego modelowania przedstawionego w tej pracy częstym punktem startowym był zakres proponowany przez TQ Analyst. Pozwalało to rozpocząć proces optymalizacji od względnie dobrego punktu startowego i ograniczało szansę na utknięcie algorytmu w minimum lokalnym.

Rozdział 5

Prezentacja wyników modelowania

Dla analizowanego zestawu danych, tj. widm ATR próbek pyłków pszczelich i wyników oznaczeń referencyjnych zawartości białka i poszczególnych aminokwasów przeprowadzono modelowanie techniką PLS z uwzględnieniem zakresów widmowych zoptymalizowanych przy użyciu algorytmu ModelHelper. Poniżej przedstawiono wartości parametrów jakościowych opracowanych modeli wraz z użytymi zakresami.

Na początkowym etapie modelowania użyto zestawu próbek walidacyjnych zaproponowanych przez TQ. W kolejnych krokach model ten był poddawany stopniowym modyfikacjom, polegających na dodawaniu kolejnych zakresów widmowych i optymalizacji ich krawędzi, zgodnie z procedurą opisaną w paragrafie 3.6.3.

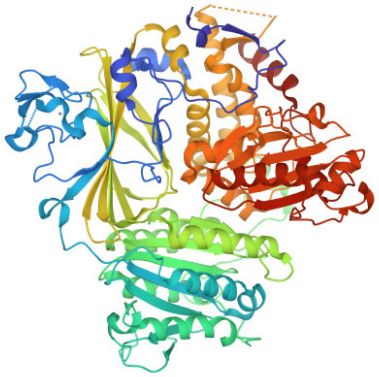
Użyte zakresy

Zakresy widmowe użyte do budowy poniższych modeli PLS zostały w całości wybrane przez wcześniej opisany algorytm heurystyczny. Oznacza to, że nie zawsze możliwa jest identyfikacja pochodzenia wybranych sygnałów - procedura wybierała dane, które pozwalały na uzyskanie najlepszych parametrów jakościowych. Mogłoby to oznaczać, że były preferowane liczby falowe, przy których nie występowała absorpcja od innych składników próbek (cukrów, wody, polifenoli, tłuszczów) oraz pasma na których ukazują się pośrednie interakcje białek/aminokwasów z innymi składnikami zawartymi w próbkach.

Zastanawiający może być fakt, że duża część modeli zawiera pasma z zakresu 1800-2200 cm^{-1}, tradycyjnie uważanego za mało interesujący z uwagi na pokrywanie się z zakresem absorpcji kryształu ATR - diamentu. Zjawisko to najprawdopodobniej wynika ze skłonności algorytmu do załączania nowych zakresów nawet przy niewielkiej zmienności spektralnej.

5.1 Modelowanie zawartości białka

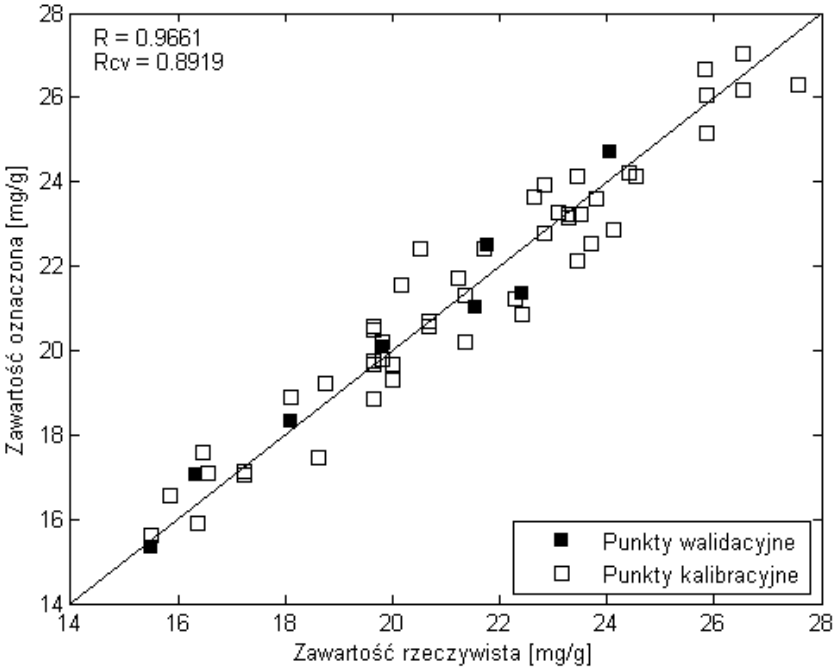
W analizowanym zestawie widm pyłków zawartość białka ogółem wahała się w zakresie 15-28mg/g suchej masy. Analiza zawartości białka była najmniej wymagającą obliczeniowo z przedstawionych analiz z uwagi na dużą zawartość składnika i relatywnie charakterystyczne udziały spektralne tej grupy składników w widmie próbki. Uzyskany model charakteryzował się bardzo dobrymi parametrami jakościowymi, a względny błąd oznaczeń nie przekraczał 2%. Na etapie kalibracji usunięto 8 próbek kalibracyjnych o skrajnie wysokich błędach predykcji, a do walidacji zastosowano osiem próbek. Krzywa PRESS charakteryzuje się poprawnym przebiegiem, więc zdecydowano o użyciu 3 faktorów do budowy modelu. W trakcie optymalizacji algorytm zaproponował 7 zakresów o zmiennej szerokości. Szczegółowe informacje dotyczące modeli przedstawiono w tabeli 5.1 oraz na rysunkach. 5.2, 5.3, 5.4, 5.5.



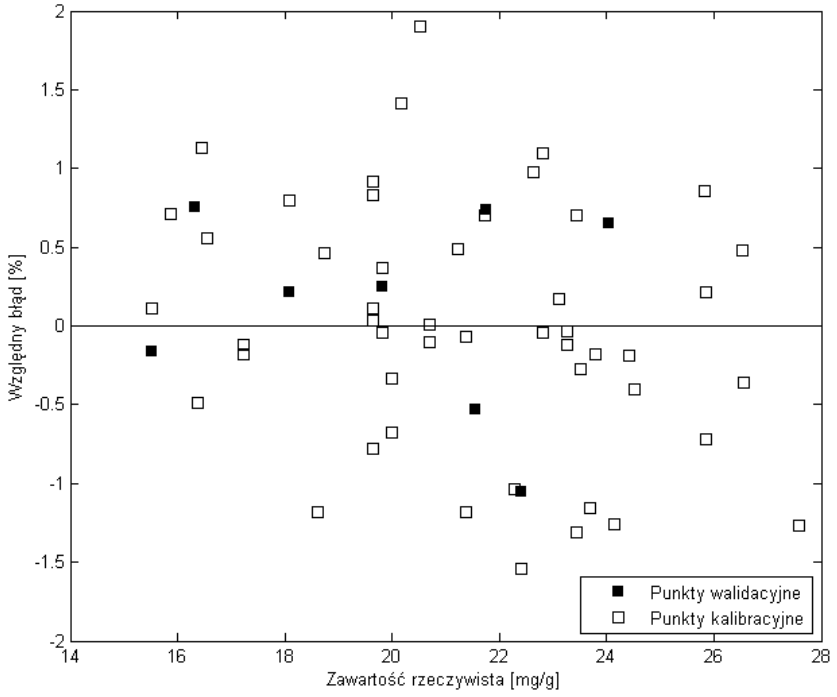
Rysunek 5.1: Struktura przykładowego białka Sec24b [54]

Białko	
Zakres stężeniowy	15-28mg/g
RMSEC	0.778
RMSEP	0.617
Użyte faktory:	3

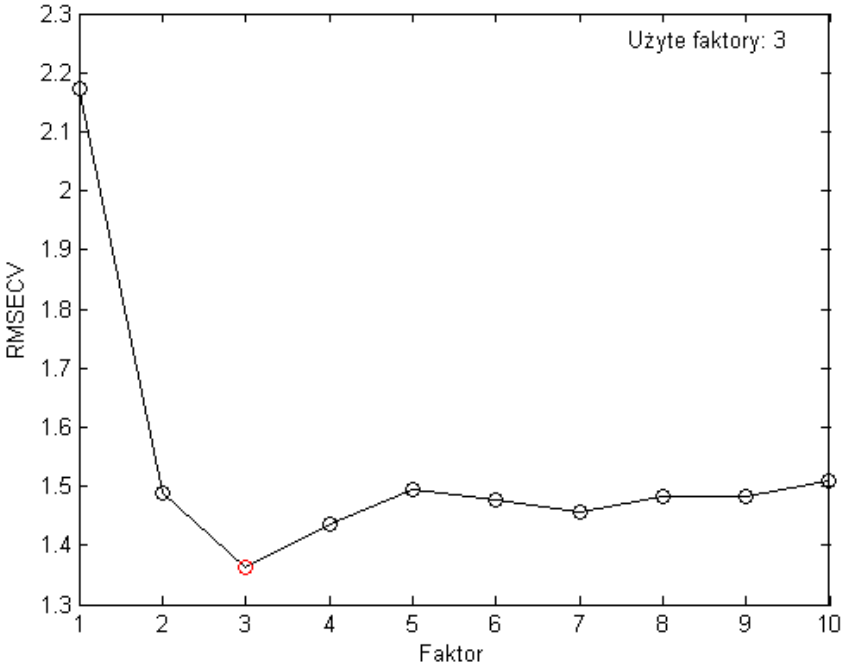
Tablica 5.1: Tabela parametrów dla modelu zawartości białka



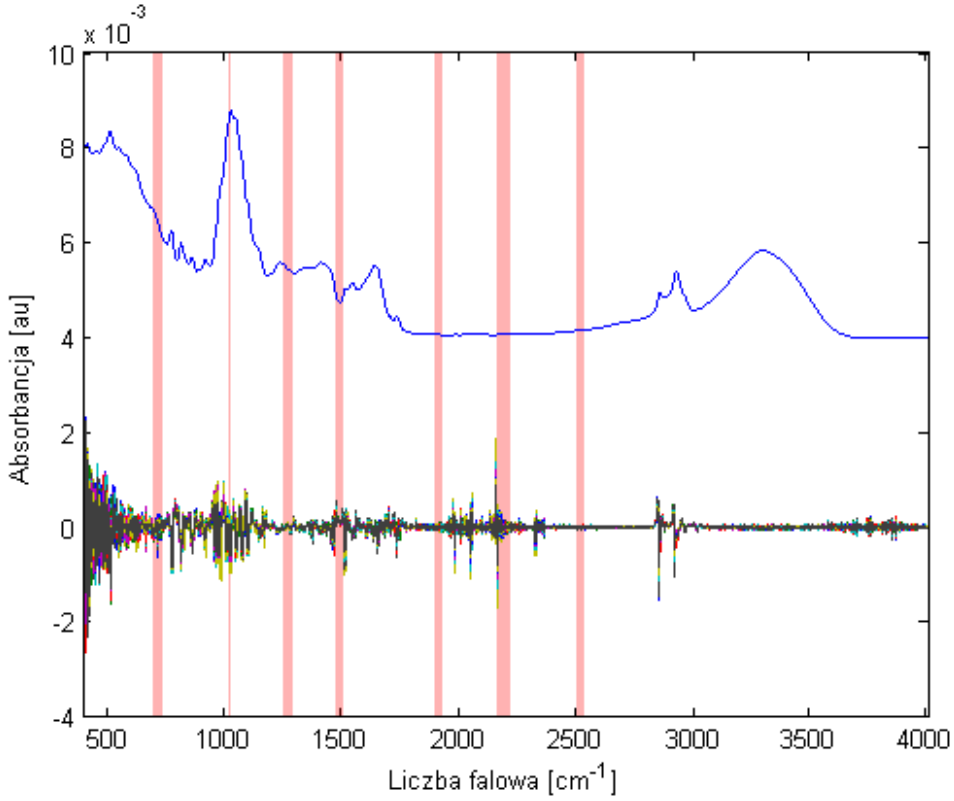
Rysunek 5.2: Krzywa predykcji oznaczeń zawartości białka w próbce



Rysunek 5.3: Względny błąd oznaczeń zawartości białka w próbce



Rysunek 5.4: Wykres RMSCEV dla modelu zawartości białka w próbce



Rysunek 5.5: Wykres zakresów użytych do budowy modelu zawartości białka w próbce

5.2 Modelowanie zawartości aminokwasów egzogennych

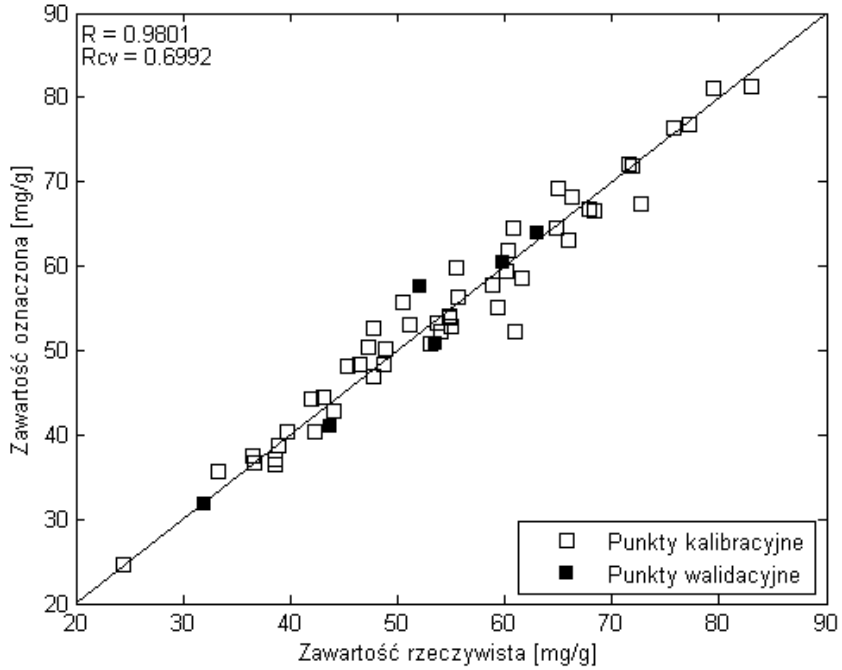
Do aminokwasów egzogennych zalicza się tradycyjnie osiem aminokwasów wymienionych w tabeli 5.2, poddawanych analizie ilościowej w celu ustalenia wartości odżywczej badanej próbki. W modelu ilościowym aminokwasów egzogennych nie uwzględniono obecności tryptofanu, ze względu na brak analizy referencyjnej tego aminokwasu. Jako, że model ten dotyczył dużej grupy związków oczekiwano otrzymania modelu charakteryzowanego przez dobre parametry jakościowe. Jego opracowanie wymagało jednak usunięcia aż 8 próbek, przy czym dalej zawiera on próbki odstające, a dobranie zakresów było wyjątkowo czasochłonne nawet przy użyciu algorytmu ModelHelper, zajmując ponad godzinę. Jak zostanie pokazane na późniejszych modelach, większość poszczególnych aminokwasów egzogennych charakteryzowała się słabymi/przeciętnymi wynikami modelowania, czego przyczyny zostaną omówione w odpowiadającym im sekcjom. Model dla ich sumy jest prawdopodobnie obarczony sumą błędów oznaczeń dla pojedynczych aminokwasów - niskie stężenia badanych związków, ich silna korelacja i podobne widma (na przykład leucyny i izoleucyny) powodują, że precyzyjne ustalenie sumy ich stężeń wiąże się z dużymi problemami (tabela 5.3). Rezultatem jest otrzymanie parametru R_{CV} na poziomie 0.699 oraz słabego przebiegu PRESS przedstawionego w tabeli 5.8. Do budowy modelu użyto ośmiu zakresów spektralnych (rysunek 5.9); przebiegi krzywej predykcji, błędów względnych i PRESS przedstawiono na figurach 5.6, 5.7, 5.8.

Aminokwasy egzogenne	
Histydyna	HIS
Izoleucyna	ILE
Leucyna	LEU
Lizyna	LYS
Metionina	MET
Fenyloalanina	PHE
Treonina	THR
~~Tryptofan~~	~~TRP~~
Walina	VAL

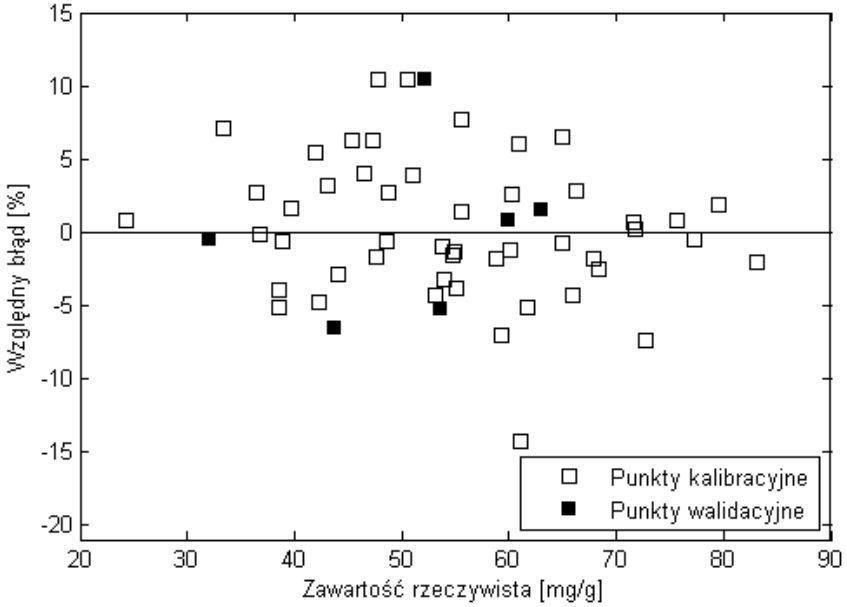
Tablica 5.2: Lista aminokwasów egzogennych wraz z akronimami. Wykreślony aminokwas nie został oznaczony referencyjnie

Aminokwasy egzogenne	
Zakres stężeniowy	25-85mg/g
RMSEC	2.59
RMSEP	2.78
Użyte faktory:	5

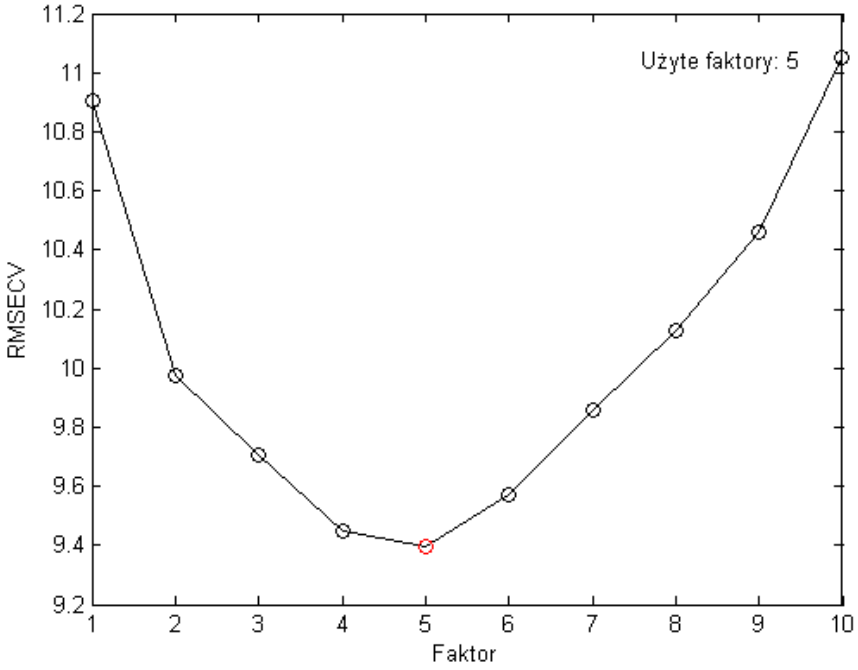
Tablica 5.3: Tabela parametrów dla modelu zawartości aminokwasów egzogennych



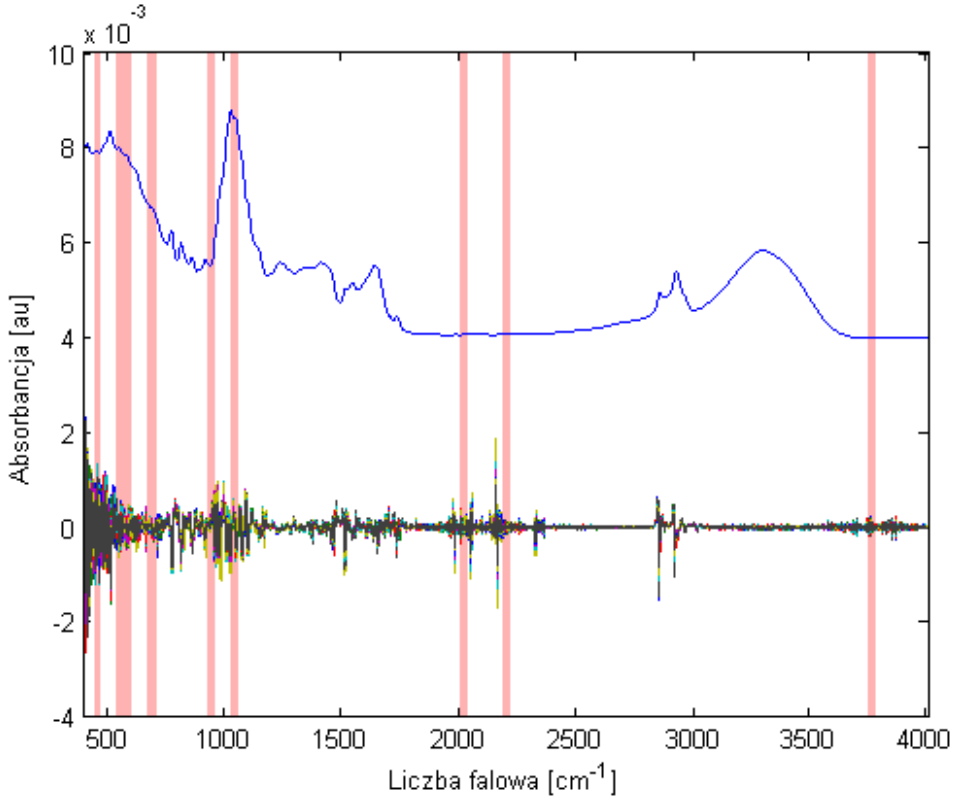
Rysunek 5.6: Krzywa predykcji oznaczeń zawartości aminokwasów egzogennych w próbce



Rysunek 5.7: Względny błąd oznaczeń zawartości aminokwasów egzogennych w próbce



Rysunek 5.8: Wykres RMSCEV dla modelu zawartości aminokwasów egzogennych w próbce



Rysunek 5.9: Wykres zakresów użytych do budowy modelu zawartości aminokwasów egzogennych w próbce

5.3 Modelowanie sumy zawartości aminokwasów

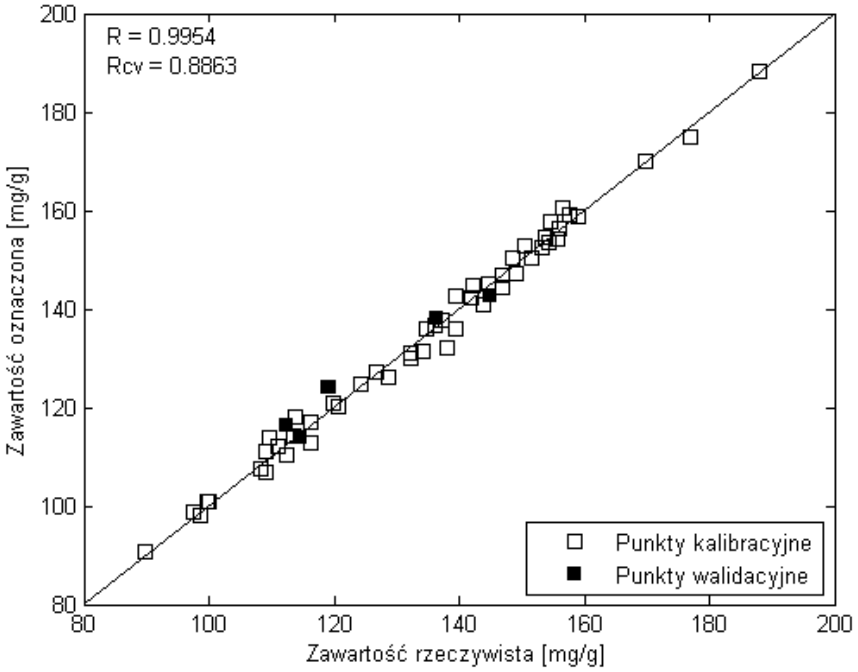
W kolejnym kroku wykonano model ilościowy przeznaczony do oznaczenia sumy stężeń wszystkich aminokwasów obecnych w analizowanych próbkach. Badany zestaw aminokwasów zamieszczono w tabeli 5.4. Opracowany Model charakteryzował się bardzo dobrymi parametrami predykcji (tabela 5.5), walidacji krzyżowej i przebiegu PRESS. Błąd względny oznaczenia mieścił się w przedziale 4%, nie przekraczając 2% dla większości oznaczanych próbek. Warto zwrócić uwagę, że uzyskane parametry jakościowe dla sumy są wyższe niż dla modeli zbudowanych dla poszczególnych aminokwasów, przy jednocześnie największej liczbie użytych zakresów spektralnych przedstawionych na rysunku 5.13. Przebieg PRESS przedstawiono na rysunku 5.12. Zmienność spektralna w badanym zestawie danych dobrze korelowała z wynikami analiz referencyjncyh, dzięki czemu uzyskanie modelu o wysokich parametrach jakościowych wiązało się z relatywnie niskim wysiłkiem obliczeniowym. Powyższe cechy świadczy to o niskim wysiłku obliczeniowym potrzebnym na budowę modelu z uwagi na abundancję informatywnych regionów widma, które dobrze opisują modelowany parametr. Krzywa przewidywania oraz względne błędy oznaczeń przedstawiono na rysunkach 5.10 i 5.11.

Aminokwasy oznaczone w sumie	
Alanina	ALA
Cysteina	CYS
Kwas Asparaginowy	ASP
Kwas Glutaminowy	GLU
Fenyloalanina	PHE
Glicyna	GLY
Histydyna	HIS
Izoleucyna	ILE
Lizyna	LYS
Leucyna	LEU
Metionina	MET
Arginina	ARG
Treonina	THR
Walina	VAL
Tyrozyna	TYR

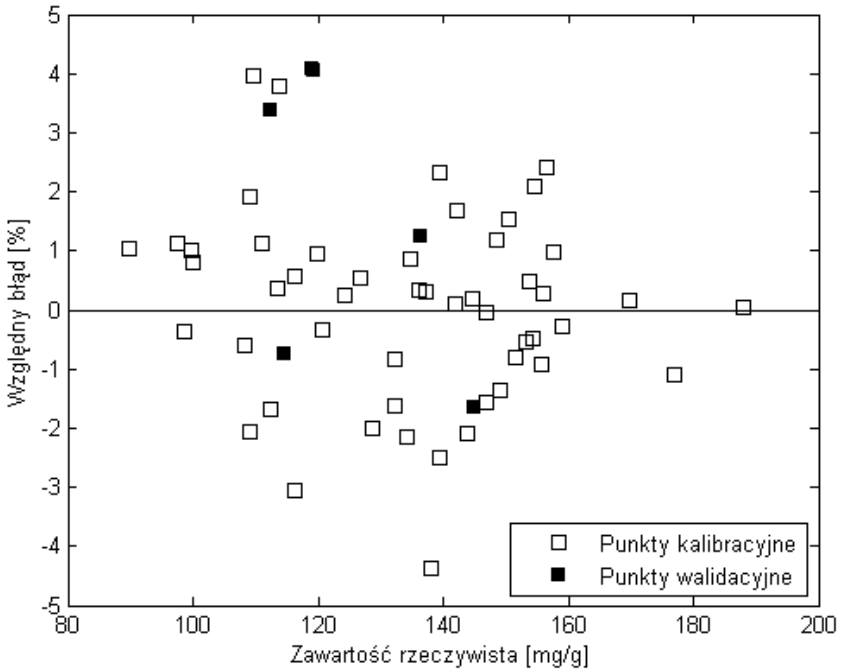
Tablica 5.4: Lista aminokwasów użytych do modelowania sumy aminokwasów

Suma aminokwasów	
Zakres stężeniowy	90-190mg/g
RMSEC	2.09
RMSEP	3.44
Użyte faktory:	6

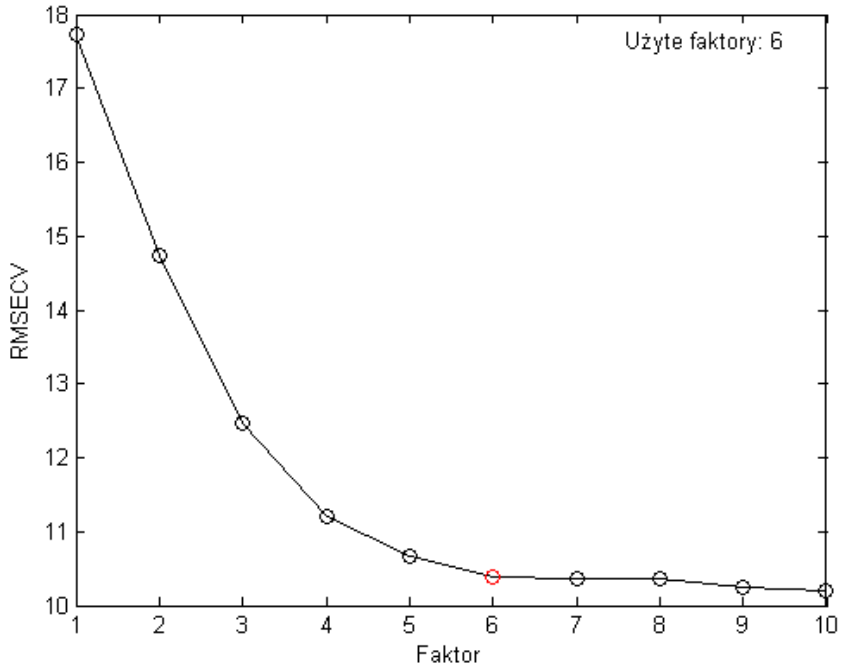
Tablica 5.5: Tabela parametrów dla modelu zawartości sumy aminokwasów



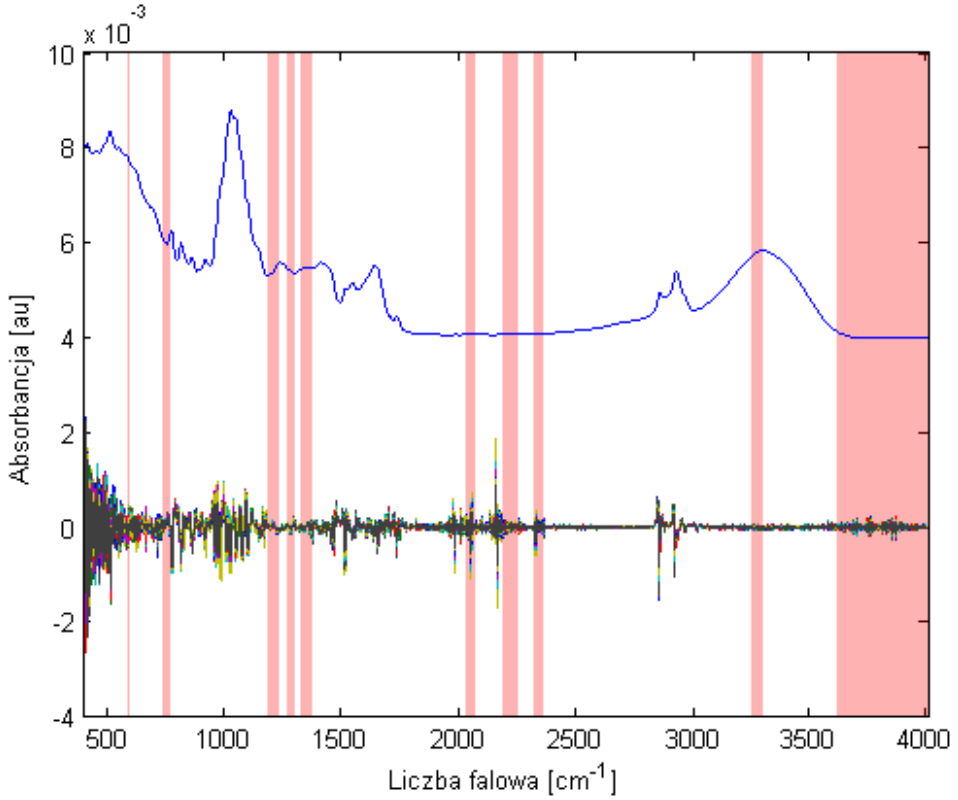
Rysunek 5.10: Krzywa predykcji oznaczeń zawartości sumy aminokwasów w próbce



Rysunek 5.11: Względny błąd oznaczeń zawartości sumy aminokwasów w próbce



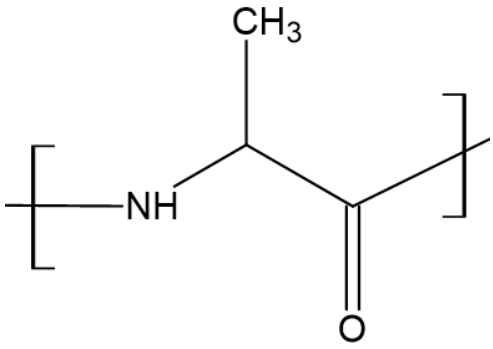
Rysunek 5.12: Wykres RMSCEV dla modelu zawartości sumy aminokwasów w próbce



Rysunek 5.13: Wykres zakresów użytych do budowy modelu zawartości sumy aminokwasów w próbce

5.4 Modelowanie zawartości alaniny

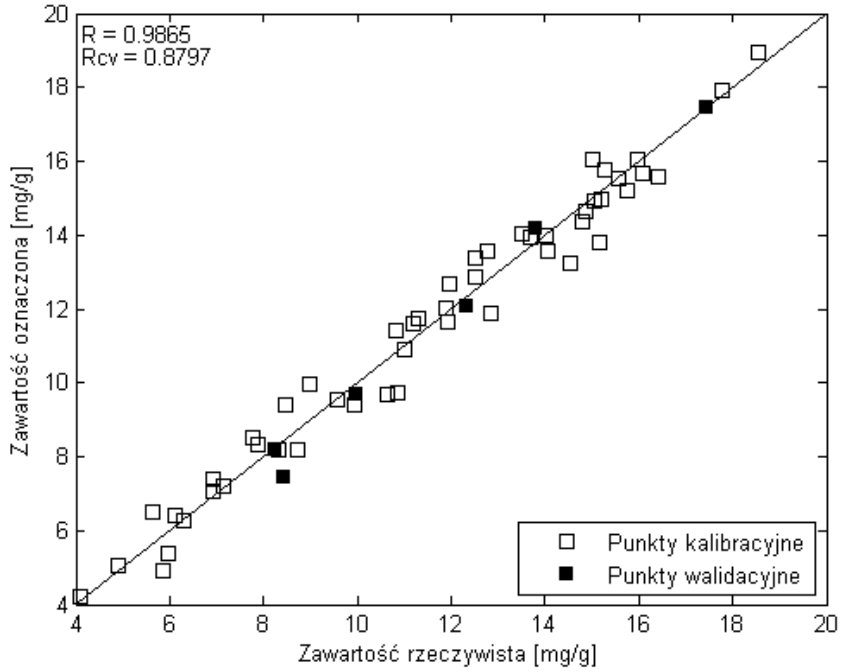
Model ilościowy opracowany przy użyciu ModelHelper opracowany dla alaniny charakteryzował się dobrymi parametrami predykcji i walidacji (tabela 5.6, rysunek 5.15). Jak przedstawiono na rysunku 5.16 błąd względny mieścił się w granicach ±10%, co jest dobrym wynikiem ze względu na małą charakterystyczność łańcucha bocznego alaniny (rys. 5.14) - będzie on w dużej mierze korelował z widmami innych aminokwasów. Dodatkowo na rysunku 5.16 jest widoczny charakterystyczny 'stożek' w układzie błędów predykcji. Świadczy on o większym błędzie metody w przedziale niskich stężeń. Na rysunku 5.18 można zwrócić uwagę na zastosowanie w budowie modelu charakterystycznych dla aminokwasów pasm Amidu I i Amidu A. Z modelu kalibracyjnego wyłączono 7 próbek odbiegających. Przebieg PRESS przedstawiono na rysunku 5.17, przyjmuje on typowy układ osypiska.



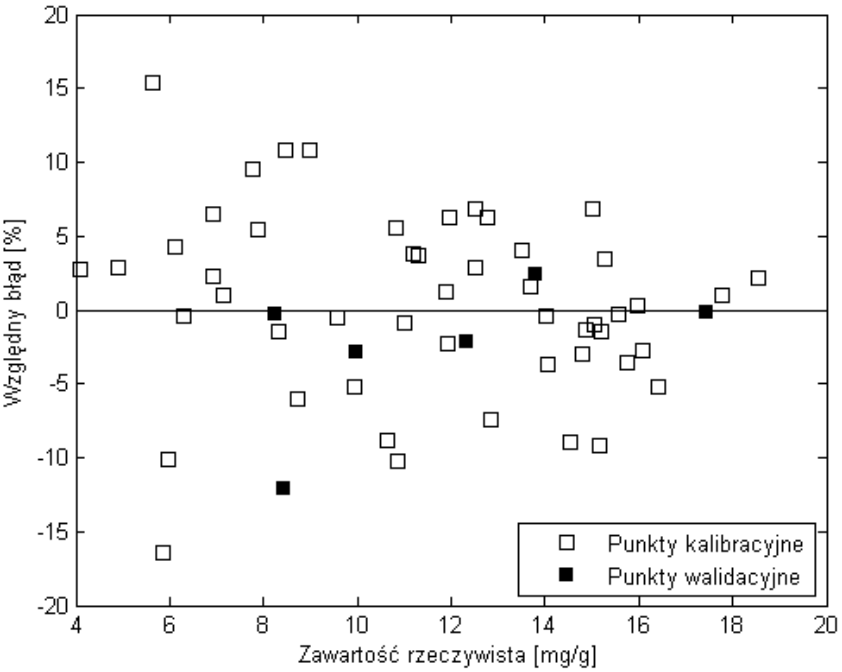
Rysunek 5.14: Struktura alaniny związanej wewnątrz białka

Alanina	
Zakres stężeniowy	4-20mg/g
RMSEC	0.611
RMSEP	0.469
Użyte faktory:	4

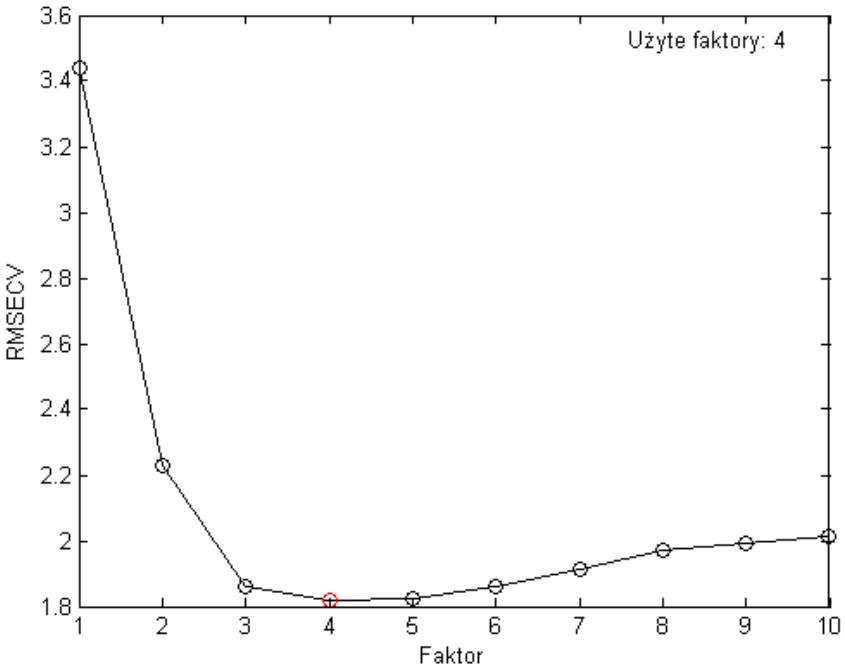
Tablica 5.6: Tabela parametrów dla modelu zawartości alaniny



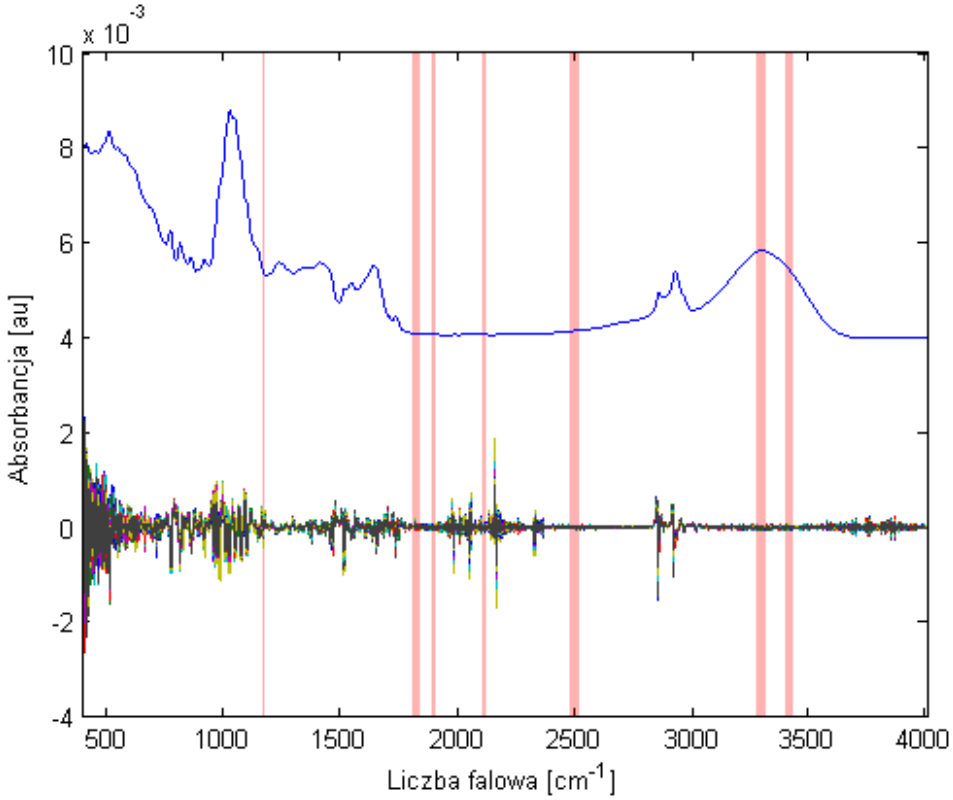
Rysunek 5.15: Krzywa predykcji oznaczeń zawartości alaniny w próbce



Rysunek 5.16: Względny błąd oznaczeń zawartości alaniny w próbce



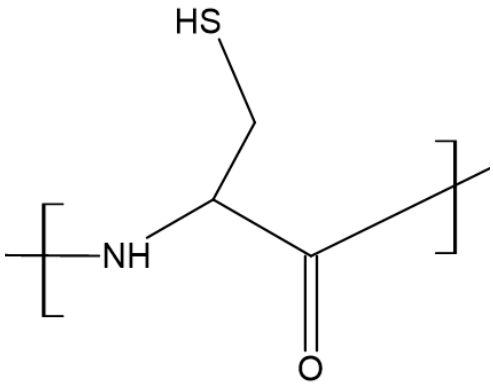
Rysunek 5.17: Wykres RMSCEV dla modelu zawartości alaniny w próbce



Rysunek 5.18: Wykres zakresów użytych do budowy modelu zawartości alaniny w próbce

5.5 Modelowanie zawartości cysteiny

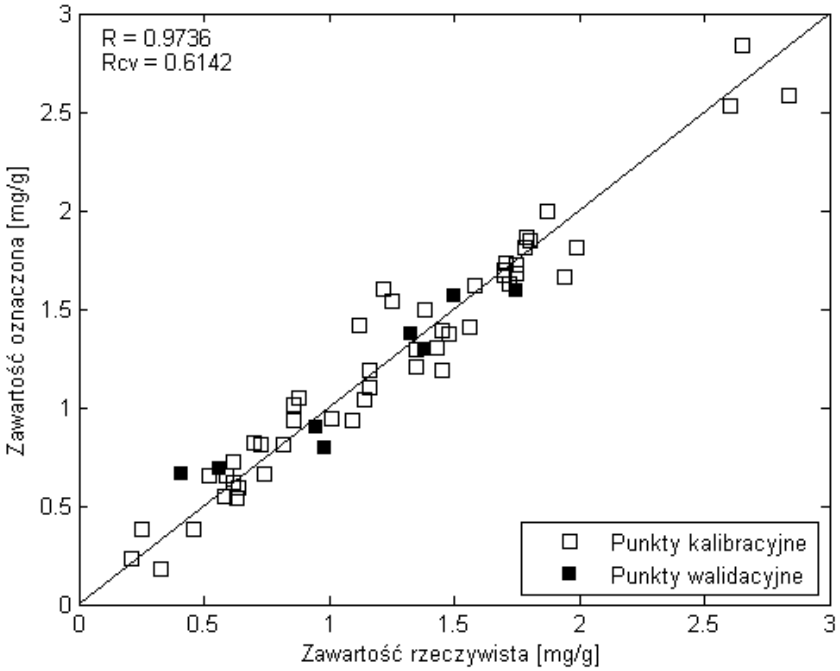
Opracowanie modelu kalibracyjnego dla cysteiny spotkało się z trudnościami, jednak definitywnie jest szansa na jego poprawę przy zastosowaniu szeregu ulepszeń, które wychodziły poza zakres tej pracy. Stworzony model charakteryzował się bardzo wysokim współczynnikiem R, jednak niski parametr walidacji krzyżowej (rysunek 5.20) wskazuje na dużą niestabilność modelu, co poparte jest jego wysokim błędem względnym w granicach 40% (rysunek 5.21) i dużą liczbą próbek odstających. Część z tych problemów wynika z niskiego stężenia tego aminokwasu w badanym zestawie próbek, jak przedstawiono w tabeli 5.7. Dodatkowo analiza zawartości tego aminokwasu w białku jest trudna z uwagi na znaczną liczbę możliwych modyfikacji, takich jak tworzenie mostków wodorowych czy siarczkowych (rysunek 5.19). Trudność w modelowaniu stężenia cysteiny metodami NIR i ATR powtarza się także w innych publikacjach, co sugeruje konieczność zastosowania alternatywnej metody, nastawionej na lepszą czułość detekcji siarki [36, 37]. Wymienione problemy znajdują także swoje odzwierciedlenie w przebiegu PRESS, przedstawionym na rysunku 5.22, który odbiega od prawidłowego. Do budowy modelu użyto dziewięciu zakresów z szerokiego przedziału liczb falowych (rysunek 5.23), co potwierdza występowanie wielu interakcji cysteiny z substancjami zawartymi w próbkach.



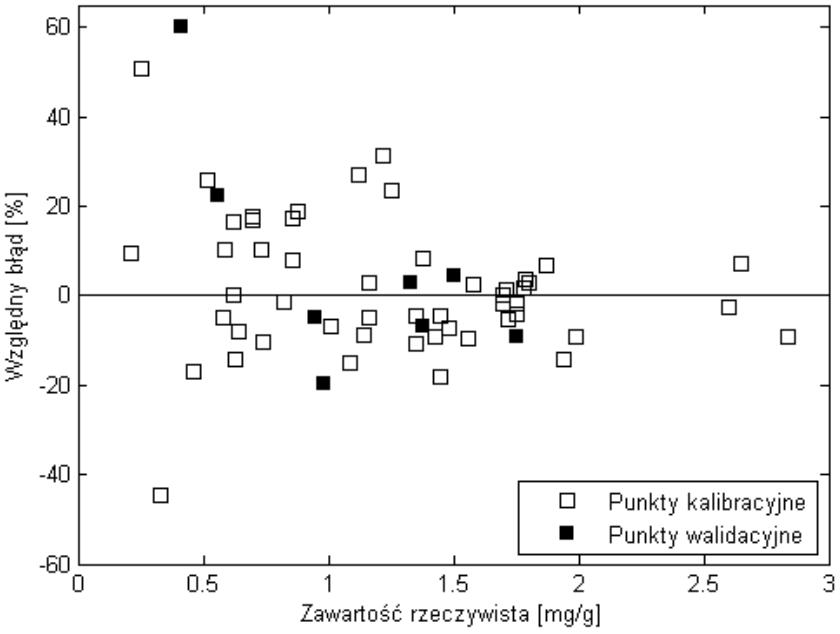
Rysunek 5.19: Struktura cysteiny związanej wewnątrz białka. Notatka: w rysunku nie uwzględniono mostków siarczkowych

Cysteina	
Zakres stężeniowy	0.2 - 2.7mg/g
RMSEC	0.1380
RMSEP	0.140
Użyte faktory:	9

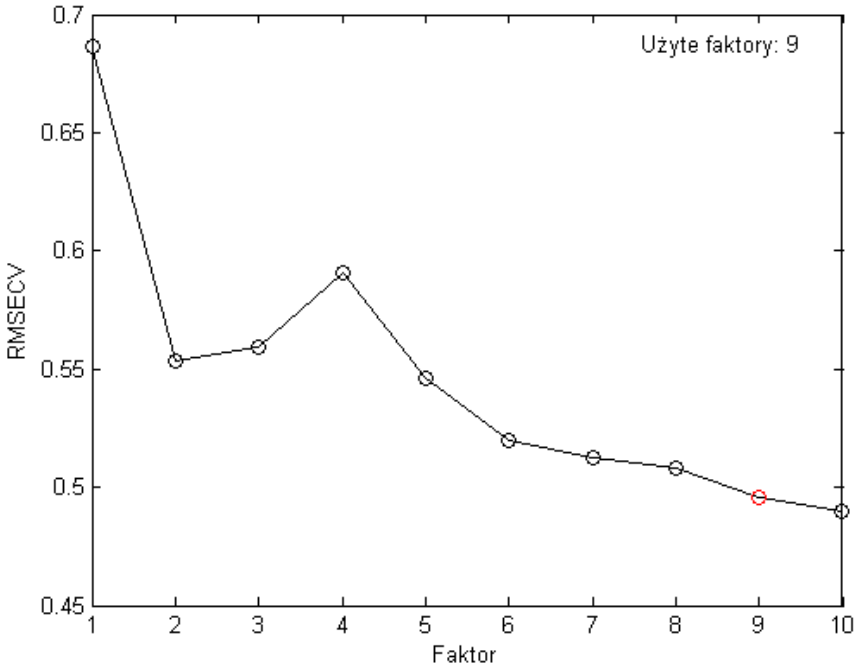
Tablica 5.7: Tabela parametrów dla modelu zawartości cysteiny



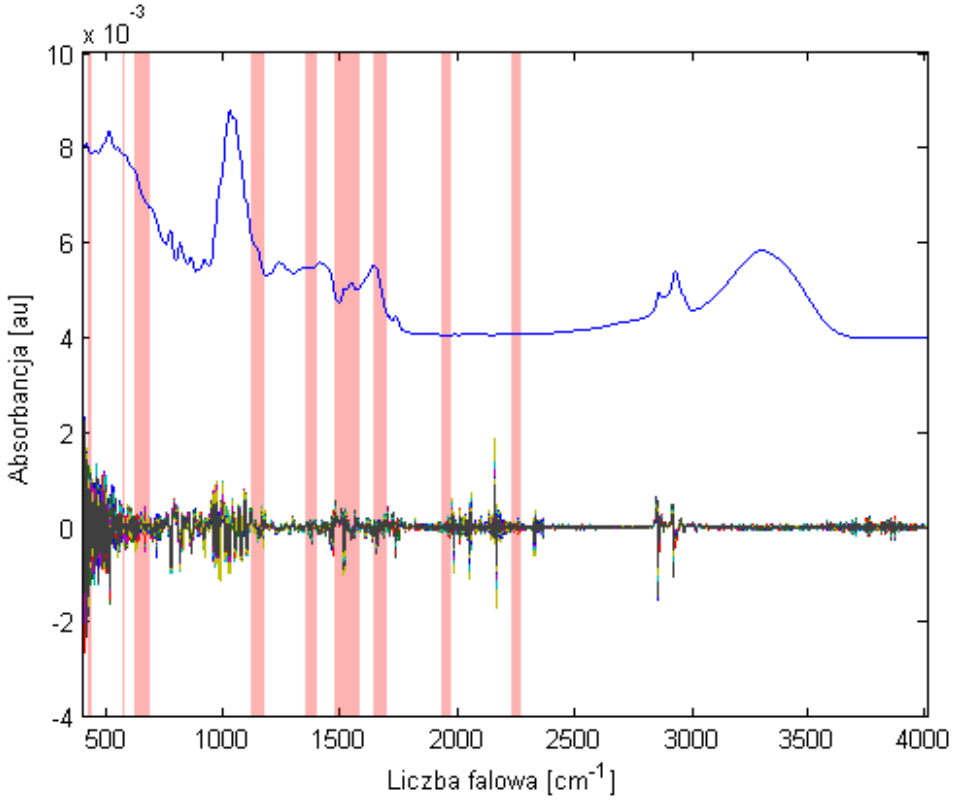
Rysunek 5.20: Krzywa predykcji oznaczeń zawartości cysteiny w próbce



Rysunek 5.21: Względny błąd oznaczeń zawartości cysteiny w próbce



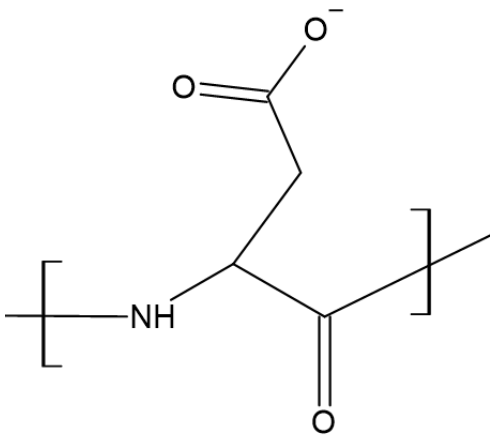
Rysunek 5.22: Wykres RMSCEV dla modelu zawartości cysteiny w próbce



Rysunek 5.23: Wykres zakresów użytych do budowy modelu zawartości cysteiny w próbce

5.6 Modelowanie zawartości kwasu asparaginowego

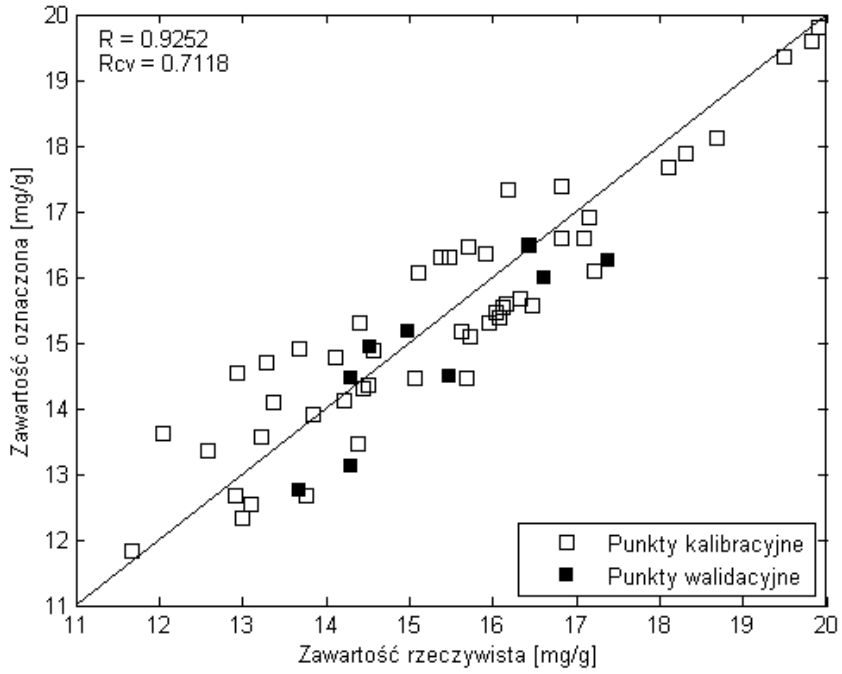
Zastosowanie techniki selekcji danych do opracowania modeli ilościowych dla kwasu asparaginowego (rysunek 5.24) jak i glutaminowego (rysunek 5.29) nie poskutkowało otrzymaniem modeli o najwyższej zdolności prognostycznej (tabela 5.8) i PRESS (rysunek 5.27). Najbardziej prawdopodobną przyczyną tego zjawiska jest silna korelacja widm tych kwasów z ich amidami - asparaginą i glutaminą. Nie udało się jednak tego potwierdzić z uwagi na brak powyższych związków w pomiarach referencyjnych. Potencjalnym czynnikiem, który pozwoliłby na poprawę tych oznaczeń byłoby wykonanie pomiarów dla amidów i próba oddzielenia ich od widm kwasów. Dodatkowym czynnikiem utrudniającym oznaczenie tych związków jest też zapewne duże stężenie kwasów organicznych w badanej próbce, których udziały spektralne mogą zakłócać zmienność w obszarze drgań rozciągających C=O. Błędy względne przedstawoino na rysunku 5.26 mieściły się w przedziale ±10% i nie charakteryzowały się się specyficznym rozkładem. Zakresy użyte do budowy modelu przedstawiono na rysunku 5.28.



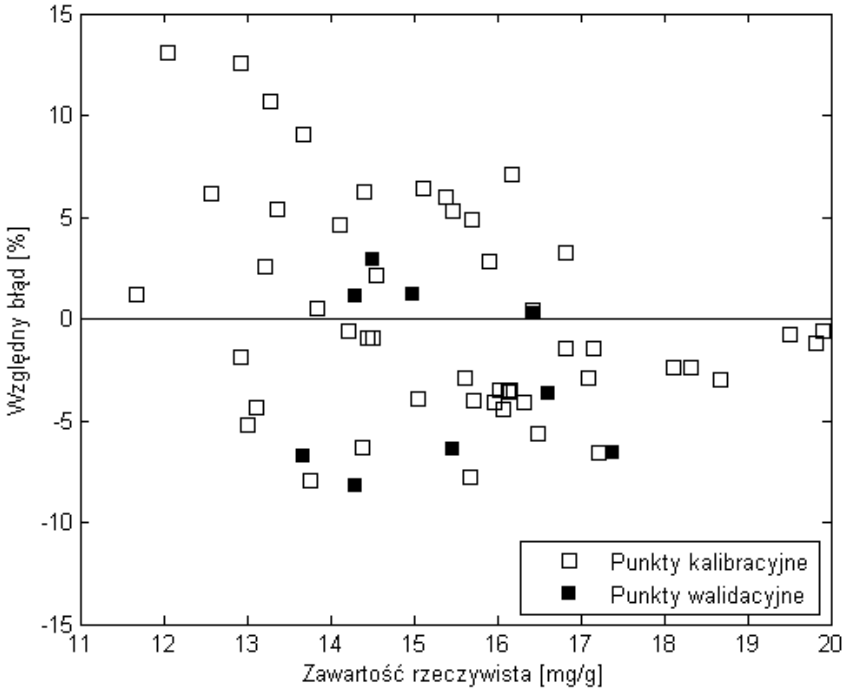
Rysunek 5.24: Struktura kwasu asparaginowego związanego wewnątrz białka

Kwas asparaginowy	
Zakres stężeniowy	11.5-20mg/g
RMSEC	0.744
RMSEP	0.755
Użyte faktory:	3

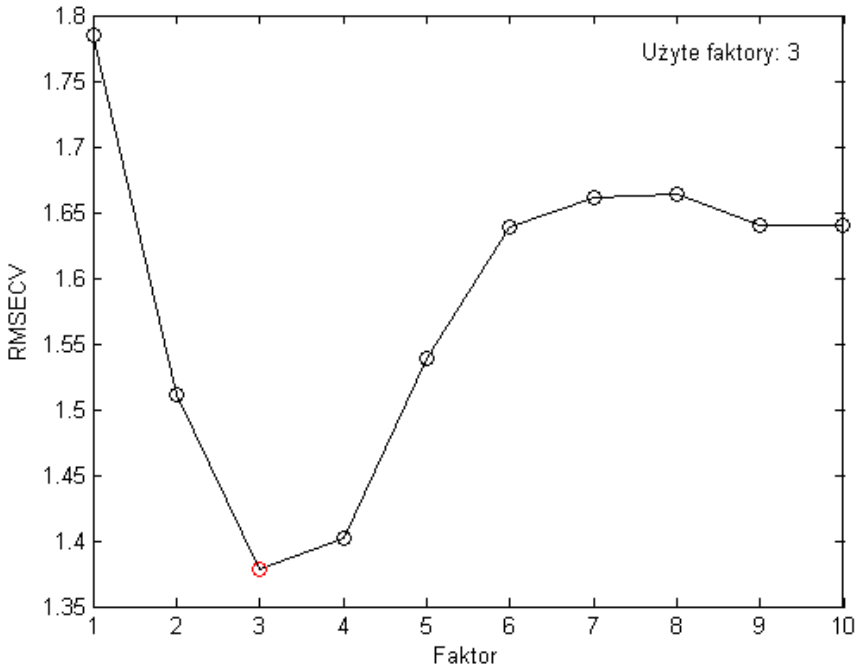
Tablica 5.8: Tabela parametrów dla modelu zawartości kwasu asparaginowego



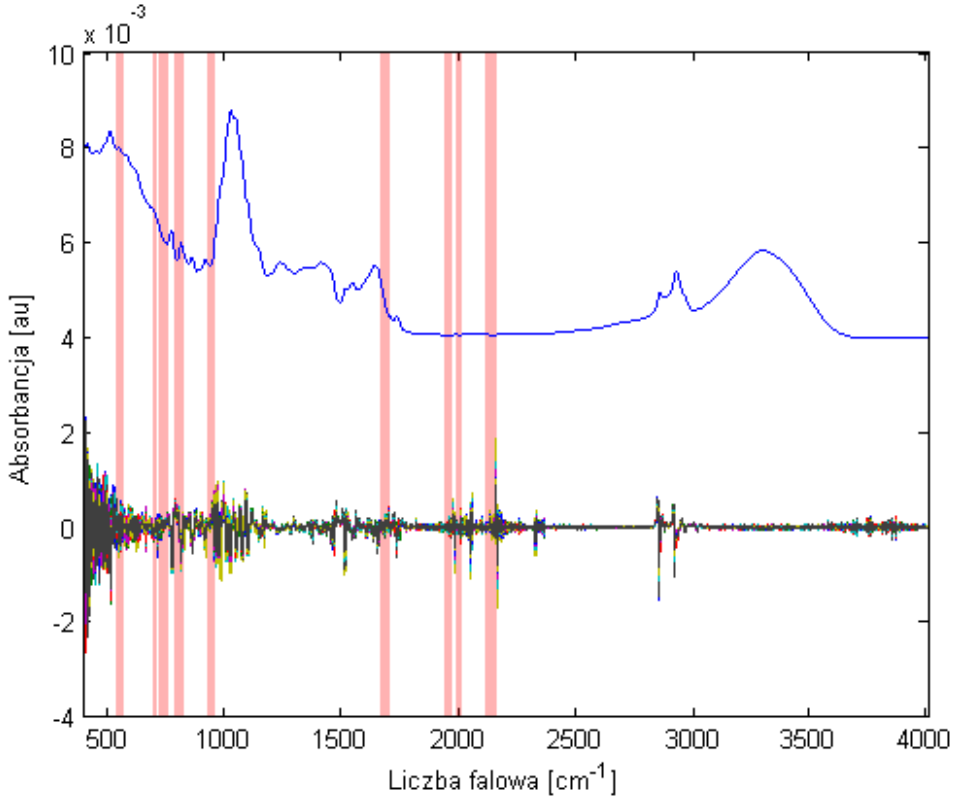
Rysunek 5.25: Krzywa predykcji oznaczeń zawartości kwasu asparaginowego w próbce



Rysunek 5.26: Względny błąd oznaczeń zawartości kwasu asparaginowego w próbce



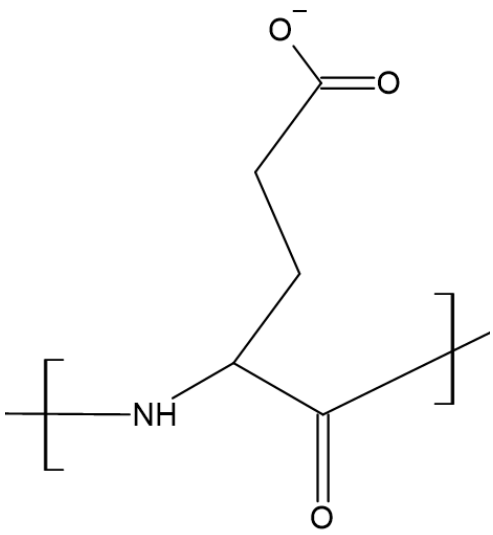
Rysunek 5.27: Wykres RMSCEV dla modelu zawartości kwasu asparaginowego w próbce



Rysunek 5.28: Wykres zakresów użytych do budowy modelu zawartości kwasu asparaginowego w próbce

5.7 Modelowanie zawartości kwasu glutaminowego

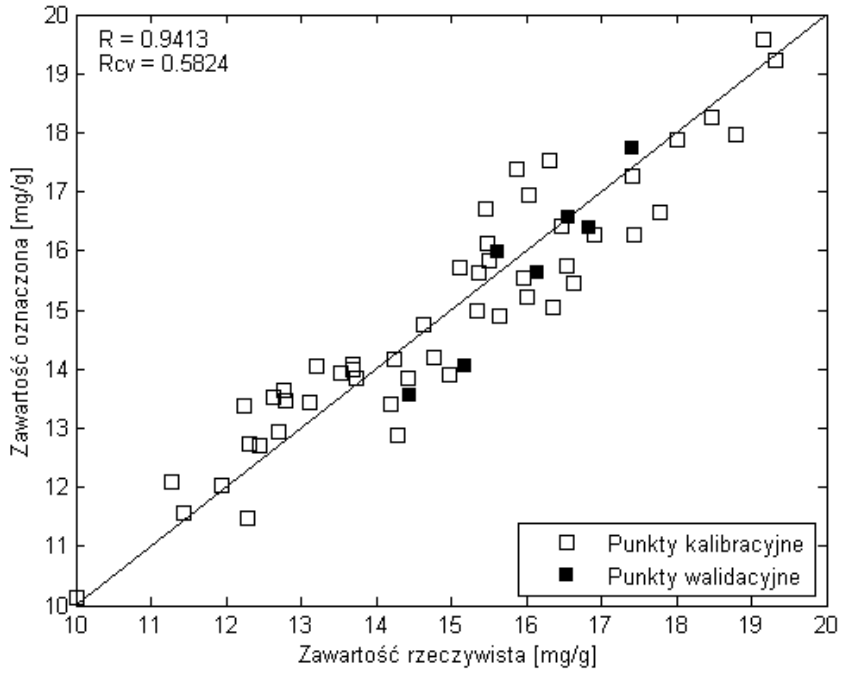
Konstrukcja modelu kalibracyjnego dla kwasu glutaminowego (rysunek 5.29) wiązała się z występowaniem podobnych problemów jak dla opisanego wyżej kwasu asparaginowego (tabela 5.9). Parametr walidacji krzyżowej o wartości $R_{CV} = 0.58$ (rysunek 5.30) świadczy o wysokim stopniu 'overfittingu' modelu. Potwierdza to przebieg PRESS przedstawiony na rysunku 5.32. W modelu zastosowano cztery faktory, mimo że przebieg PRESS sugerowałby użycie trzech. Wynika to z otrzymania lepszych wartości zarówno dla walidacji jak i walidacji krzyżowej przy zastosowaniu czterech faktorów. Błędy względne oznaczeń kwasu glutaminowego przedstawiono na rysunku 5.31.



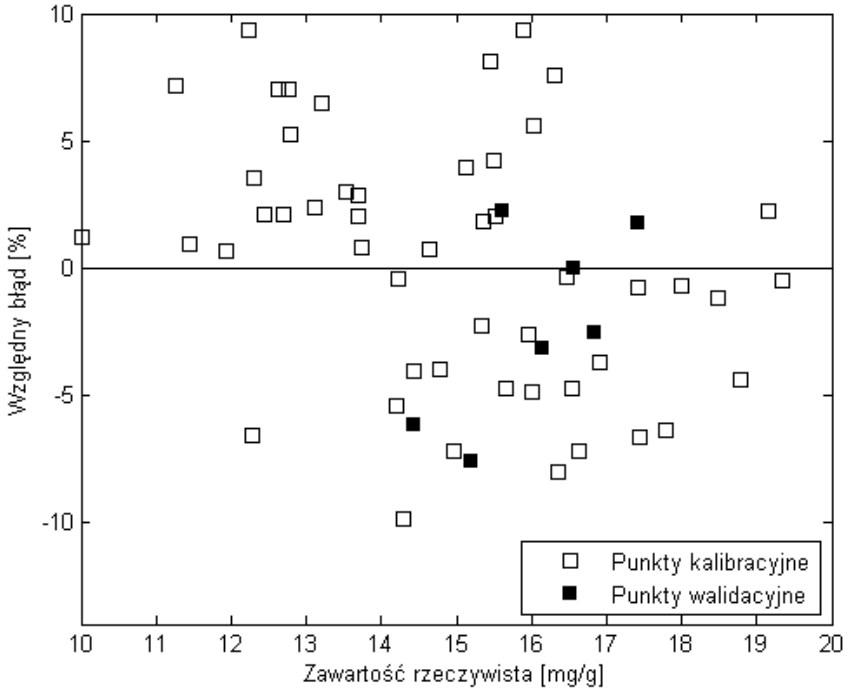
Rysunek 5.29: Struktura kwasu glutaminowego związanego wewnątrz białka

Kwas glutaminowy	
Zakres stężeniowy	10-19.5mg/g
RMSEC	0.735
RMSEP	0.631
Użyte faktory:	4

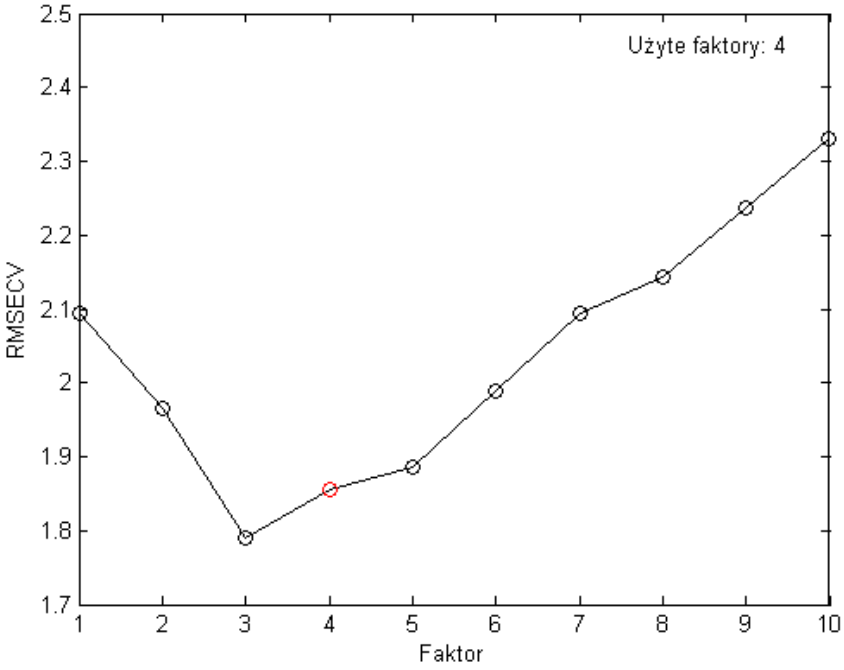
Tablica 5.9: Tabela parametrów dla modelu zawartości kwasu glutaminowego



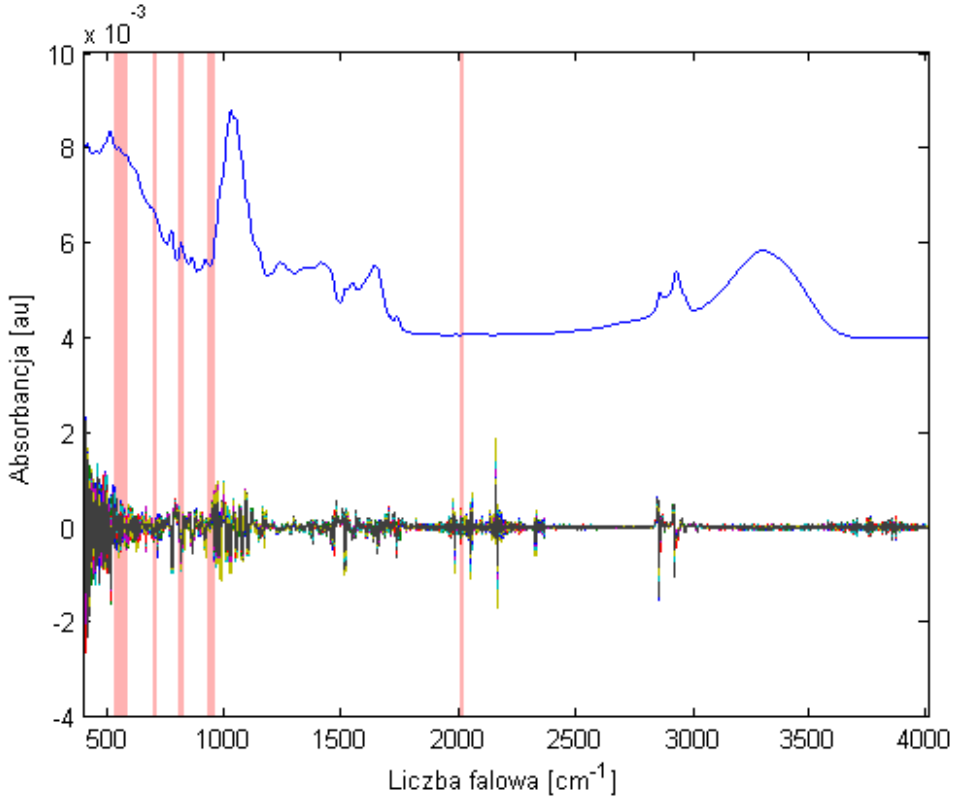
Rysunek 5.30: Krzywa predykcji oznaczeń zawartości kwasu glutaminowego w próbce



Rysunek 5.31: Względny błąd oznaczeń zawartości kwasu glutaminowego w próbce



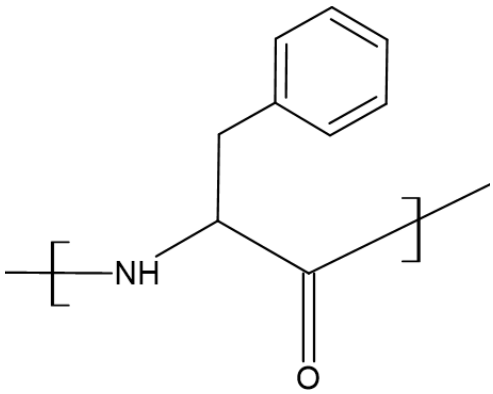
Rysunek 5.32: Wykres RMSCEV dla modelu zawartości kwasu glutaminowego w próbce



Rysunek 5.33: Wykres zakresów użytych do budowy modelu zawartości glutaminowego w próbce

5.8 Modelowanie zawartości fenyloalaniny

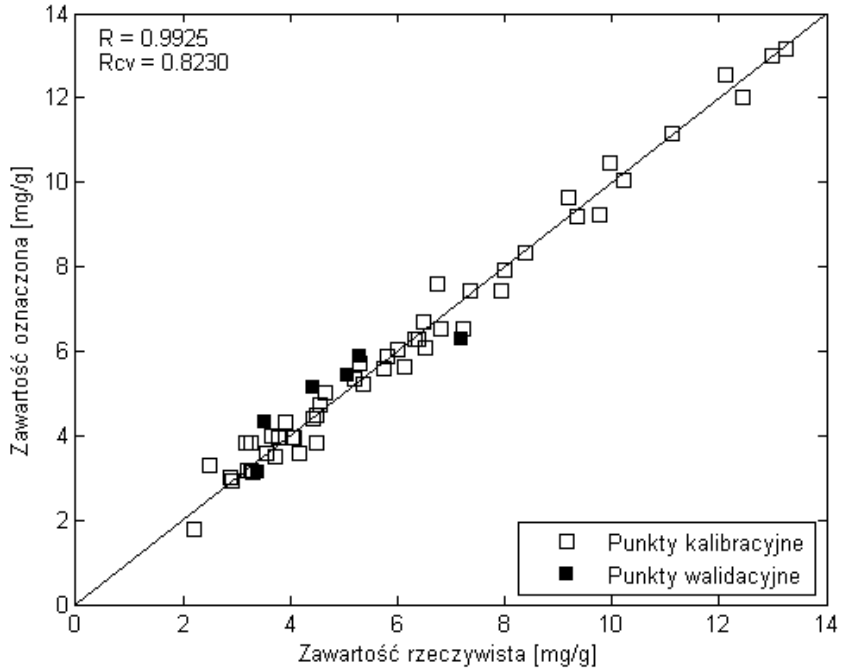
Opracowany przy użyciu algorytmu ModelHelper model ilościowy dla fenyloalaniny charakteryzował się wysoką zdolnością prognostyczną. Było to najprawdopodobniej związane z jej charakterystycznym udziałem spektralnym pochodzącym od drgań łańcucha bocznego zawierającego pierścień aromatyczny (rysunek 5.34) i względnie wysokie stężenie w próbkach (tabela 5.10, rysunek 5.35). Warto zauważyć, że błąd względny (rysunek 5.36) znacząco rośnie dla próbek o niskiej zawartości rzeczywistej, przybierając kształt stożka - wynika to najprawdopodobniej z ograniczeń metody referencyjnej w niskim zakresie stężeń. Przebieg PRESS ma typowy kształt osypiska (rysunek 5.37), a do modelowania użyto sześciu zakresów spektralnych przedstawionych na rysunku 5.38.



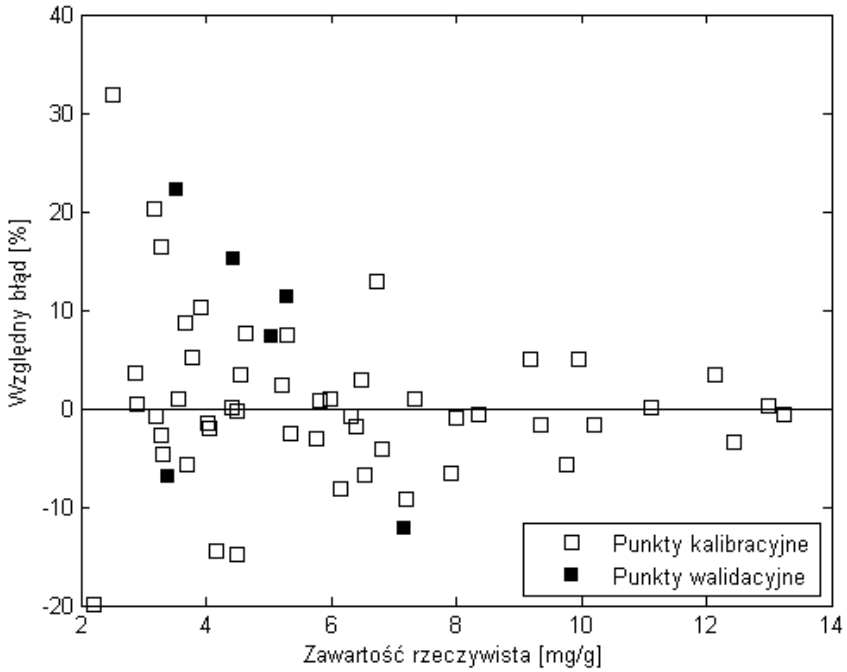
Rysunek 5.34: Struktura fenyloalaniny związanej wewnątrz białka

Fenyloalanina	
Zakres stężeniowy	2-14mg/g
RMSEC	0.357
RMSEP	0.631
Użyte faktory:	7

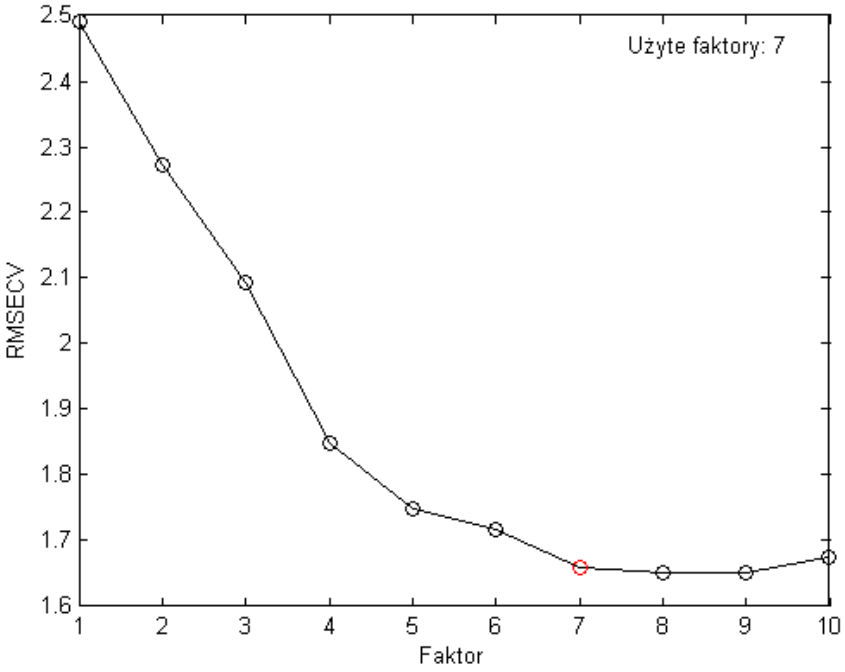
Tablica 5.10: Tabela parametrów dla modelu zawartości fenyloalaniny



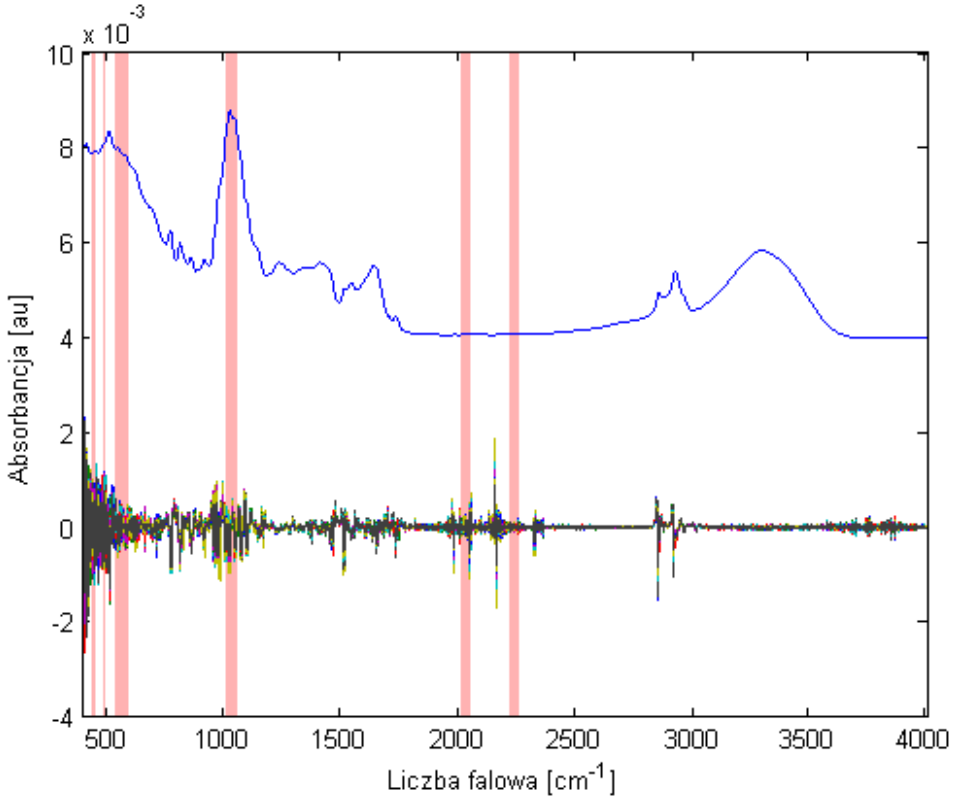
Rysunek 5.35: Krzywa predykcji oznaczeń zawartości fenyloalaniny w próbce



Rysunek 5.36: Względny błąd oznaczeń zawartości fenyloalaniny w próbce



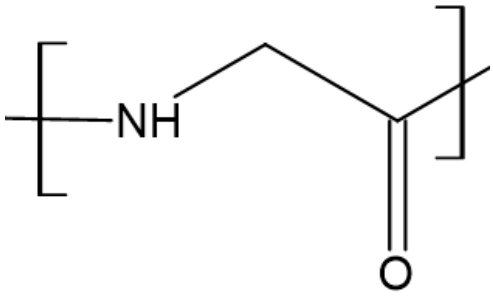
Rysunek 5.37: Wykres RMSCEV dla modelu zawartości fenyloalaniny w próbce



Rysunek 5.38: Wykres zakresów użytych do budowy modelu zawartości fenyloalaniny w próbce

5.9 Modelowanie zawartości glicyny

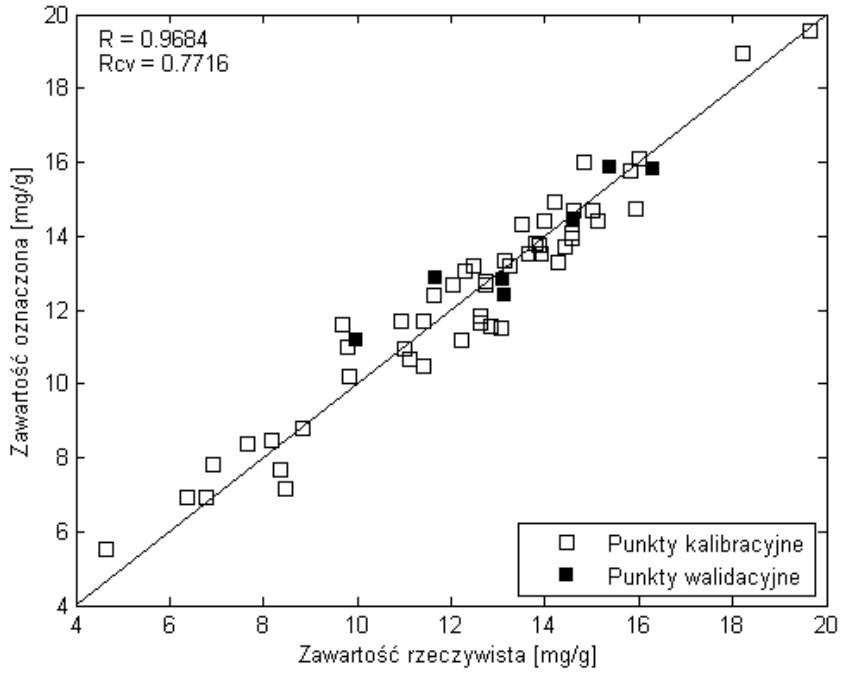
Widmo glicyny, podobnie jak alaniny, jest mało charakterystyczne, co prawdopodobnie wynika z braku łańcucha bocznego (rysunek 5.39). Utrudniało to opracowanie modelu przewidującego stężenie tego aminokwasu. Na przebiegu PRESS widoczne są niewielkie fluktuacje dla wyższych faktorów PLS, co przedstawiono na rysunku 5.42. W toku modelowania udało się uzyskać bardzo dobre rezultaty dla kalibracji i walidacji krzyżowej (tabela 5.11, rysunek 5.40). Uzyskany błąd względny na poziomie ±15% (rysunek 5.41) jest zadowalający biorąc pod uwagę dokładność metody referencyjnej. Ponownie można tutaj zaobserwować charakterystyczny rozkład względnych błędów oznaczeń układających się na kształt stożka. Świadczą one o słabej dokładności metody w zakresie niskich stężeń.



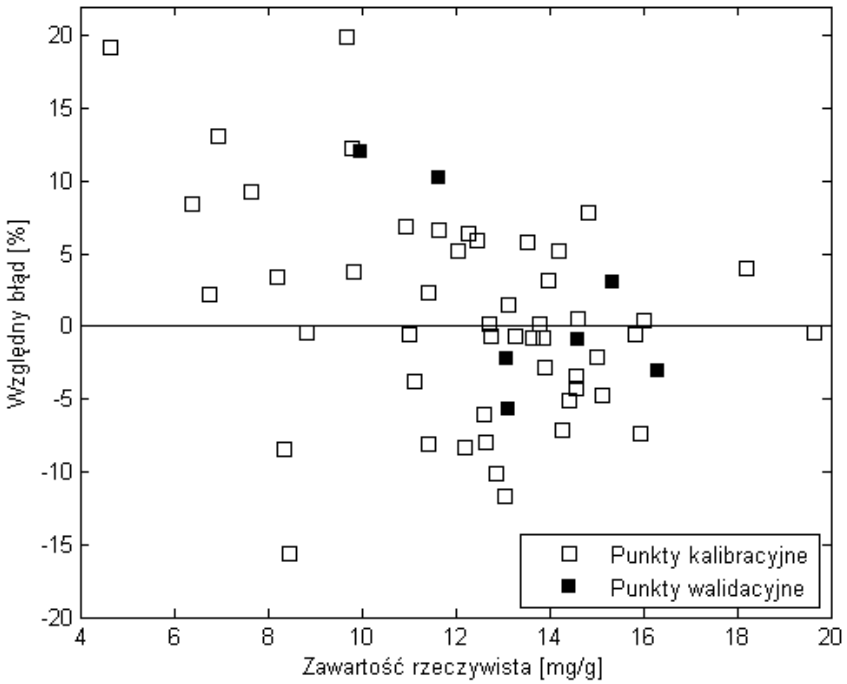
Rysunek 5.39: Struktura glicyny związanej wewnątrz białka

Glicyna	
Zakres stężeniowy	4-20mg/g
RMSEC	1.140
RMSEP	2.48
Użyte faktory:	6

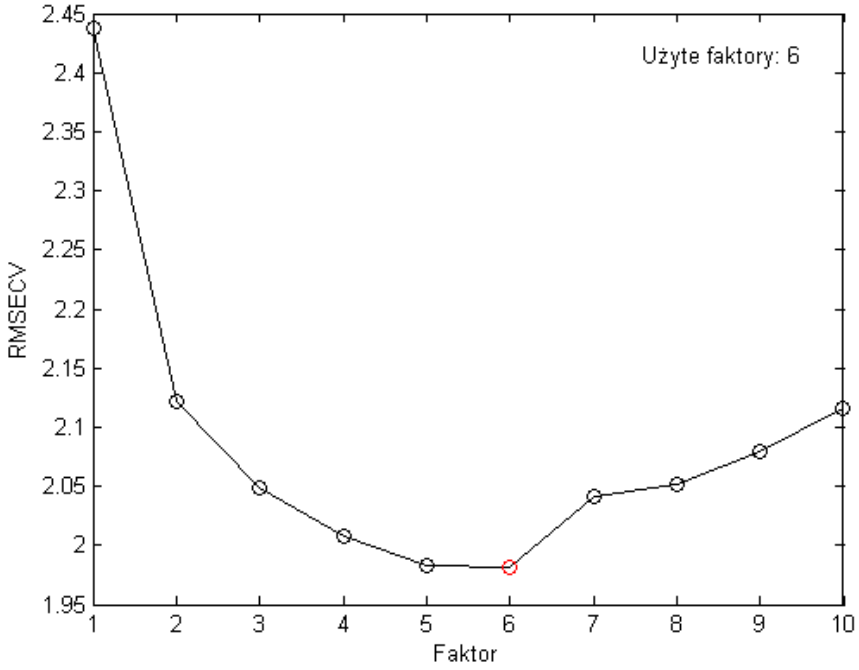
Tablica 5.11: Tabela parametrów dla modelu zawartości glicyny



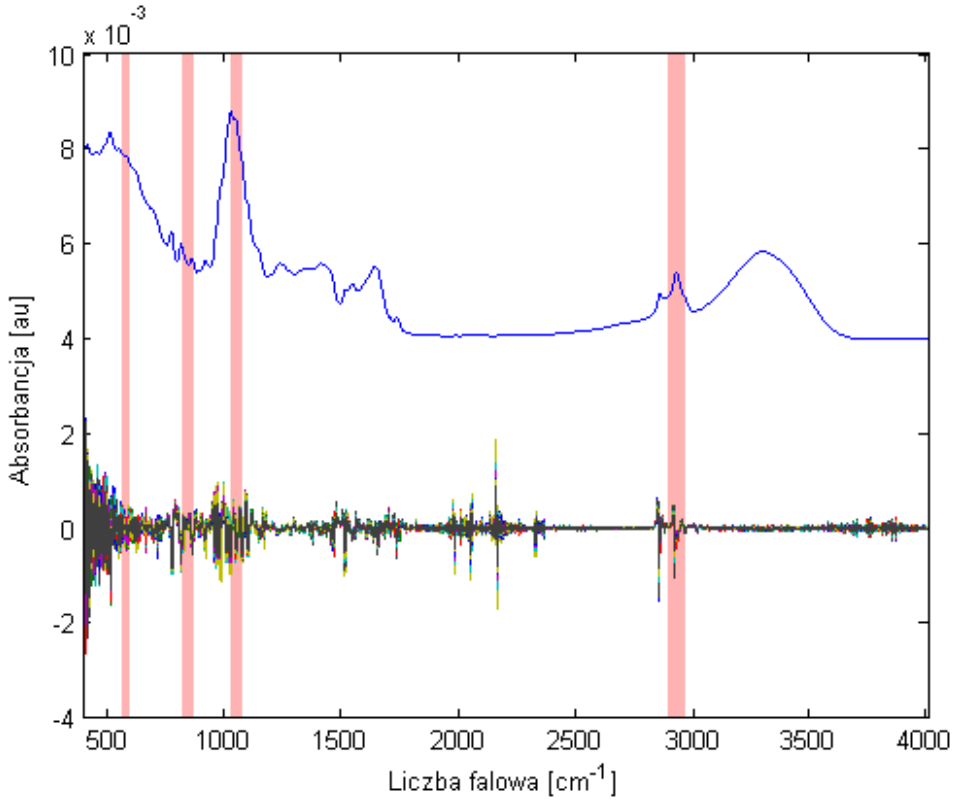
Rysunek 5.40: Krzywa predykcji oznaczeń zawartości glicyny w próbce



Rysunek 5.41: Względny błąd oznaczeń zawartości glicyny w próbce



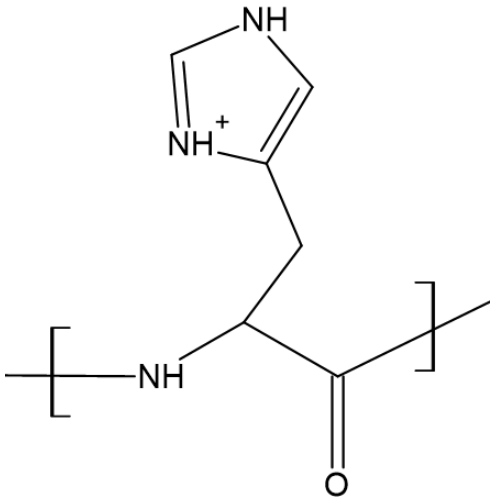
Rysunek 5.42: Wykres RMSCEV dla modelu zawartości glicyny w próbce



Rysunek 5.43: Wykres zakresów użytych do budowy modelu zawartości glicyny w próbce

5.10 Modelowanie zawartości histydyny

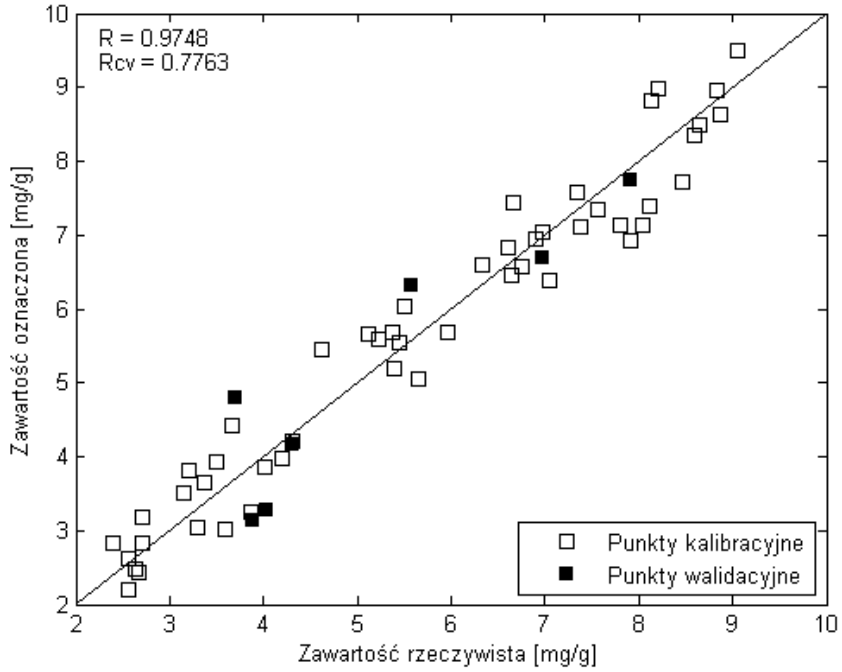
Opracowany model predykcyjny dla histydyny, której strukturę przedstawiono na rysunku 5.44, charakteryzuje się bardzo dobrymi parametrami jakości, zawartymi w tabeli 5.12 oraz wysokim parametrem walidacji krzyżowej, o wartości 0.78 (rysunek 5.45). Błędy względne przedstawione na rysunku 5.46 ponownie układają się w charakterystyczny kształt stożka, informując o malejącej zdolności prognostycznej metody przy niższych stężeniach, co wynika z niższego stosunku sygnału do szumu i prawdopodobnie niższej dokładności oznaczeń referencyjnych dla niskich stężeń tego aminokwasu. Problem ten znajduje swoje odzwierciedlenie w przebiegu PRESS (rysunek 5.47), gdzie błędy walidacji są zauważalnie wyższe powyżej 5 faktorów. Zakresy użyte do budowy opisanego modelu przedstawiono na rysunku 5.48.



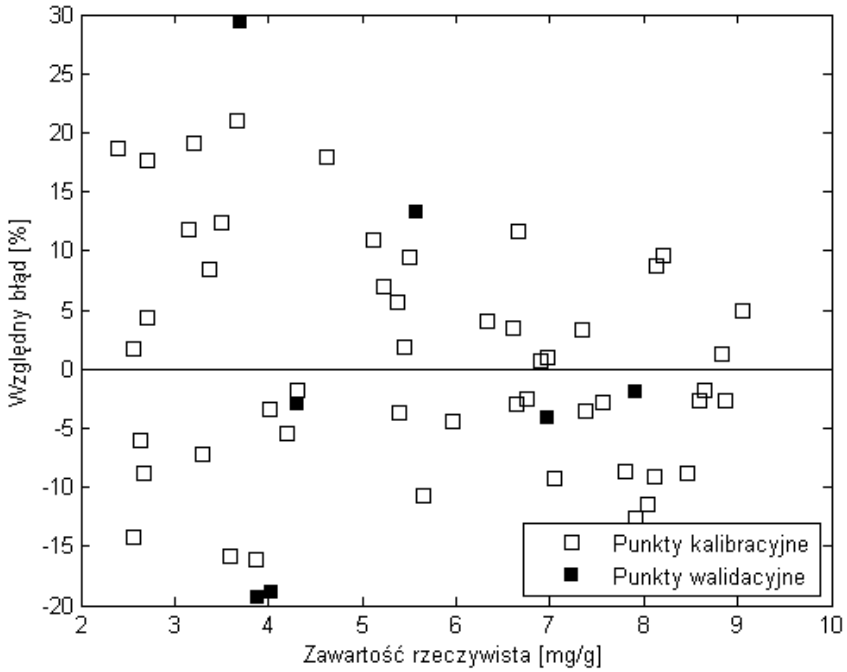
Rysunek 5.44: Struktura histydyny związanej wewnątrz białka

Histydyna	
Zakres stężeniowy	2.5-9mg/g
RMSEC	0.472
RMSEP	0.655
Użyte faktory:	4

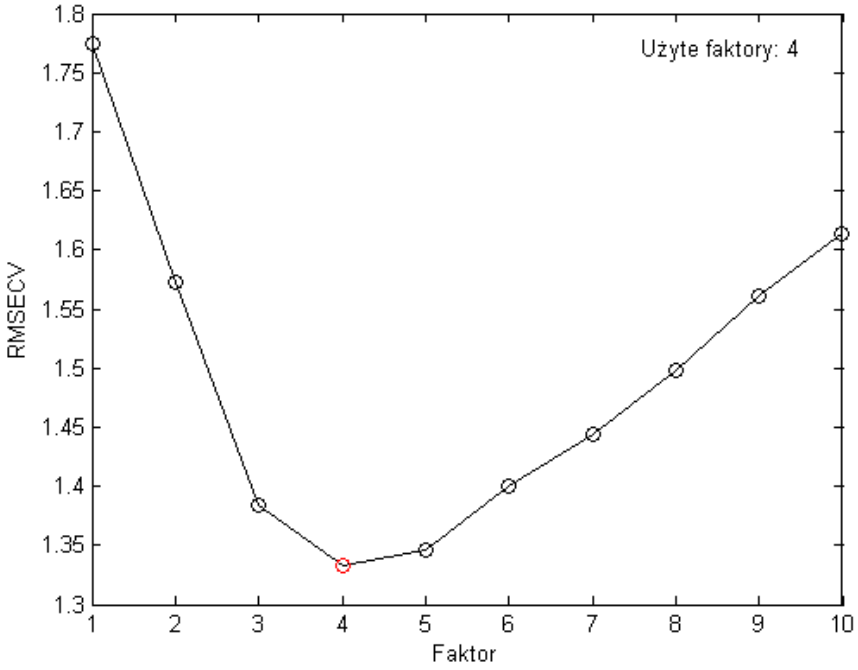
Tablica 5.12: Tabela parametrów dla modelu zawartości histydyny



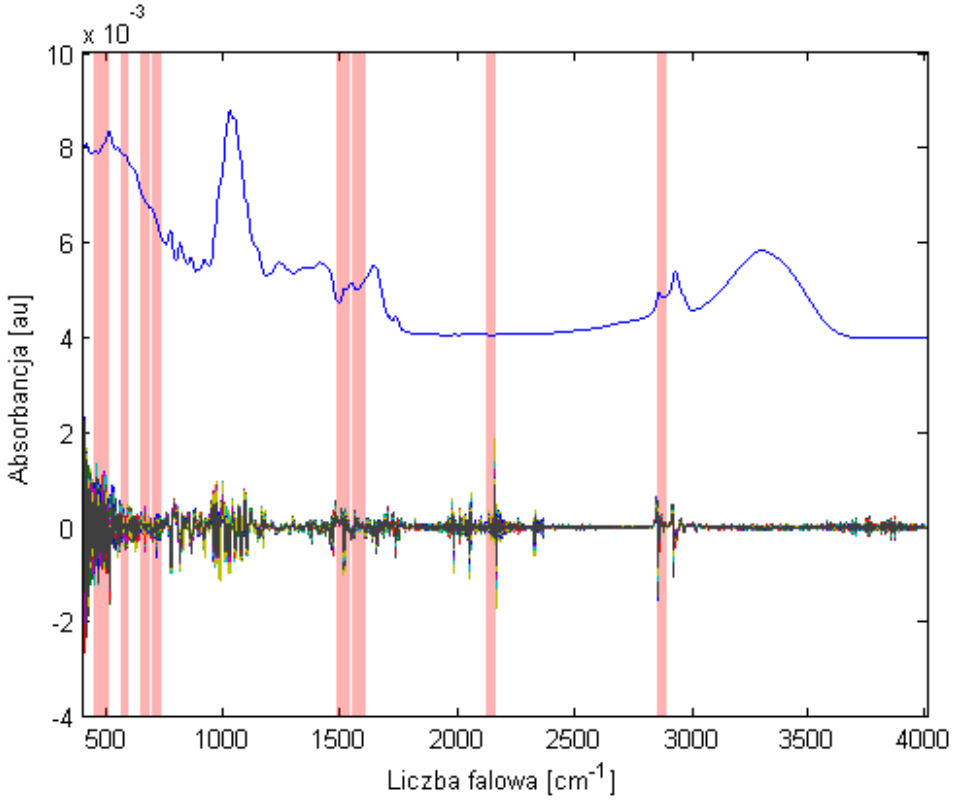
Rysunek 5.45: Krzywa predykcji oznaczeń zawartości histydyny w próbce



Rysunek 5.46: Względny błąd oznaczeń zawartości histydyny w próbce



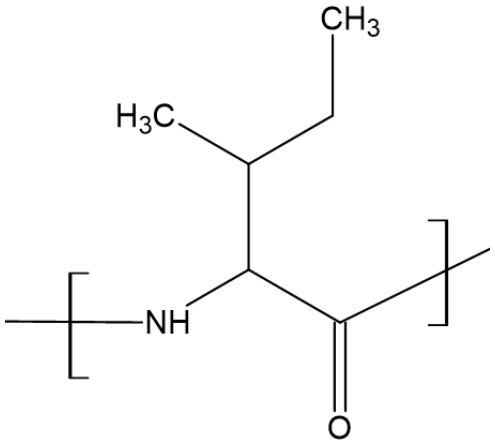
Rysunek 5.47: Wykres RMSCEV dla modelu zawartości histydyny w próbce



Rysunek 5.48: Wykres zakresów użytych do budowy modelu zawartości histydyny w próbce

5.11 Modelowanie zawartości izoleucyny

Modele dla izoleucyny (rysunek 5.49) i leucyny (rysunek 5.59) charakteryzowały się jednymi z najsłabszych parametrów jakości (tabela 5.13) uzyskanych w przebiegu badań. Wynika to z bardzo silnej korelacji ich stężeń i podobieństwa widm (rysunek 3.1), co utrudnia skuteczną separację ich zmienności na etapie konstrukcji modelu PLS. Rezultatem jest brak poprawy jakości modelu przy dodawaniu kolejnych zakresów, co widać na rysunku 5.53 (jako, że wszystkie się pokrywają) i znaczne pogorszenie parametrów walidacji krzyżowej (rysunek 5.50) i przebiegu PRESS (rysunek 5.52). W nadziei na uzyskanie stabilniejszego modelu wykonano procedurę dla większego zestawu próbek kalibracyjnych, ale działania te, co widać na rysunku 5.51, nie przyniosły znaczącej poprawy jakości modelu.

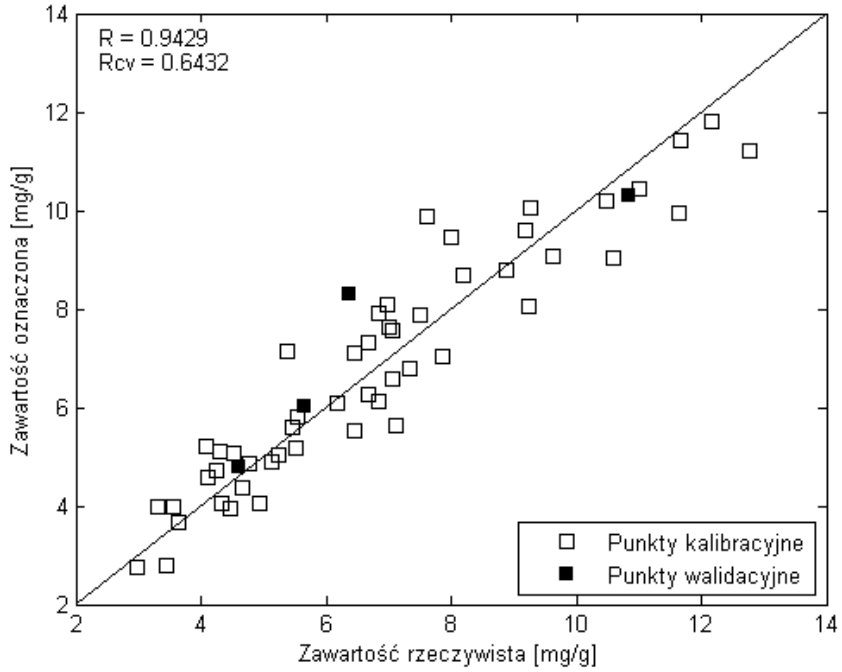


Rysunek 5.49: Struktura izoleucyny związanej wewnątrz białka

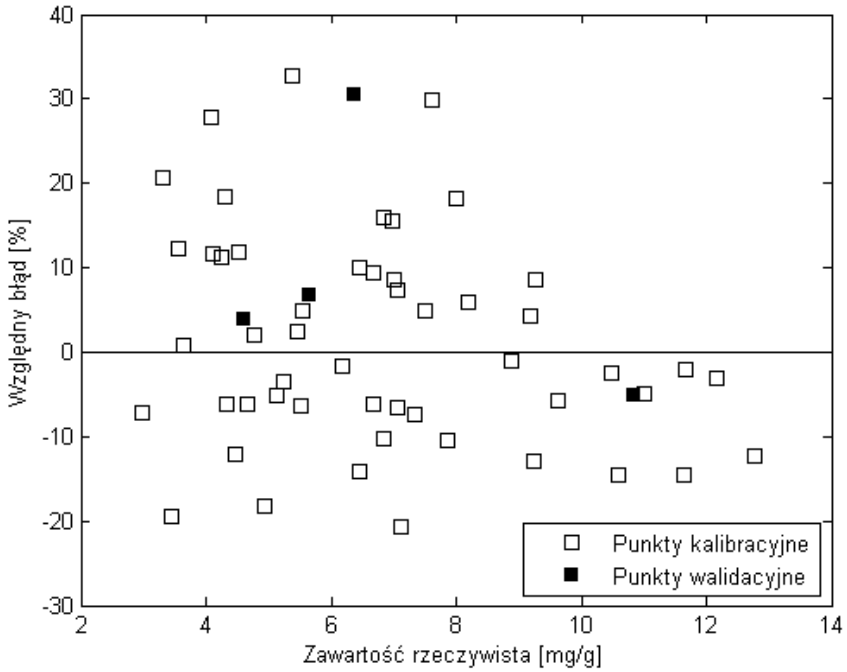
Problem korelacji rozwiązano poprzez wykonanie wspólnego modelu ILE+LEU, w którym modelowano sumę ich stężeń w badanych próbkach. Model ten zostanie opisany w dalszej części pracy.

Izoleucyna	
Zakres stężeniowy	3-13mg/g
RMSEC	0.835
RMSEP	1.03
Użyte faktory:	7

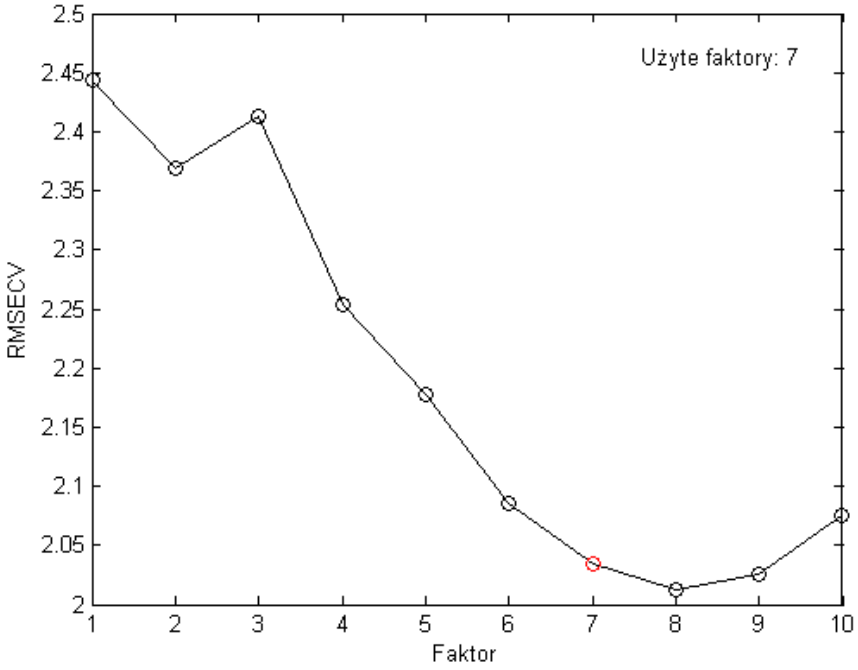
Tablica 5.13: Tabela parametrów dla modelu zawartości izoleucyny



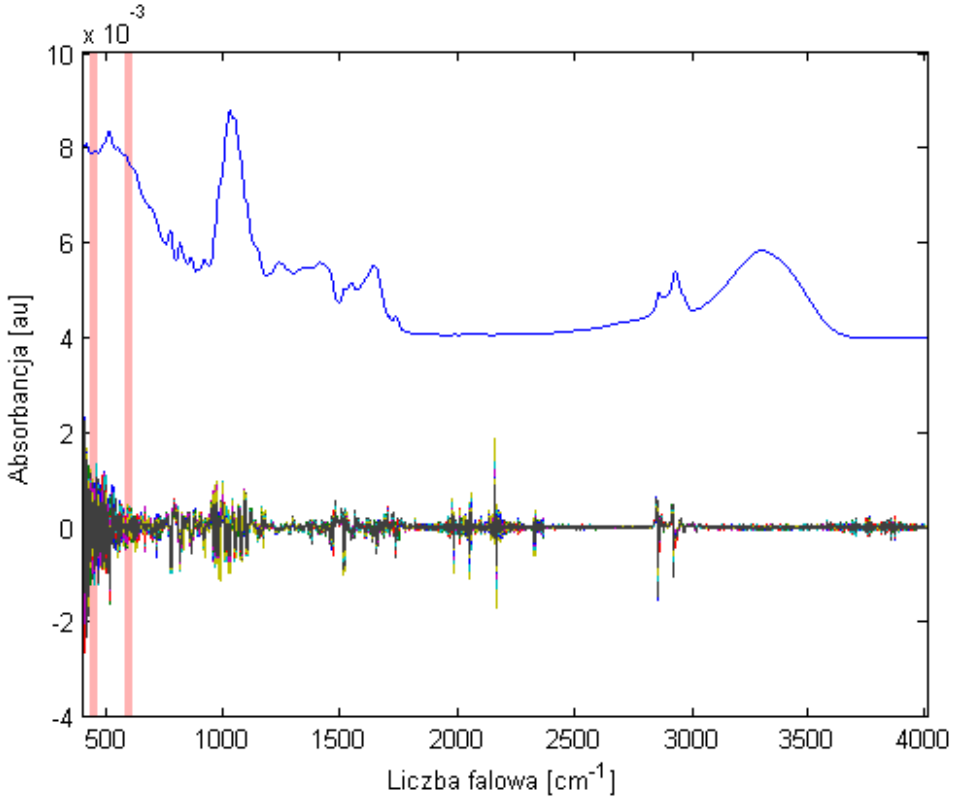
Rysunek 5.50: Krzywa predykcji oznaczeń zawartości izoleucyny w próbce



Rysunek 5.51: Względny błąd oznaczeń zawartości izoleucyny w próbce



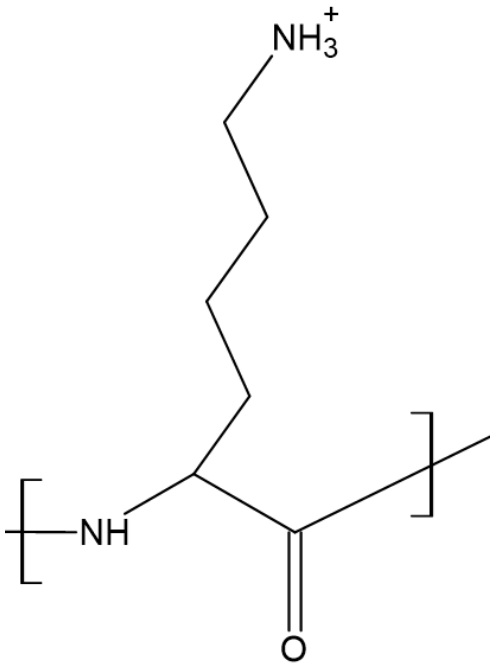
Rysunek 5.52: Wykres RMSCEV dla modelu zawartości izoleucyny w próbce



Rysunek 5.53: Wykres zakresów użytych do budowy modelu zawartości izoleucyny w próbce

5.12 Modelowanie zawartości lizyny

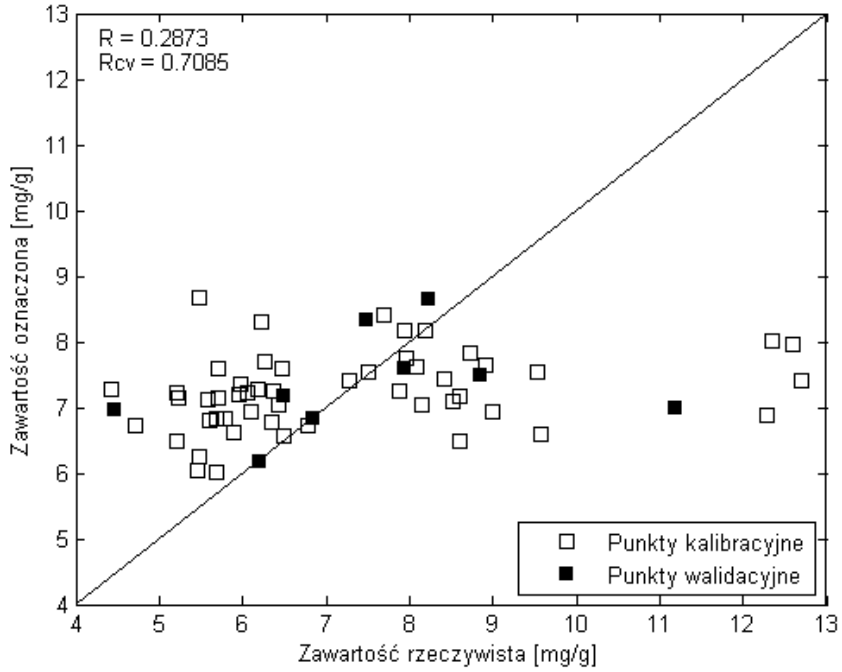
W przypadku modelowania stężeń lizyny (rysunek 5.54) w badanych próbkach pyłku pszczelego algorytm zajmujący się wyborem i optymalizacją zakresów spektralnych nie był w stanie znaleźć obszarów w widmie korelujących z jej stężeniem. Nie udało się ustalić przyczyny tego zjawiska - możliwe jest tutaj pełne przysłonięcie udziałów spektralnych lizyny przez inne związki obecne w próbce lub błędne pomiary referencyjne. Jest to jedyny aminokwas, dla którego nie powiódł się proces kalibracji. Rysunki 5.55, 5.56, 5.57, 5.58 przedstawiają krzywe predykcji, błędów względnych, przebiegu PRESS i użytych zakresów dla tego modelu. Tabela 5.14 zawiera parametry jakości modelu.



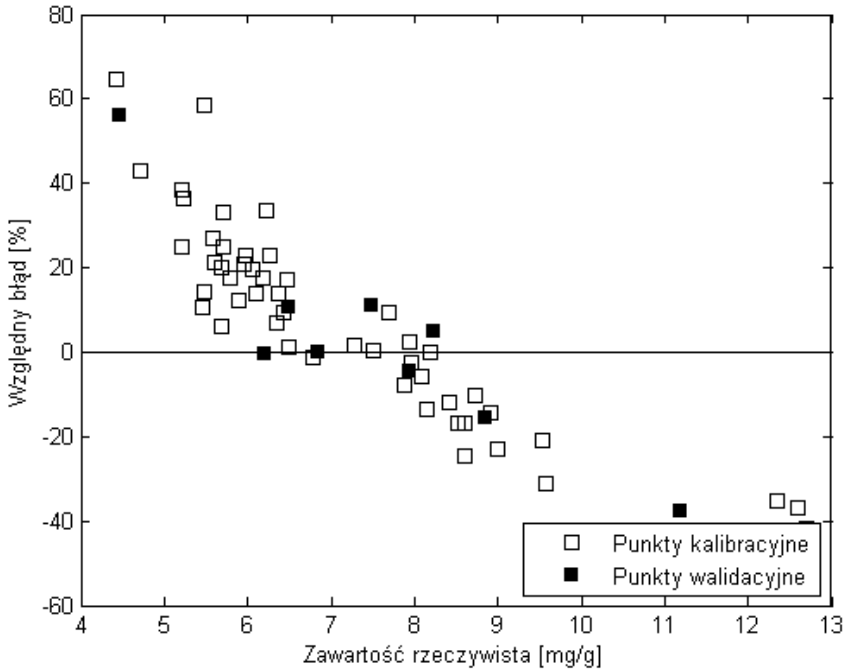
Rysunek 5.54: Struktura lizyny związanej wewnątrz białka

Lizyna	
Zakres stężeniowy	4-13mg/g
RMSEC	1.93
RMSEP	1.74
Użyte faktory:	4

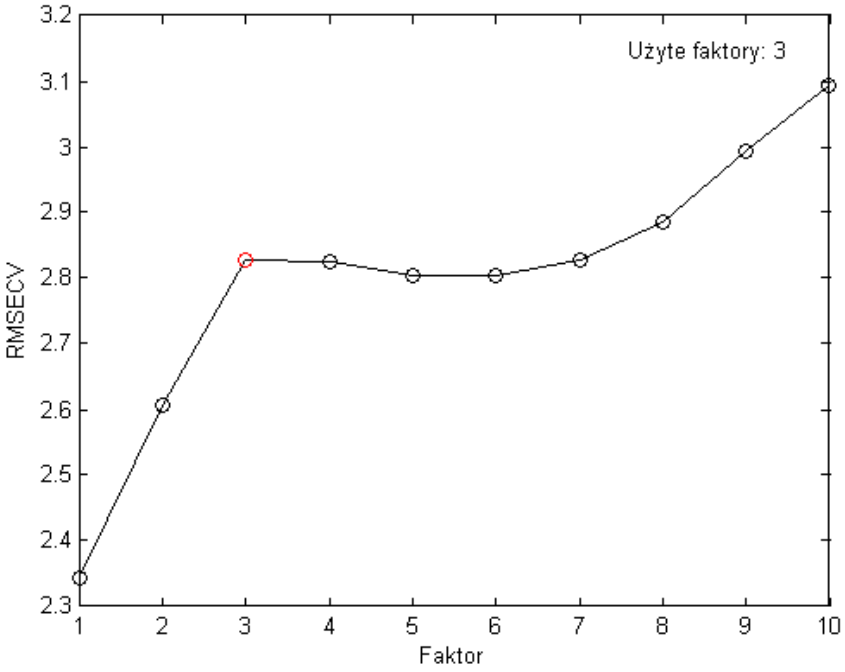
Tablica 5.14: Tabela parametrów dla modelu zawartości lizyny



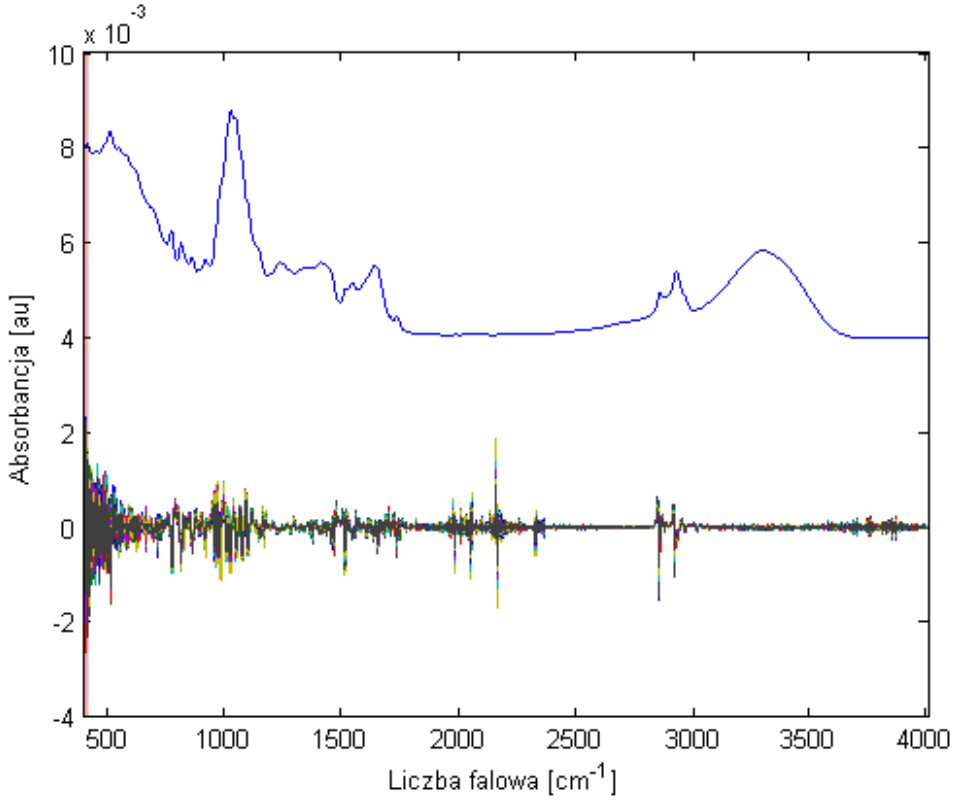
Rysunek 5.55: Krzywa predykcji oznaczeń zawartości lizyny w próbce



Rysunek 5.56: Względny błąd oznaczeń zawartości lizyny w próbce



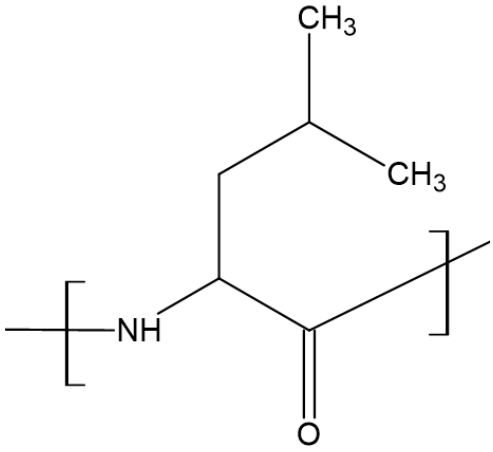
Rysunek 5.57: Wykres RMSCEV dla modelu zawartości lizyny w próbce



Rysunek 5.58: Wykres zakresów użytych do budowy modelu zawartości lizyny w próbce

5.13 Modelowanie zawartości leucyny

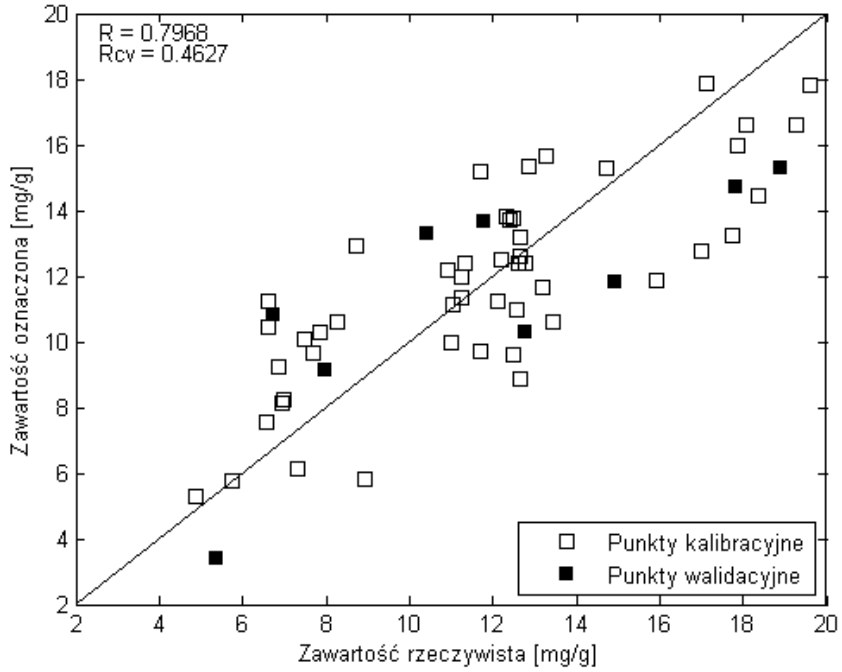
Opracowanie modelu kalibracyjnego dla leucyny (rysunek 5.59) wiązało się z problemami, które opisano wcześniej w dziale poświęconym izoleucynie. Silna korelacja stężeniowa i widmowa (tabela 3.1), wynikająca z niemal identycznej struktury, uniemożliwia skuteczną selekcję udziałów spektralnych i ich korelację ze stężeniem. Rezultatem jest model o niskich parametrach jakości (tabela 5.15, rysunek 5.60), gdzie algorytm ModelHelper nie był w stanie znaleźć użytecznych zakresów widma. Rozwiązanie tego problemu omówiono dalej w dziale ILE+LEU. Na rysunkach 5.62 i 5.63 przedstawiono przebieg PRESS i zakres użyty do stworzenia modelu dla leucyny.



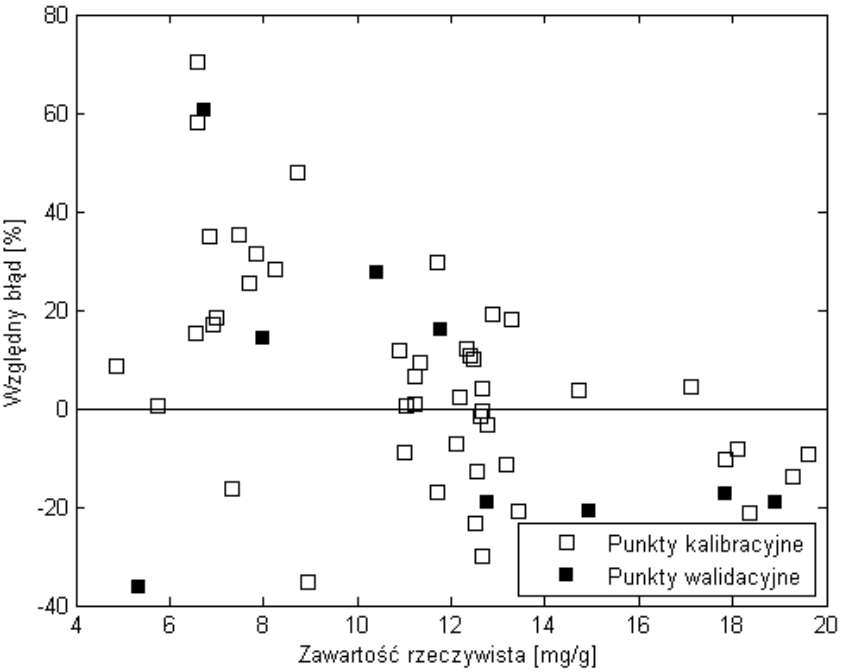
Rysunek 5.59: Struktura leucyny związanej wewnątrz białka

Leucyna	
Zakres stężeniowy	5-20mg/g
RMSEC	2.30
RMSEP	2.82
Użyte faktory:	3

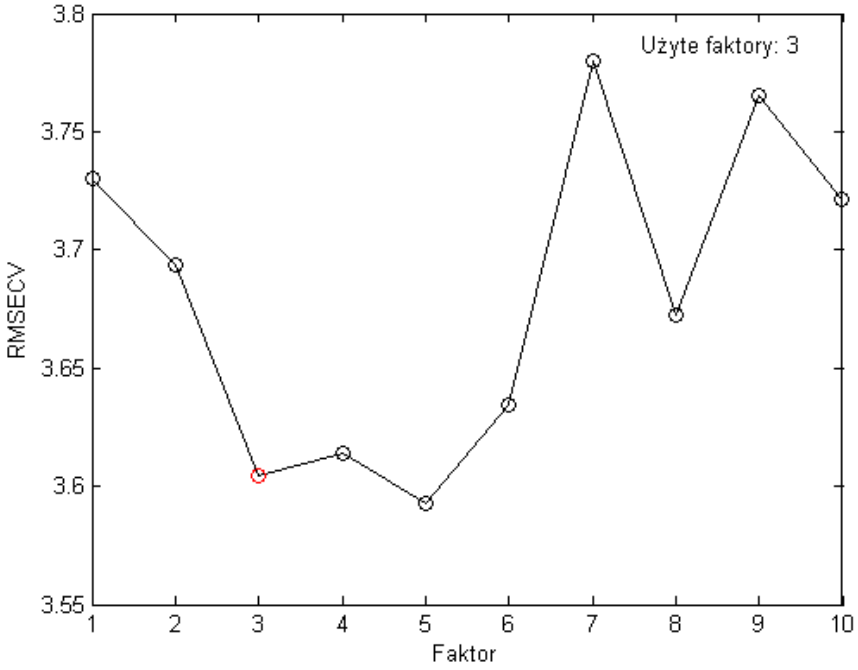
Tablica 5.15: Tabela parametrów dla modelu zawartości leucyny



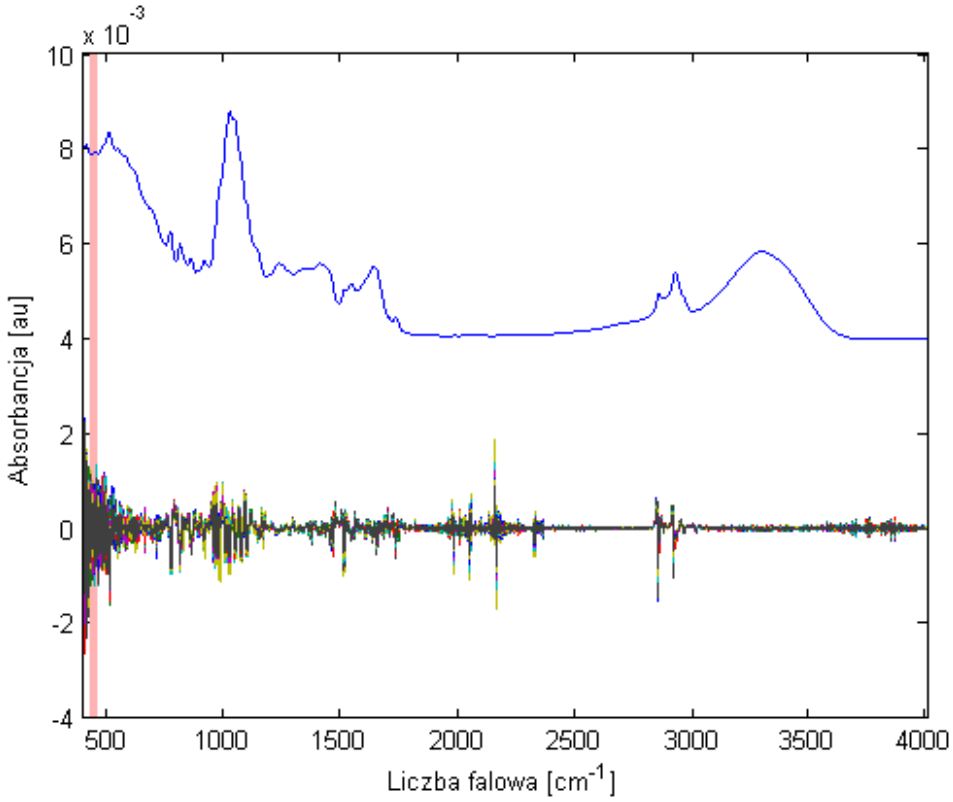
Rysunek 5.60: Krzywa predykcji oznaczeń zawartości leucyny w próbce



Rysunek 5.61: Względny błąd oznaczeń zawartości leucyny w próbce



Rysunek 5.62: Wykres RMSCEV dla modelu zawartości leucyny w próbce



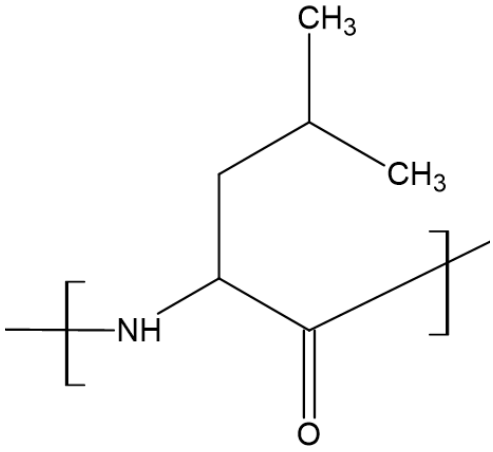
Rysunek 5.63: Wykres zakresów użytych do budowy modelu zawartości leucyny w próbce

5.14 Modelowanie sumy zawartości leucyny i izoleucyny

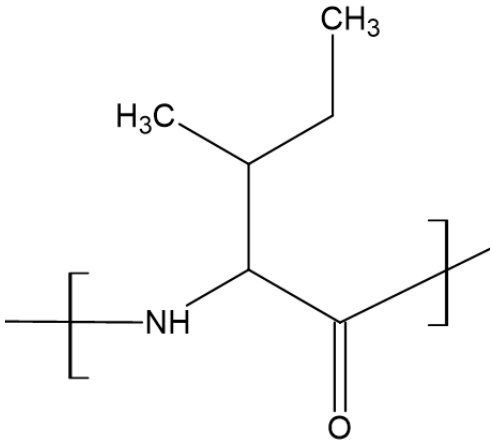
Rozwiązaniem problemu silnej korelacji leucyny i izoleucyny, przedstawionych na rysunkach 5.64, 5.65 okazało się stworzenie wspólnego modelu dla sumy ich stężeń, co pozwoliło na znaczną poprawę parametrów jakości (tabela 5.16) i stabilności modelu. Uważa się, że podobne podejście można by zastosować dla modelowania innych, silnie skorelowanych aminokwasów, przedstawionych w tabeli 3.1. W wyniku tej operacji poprawiono walidację krzyżową do wartości 0.78, ograniczono błędy względne do wartości ±20% (rysunek 5.21) i uzyskano poprawny przebieg PRESS (rysunek 5.68). Podejście to poprawiło zdolność algorytmu ModelHelper do znalezienia informatywnych zakresów spektralnych, zwiększając ich liczbę do siedmiu, co przedstawiono na rysunku 5.69.

Leucyna i Izoleucyna	
Zakres stężeniowy	7-33mg/g
RMSEC	0.179
RMSEP	0.129
Użyte faktory:	4

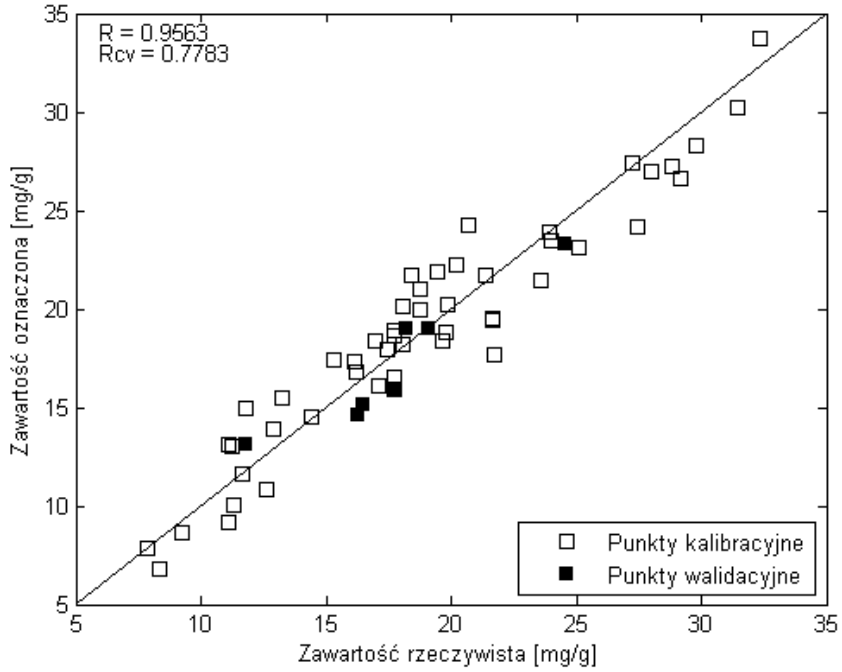
Tablica 5.16: Tabela parametrów dla modelu zawartości leucyny i izoleucyny



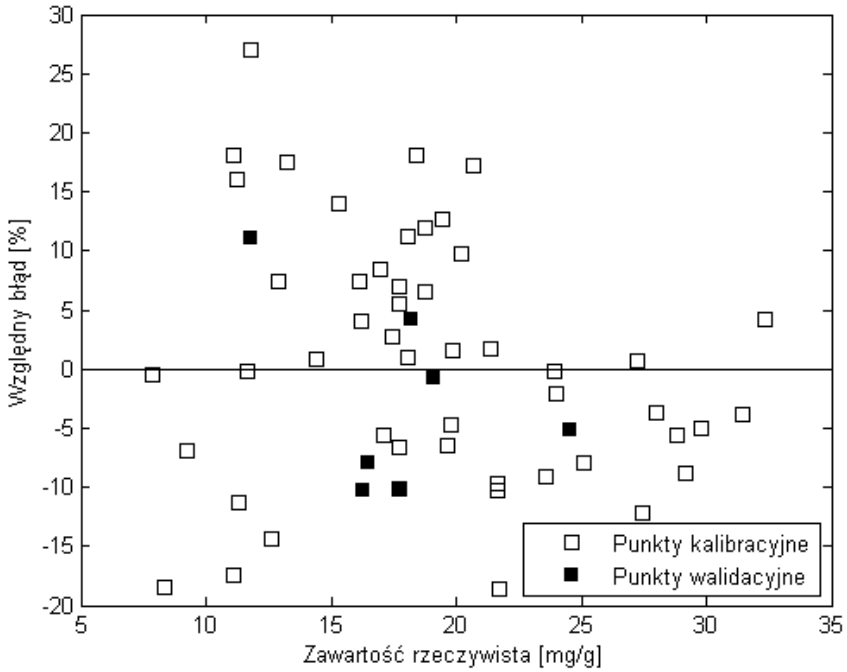
Rysunek 5.64: Struktura leucyny związanej wewnątrz białka



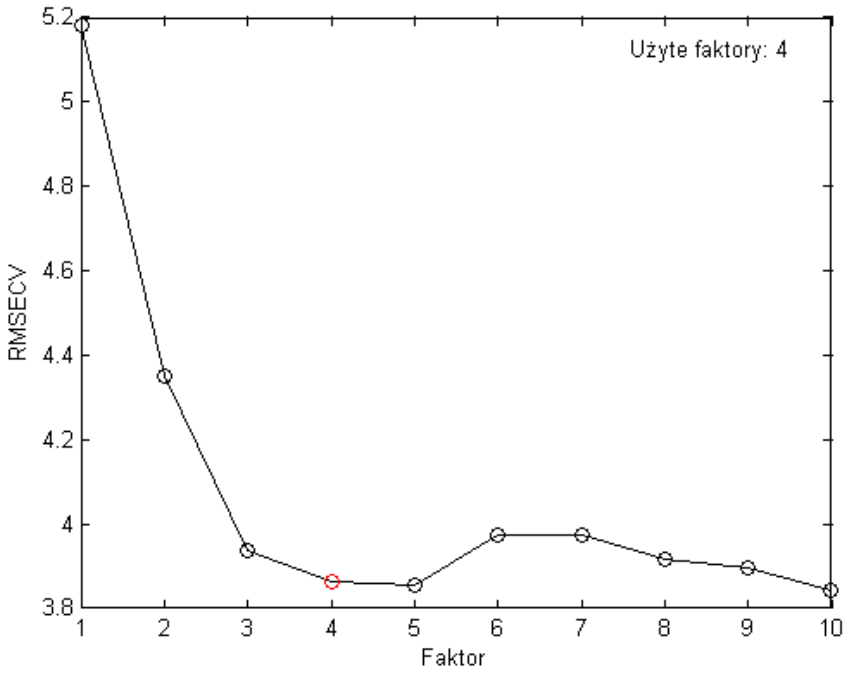
Rysunek 5.65: Struktura izoleucyny związanej wewnątrz białka



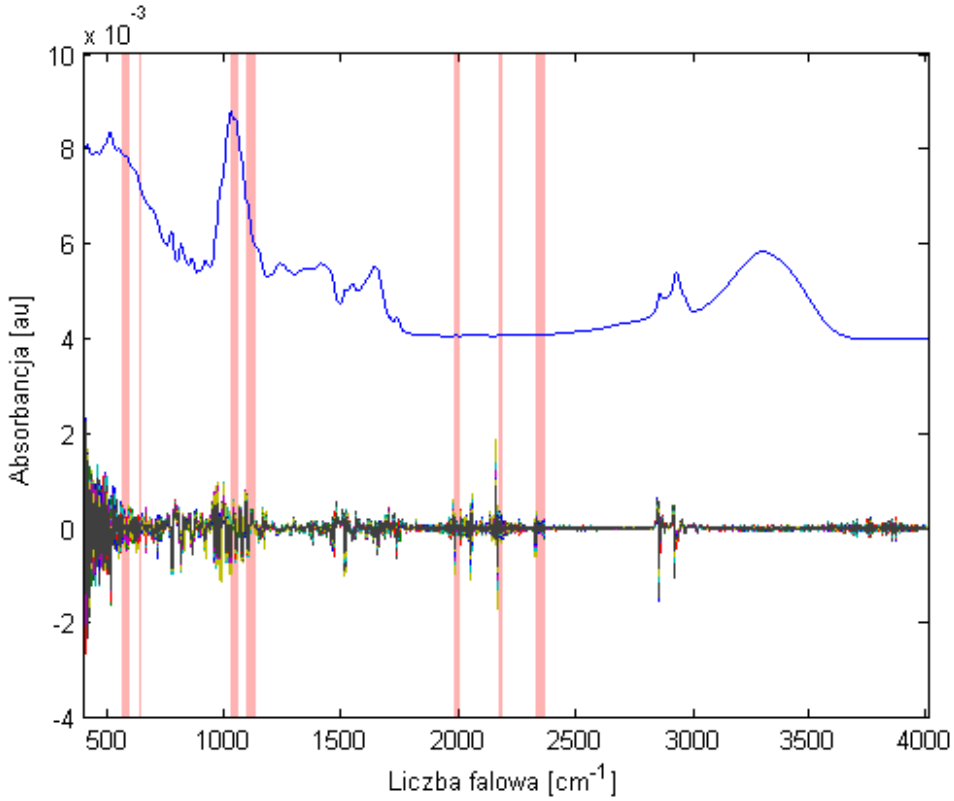
Rysunek 5.66: Krzywa predykcji oznaczeń zawartości sumy izoleucyny i leucyny w próbce



Rysunek 5.67: Względny błąd oznaczeń zawartości sumy izoleucyny i leucyny w próbce



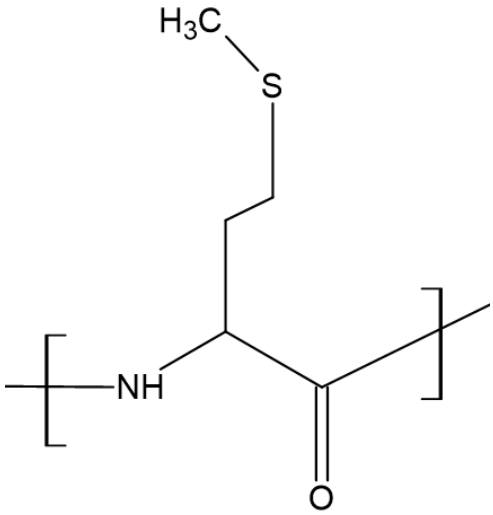
Rysunek 5.68: Wykres RMSCEV dla modelu zawartości sumy izoleucyny i leucyny w próbce



Rysunek 5.69: Wykres zakresów użytych do budowy modelu zawartości sumy izoleucyny i leucyny w próbce

5.15 Modelowanie zawartości metioniny

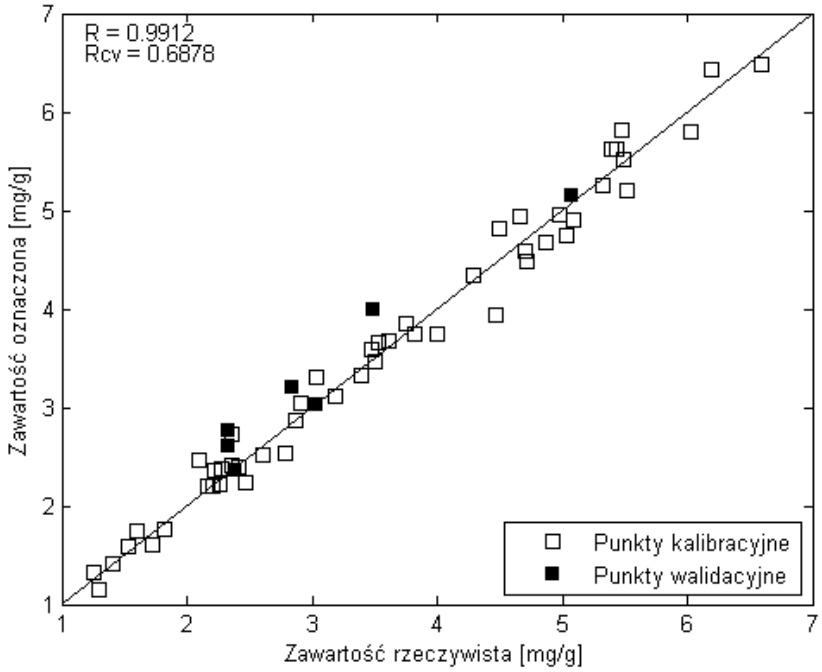
Model ilościowy wykonany dla metioniny (rysunek 5.70) charakteryzuje się przeciętnymi wartościami parametrów jakości (rysunek 5.71). Przyczyną jest jej niezwykle niskie stężenie w próbce, wynoszące poniżej 6mg/g (tabela 5.17). Może to skutkować trudnością w identyfikacji zmienności spektralnej pochodzącej od metioniny w widmie pyłku i nakładanie się jej z udziałami innych aminokwasów obecnych w próbkach. Dodatkowo, niskie stężenie może negatywnie wpływać na dokładność metody referencyjnej. Zależność błędów względnych od niskiego stężenia można zaobserwować na rysunku 5.72, gdzie błędy tworzą charakterystyczny 'stożek', skierowany w stronę wyższych stężeń. Jak widać na rysunku 5.73, przebieg PRESS jest poprawny. Biorąc pod uwagę niezwykle niską zawartość metioniny w analizowanym układzie otrzymane wyniki są satysfakcjonujące. Zakresy zastosowane przy tworzeniu modelu można zaobserwować na rysunku 5.74.



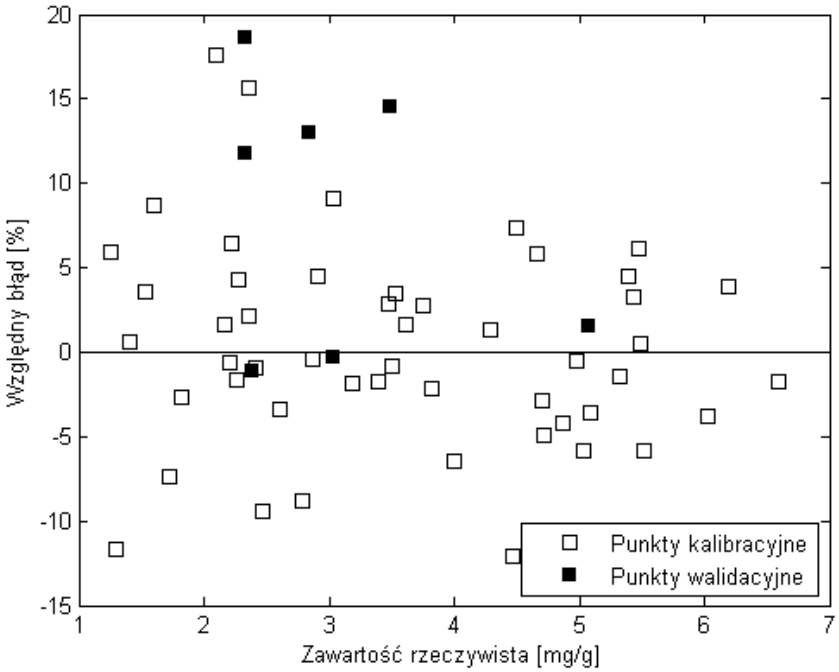
Rysunek 5.70: Struktura metioniny związanej wewnątrz białka

Metionina	
Zakres stężeniowy	1-7mg/g
RMSEC	0.193
RMSEP	0.308
Użyte faktory:	8

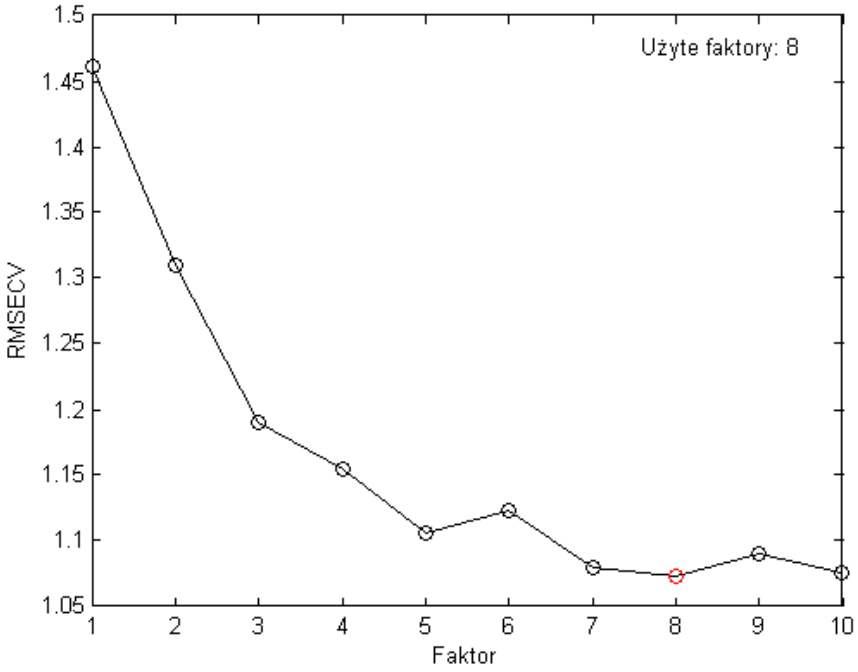
Tablica 5.17: Tabela parametrów dla modelu zawartości metioniny



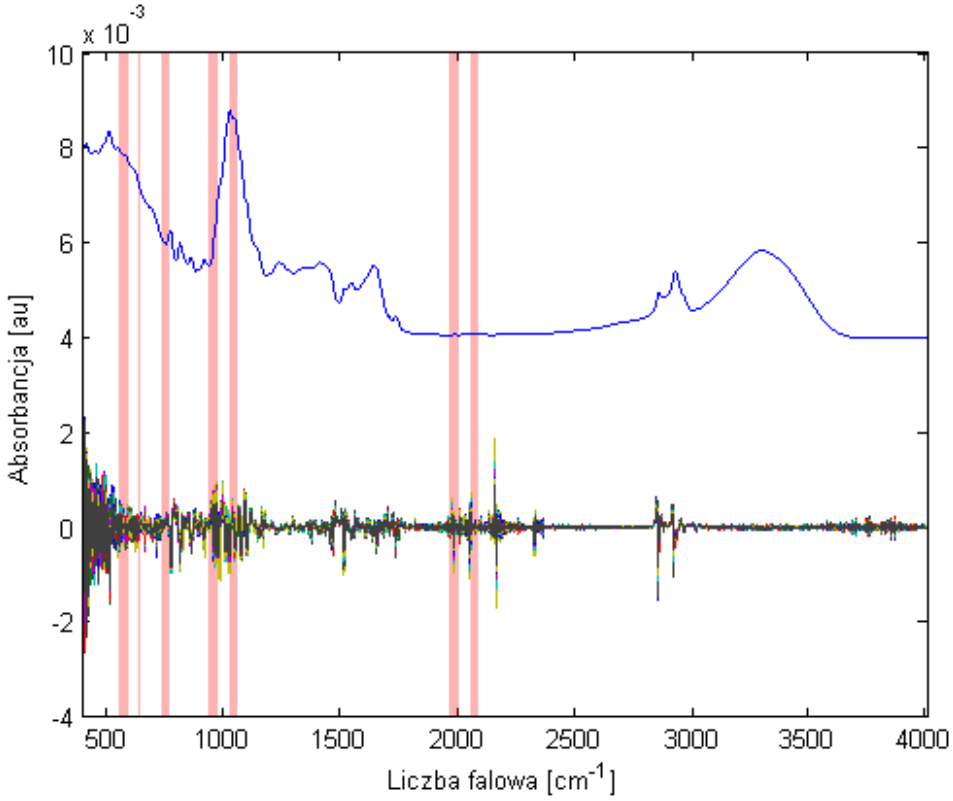
Rysunek 5.71: Krzywa predykcji oznaczeń zawartości metioniny w próbce



Rysunek 5.72: Względny błąd oznaczeń zawartości metioniny w próbce



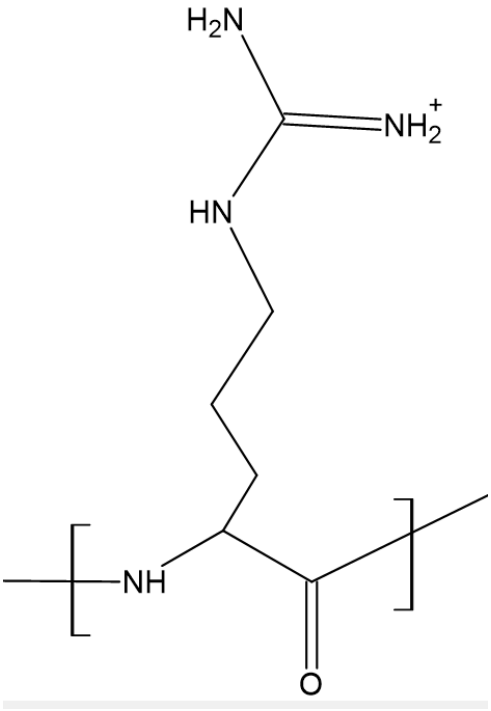
Rysunek 5.73: Wykres RMSCEV dla modelu zawartości metioniny w próbce



Rysunek 5.74: Wykres zakresów użytych do budowy modelu zawartości metioniny w próbce

5.16 Modelowanie zawartości argininy

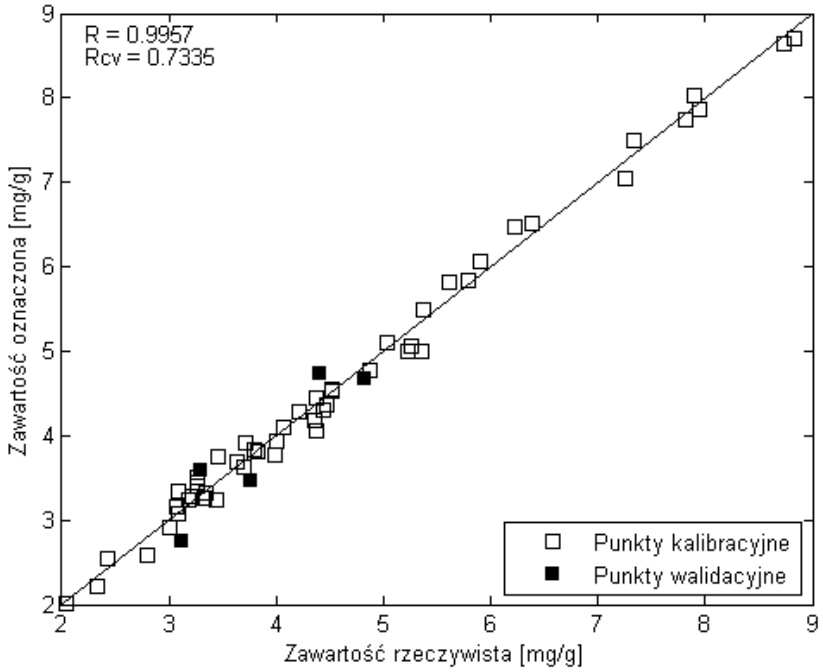
Model predykcyjny opracowany dla argininy (rysunek 5.75), w zakresie stężeniowym podobnym do poprzednio analizowanej metioniny, charakteryzuje się dobrymi parametrami jakości i predykcji (rysunek 5.76, tabela 5.18). Na rysunku 5.77 ponownie widać, że rozkład błędów względnych oznaczeń poszczególnych próbek formuje charakterystyczny stożek, którego formacja jest rezultatem wyższych błędów predykcji w niskim zakresie stężeń. Warto zauważyć, że obecność bardziej charakterystycznego łańcucha bocznego może być źródłem dodatkowej zmienności spektralnej, która byłaby odpowiedzialna za lepsze parametry otrzymanego modelu. Przebieg PRESS (rysunek 5.78) przyjmuje klasyczny układ osypiska. Do budowy modelu użyto siedmiu zakresów, które przedstawiono na rysunku 5.79.



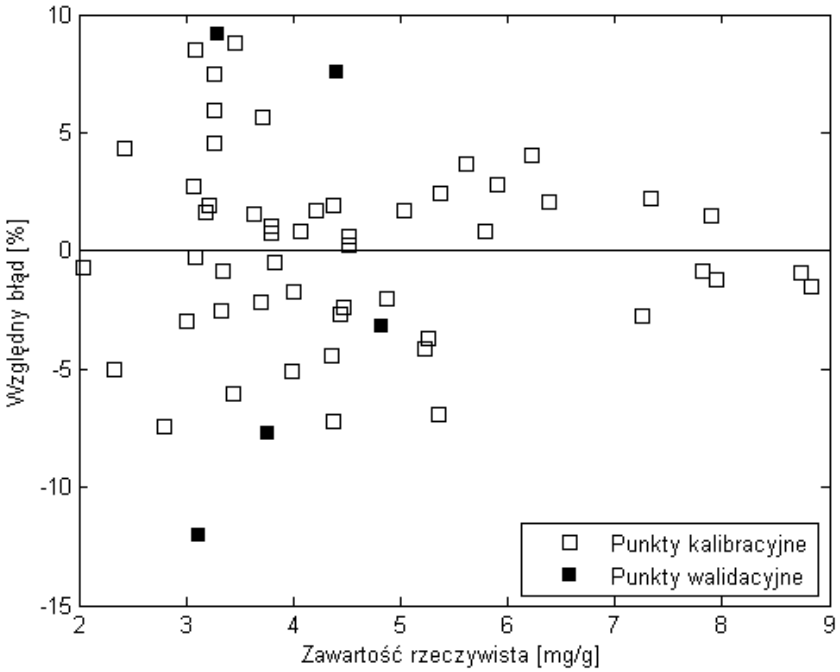
Rysunek 5.75: Struktura argininy związanej wewnątrz białka

Arginina	
Zakres stężeniowy	2-9mg/g
RMSEC	0.158
RMSEP	0.293
Użyte faktory:	8

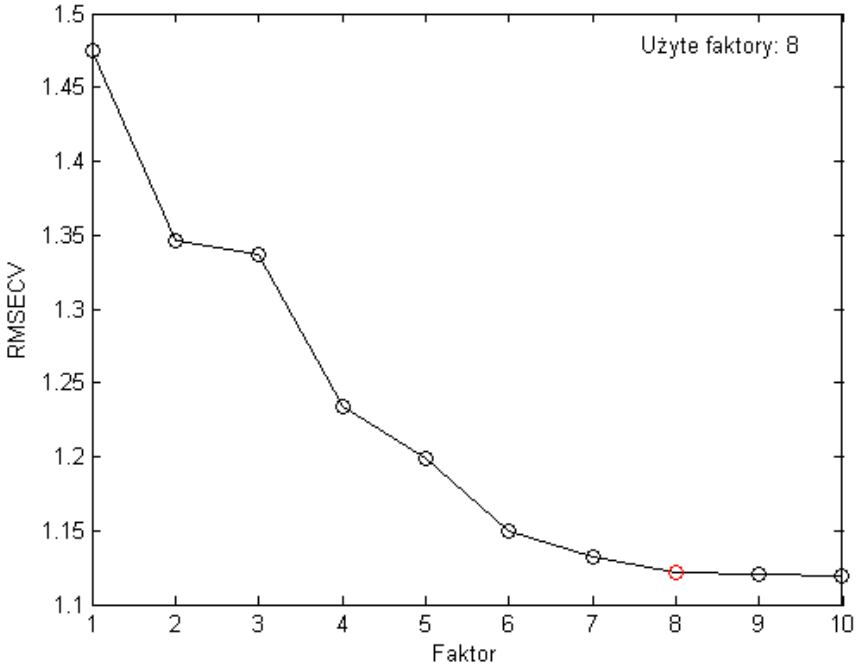
Tablica 5.18: Tabela parametrów dla modelu zawartości argininy



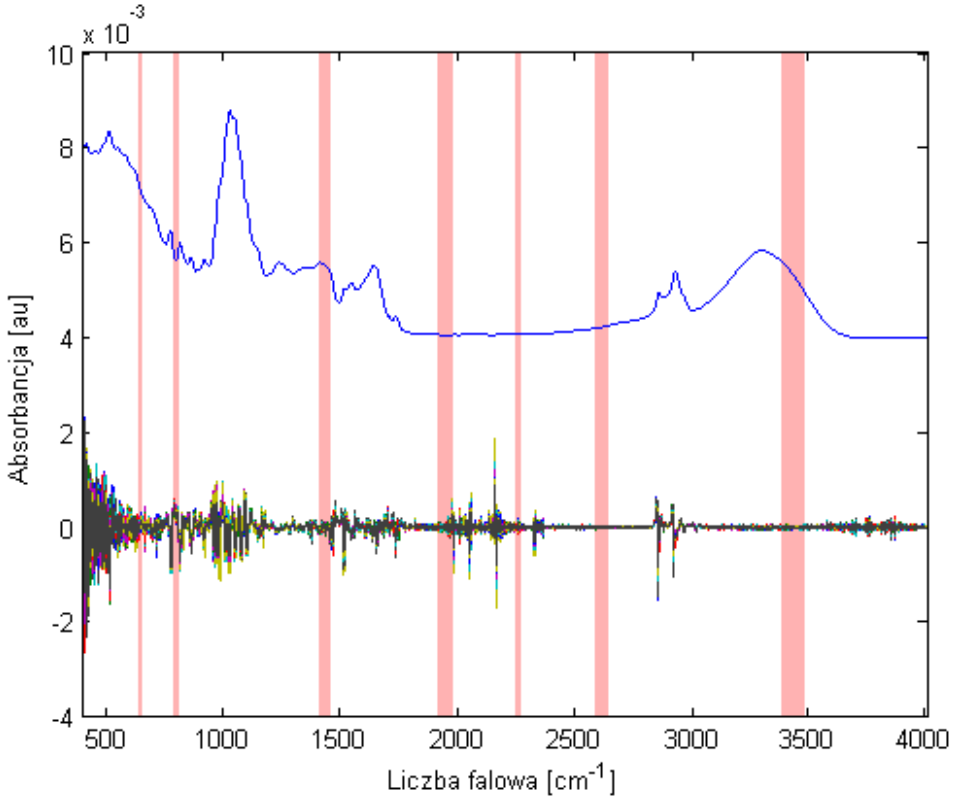
Rysunek 5.76: Krzywa predykcji oznaczeń zawartości argininy w próbce



Rysunek 5.77: Względny błąd oznaczeń zawartości argininy w próbce



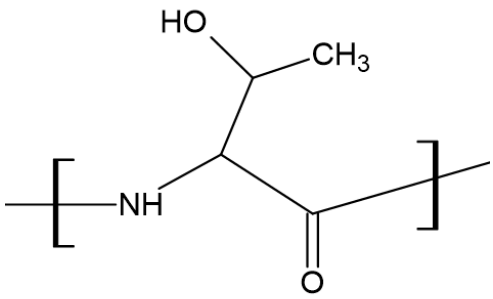
Rysunek 5.78: Wykres RMSCEV dla modelu zawartości argininy w próbce



Rysunek 5.79: Wykres zakresów użytych do budowy modelu zawartości argininy w próbce

5.17 Modelowanie zawartości treoniny

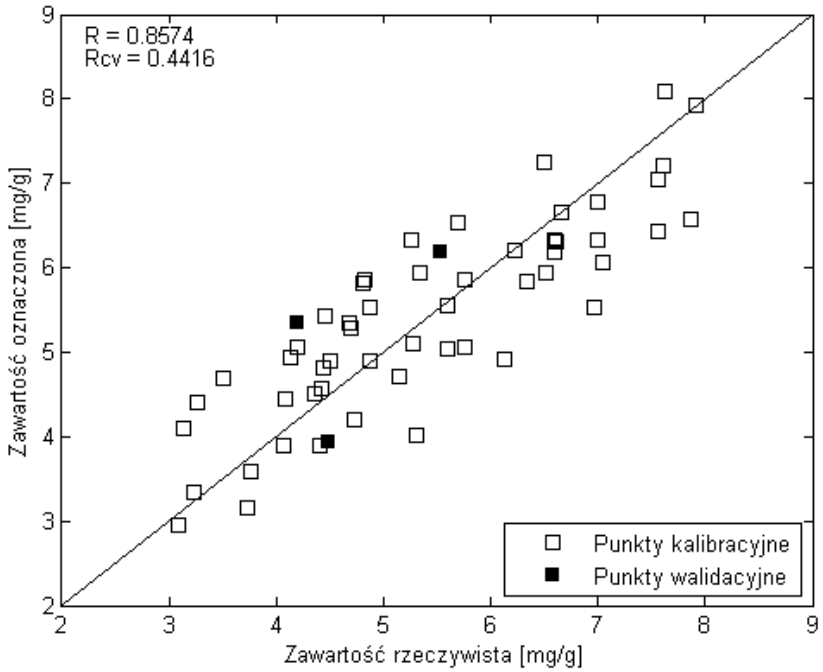
Jak przedstawiono na rysunku 5.80 model dla treoniny charakteryzuje się niską stabilnością i słabą zdolnością predykcji. Najbardziej prawdopodobną przyczyną jest niska charakterystyczność strukturalna jej łańcucha bocznego i wynikająca z niej niewielka zmienność spektralna (rysunek 5.82, tabela 5.19). Ten problem jest też prawdopodobną przyczyną nietypowego przebiegu PRESS, przedstawionego na rysunku 5.83. Użycie trzech faktorów generowało lepsze wartości parametrów jakości opracowanego modelu. Do jego budowy użyto ośmiu zakresów, które zaprezentowano na rysunku 5.84.



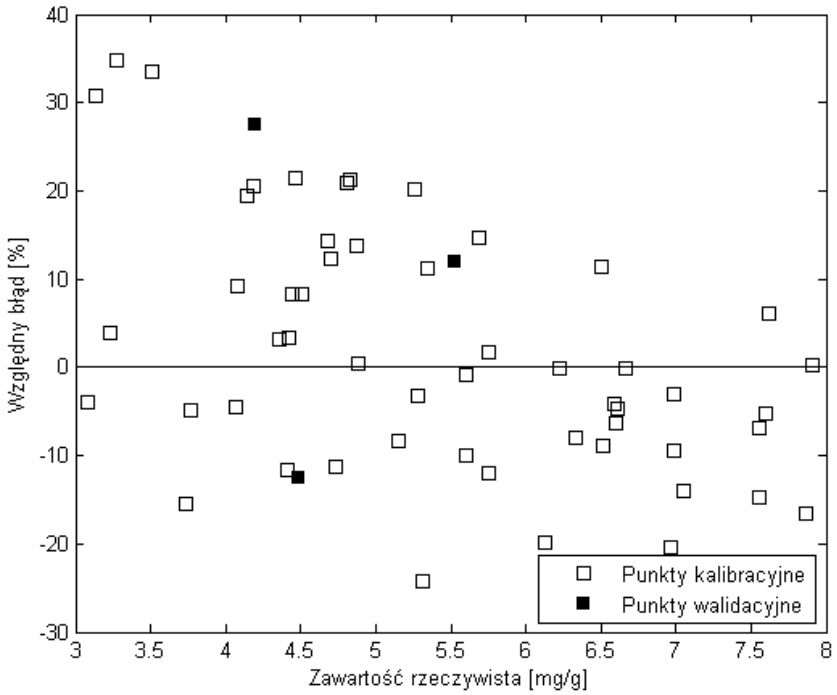
Rysunek 5.80: Struktura treoniny związanej wewnątrz białka

Treonina	
Zakres stężeniowy	2-9mg/g
RMSEC	0.694
RMSEP	0.834
Użyte faktory:	3

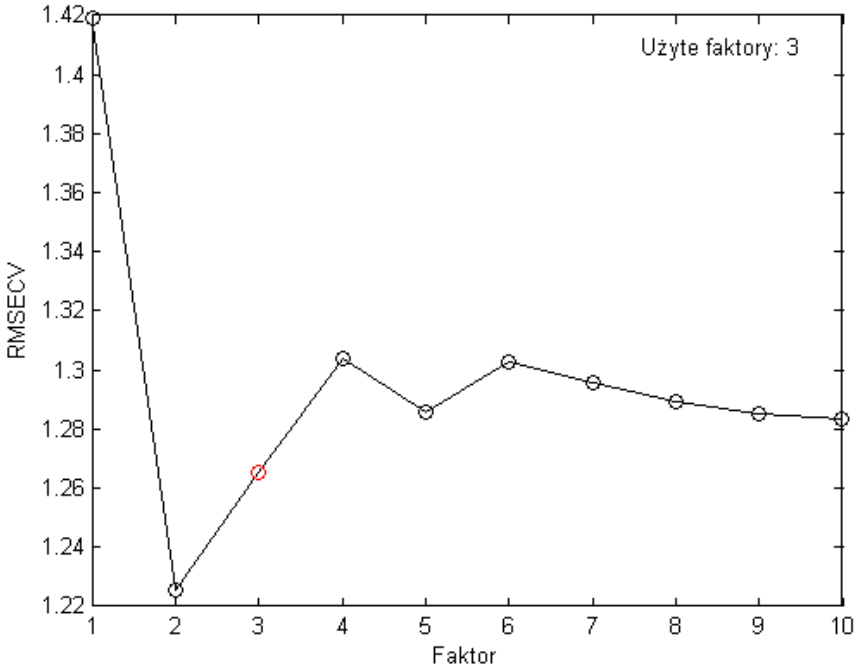
Tablica 5.19: Tabela parametrów dla modelu zawartości treoniny



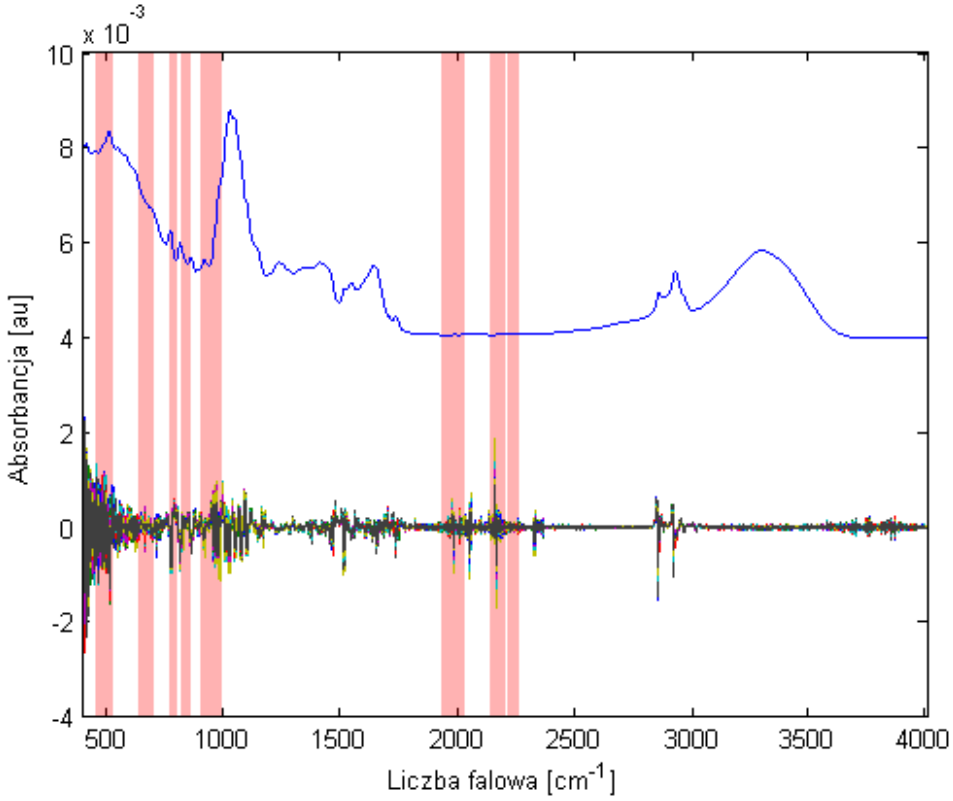
Rysunek 5.81: Krzywa predykcji oznaczeń zawartości treoniny w próbce



Rysunek 5.82: Względny błąd oznaczeń zawartości treoniny w próbce



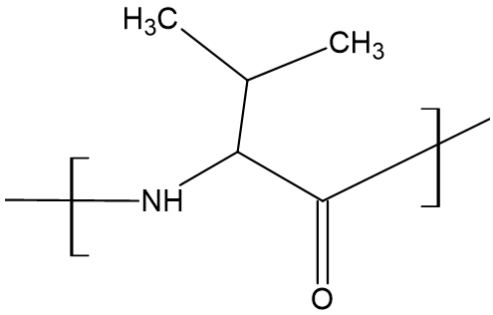
Rysunek 5.83: Wykres RMSCEV dla modelu zawartości treoniny w próbce



Rysunek 5.84: Wykres zakresów użytych do budowy modelu zawartości treoniny w próbce

5.18 Modelowanie zawartości waliny

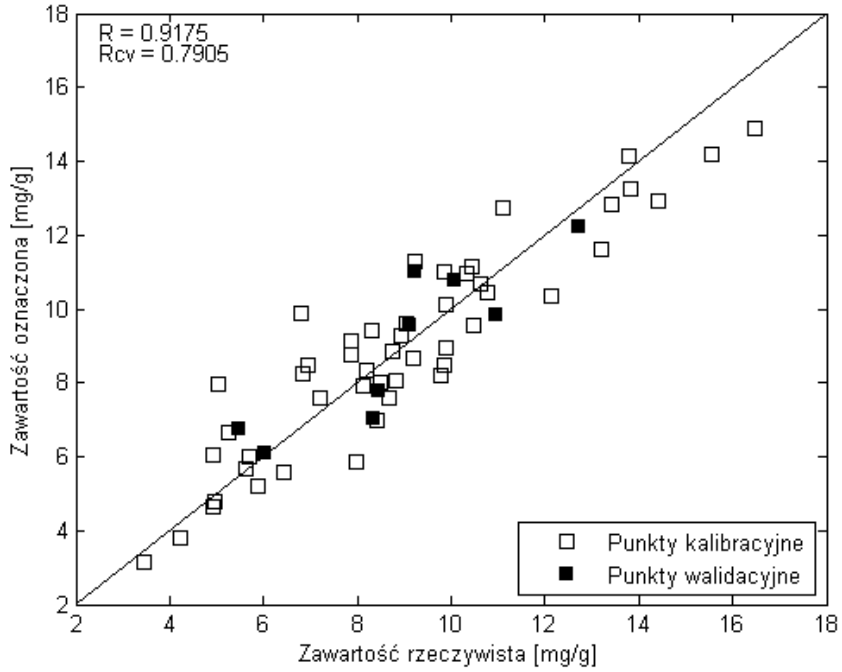
Na rysunku 5.85 przedstawiono krzywą regresji i względne błędy oznaczeń modelu PLS opracowanego dla waliny. Charakteryzuje się on bardzo dobrymi parametrami R i R_{CV} (rysunek 5.86), problem stanowią natomiast słaby przebieg PRESS (rysunek 5.88) oraz obecność kilku próbek wyraźnie odstających w opracowanym modelu. Najbardziej prawdopodobną przyczyną tych zjawisk są duże różnice w zgodności oznaczeń pomiędzy metodą referencyjną a widmową dla próbek o stężeniach poniżej 8mg/g, co pokrywałoby się z obserwacjami błędów dla poprzednich próbek. Do budowy modelu użyto pięciu zakresów widmowych, przedstawionych na rysunku 5.89. Parametry modelu umieszczono w tabeli 5.20.



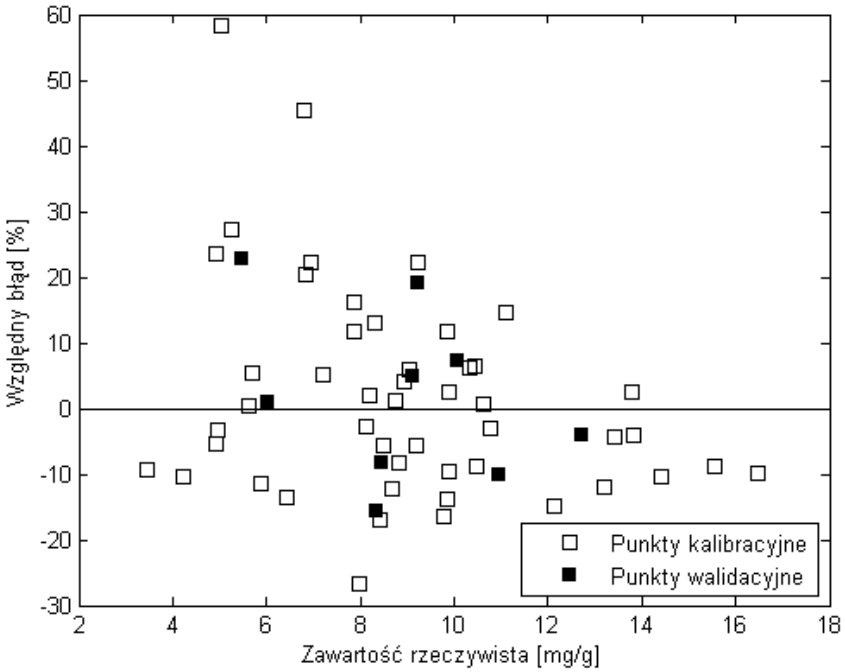
Rysunek 5.85: Struktura waliny związanej wewnątrz białka

Walina	
Zakres stężeniowy	4 - 17mg/g
RMSEC	1.190
RMSEP	1.010
Użyte faktory:	3

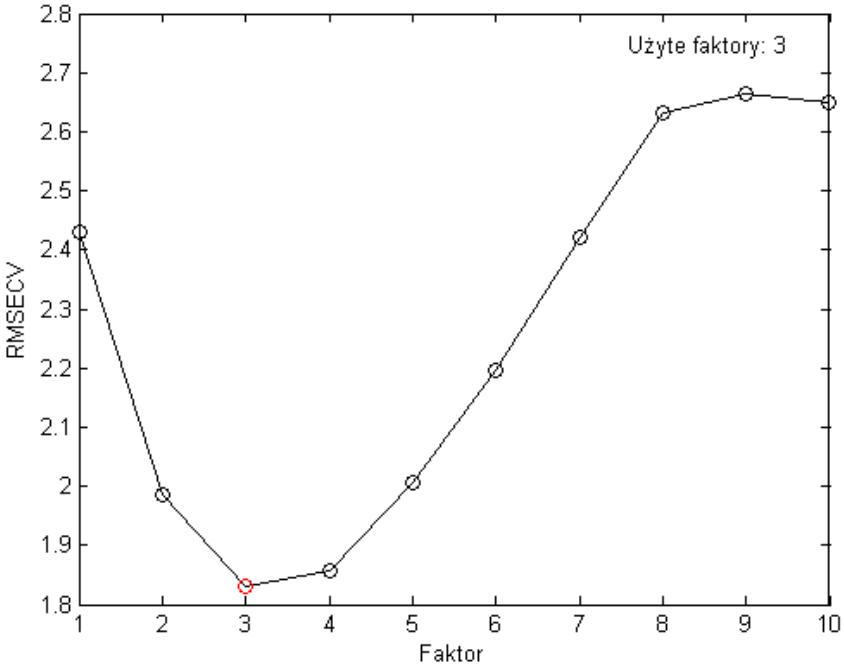
Tablica 5.20: Tabela parametrów dla modelu zawartości waliny



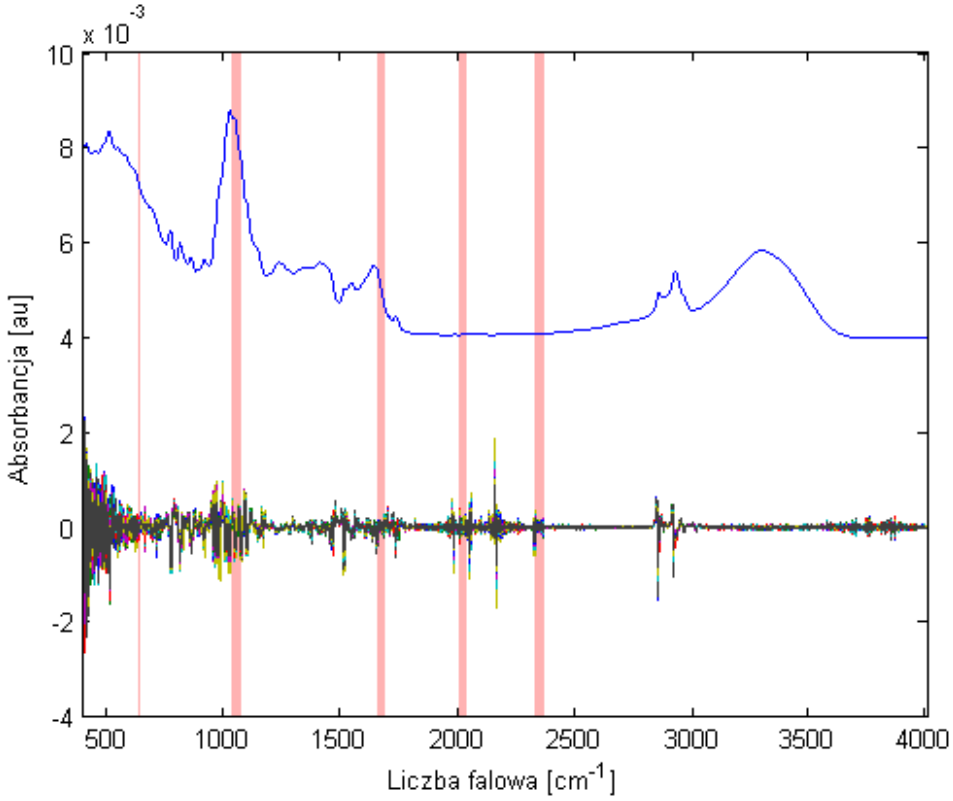
Rysunek 5.86: Krzywa predykcji oznaczeń zawartości waliny w próbce



Rysunek 5.87: Względny błąd oznaczeń zawartości waliny w próbce



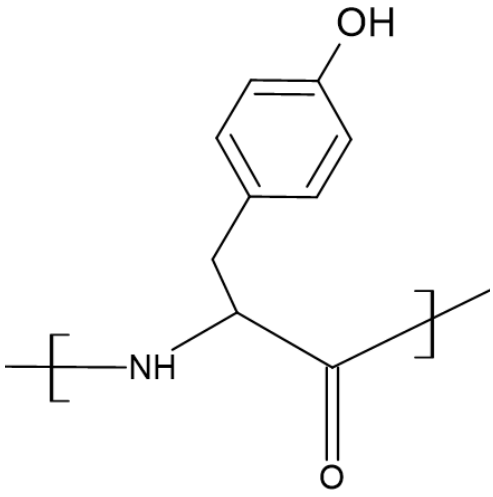
Rysunek 5.88: Wykres RMSCEV dla modelu zawartości waliny w próbce



Rysunek 5.89: Wykres zakresów użytych do budowy modelu zawartości waliny w próbce

5.19 Modelowanie zawartości tyrozyny

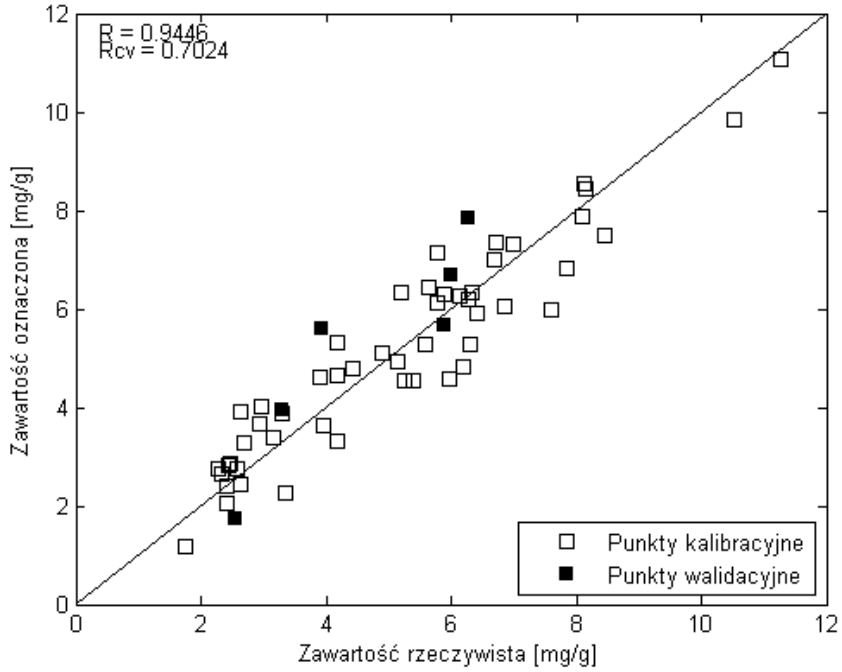
Model predykcyjny opracowany dla tyrozyny (rysunek 5.90) przy użyciu zakresów spektralnych wybranych przez ModelHelper charakteryzuje się dobrymi parametrami przewidywania (rysunek 5.91), lecz przeciętnymi wynikami dla walidacji krzyżowej, co widać na przebiegu PRESS przedstawionym na rysunku 5.93, pomimo obecności charakterystycznej grupy fenolowej. Najbardziej prawdopodobną przyczyną jest wysoka zawartość związków polifenolowych w badanych próbkach pyłku pszczelego, których udziały spektralne maskują zmienność spektralną pochodzącą od tyrozyny. Błędy względne oznaczeń przedstawiono na rysunku 5.92. Do budowy modelu użyto siedmiu zakresów widmowych przedstawionych na rysunku 5.94. Parametry modelu umieszczono w tabeli 5.21.



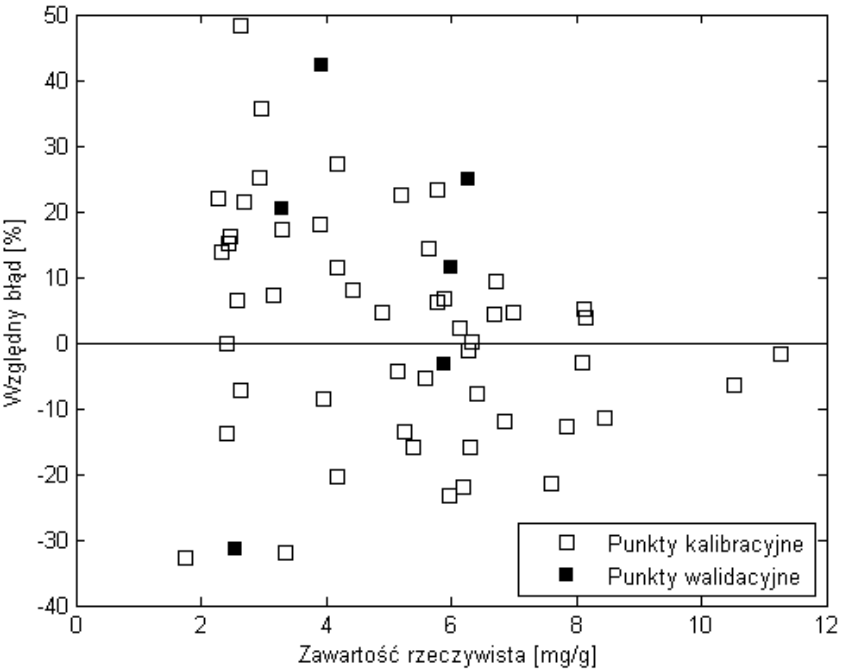
Rysunek 5.90: Struktura tyrozyny związanej wewnątrz białka

Tyrozyna	
Zakres stężeniowy	2-12mg/g
RMSEC	0.723
RMSEP	1.060
Użyte faktory:	3

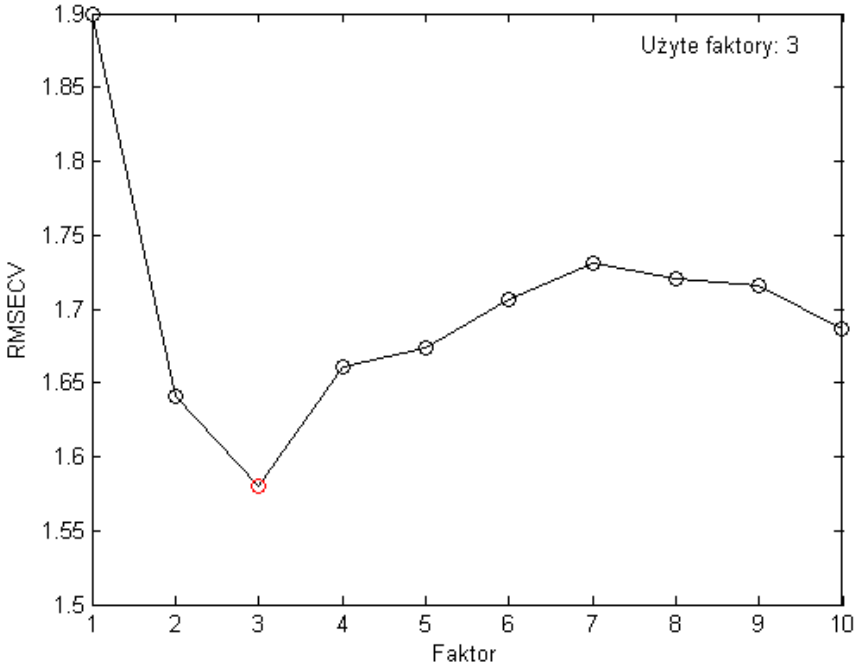
Tablica 5.21: Tabela parametrów dla modelu zawartości tyrozyny



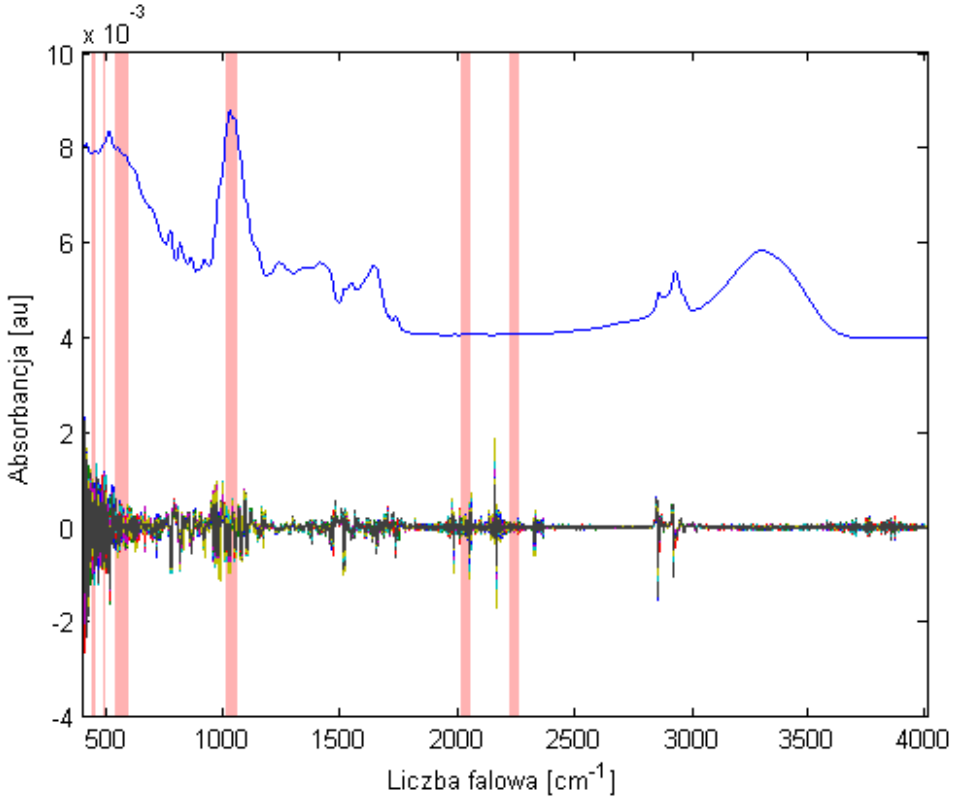
Rysunek 5.91: Krzywa predykcji oznaczeń zawartości tyrozyny w próbce



Rysunek 5.92: Względny błąd oznaczeń zawartości tyrozyny w próbce



Rysunek 5.93: Wykres RMSCEV dla modelu zawartości tyrozyny w próbce



Rysunek 5.94: Wykres zakresów użytych do budowy modelu zawartości tyrozyny w próbce

Rozdział 6

Podsumowanie i wnioski

W toku pracy opracowano nowy algorytm selekcji zakresów spektralnych w widmach IR próbek pyłków pszczelich, który wykorzystano w modelowaniu profilu aminokwasowego. Udowodniono jego przewagę nad algorytmem iPLS implementowanym w PLS Toolbox, poprzez otrzymanie modeli ilościowych o wyraźnie wyższej zdolności zdolności prognostycznej w znacząco krótszym czasie optymalizacji zakresów. Zautomatyzowano w znacznym stopniu proces budowy modeli, umożliwiając łatwe przeprowadzenie analiz dla dużej grupy składników.

Stosując zaproponowany algorytm dla analizowanego zestawu danych opracowano 19 modeli chemometrycznych, z których aż 12 charakteryzowało się dobrymi parametrami predykcji i jakości (R_{CV} w zakresie od 0.7 do 0.89), a kolejne cztery miały potencjał do poprawy przy zwiększeniu populacji próbek i wykonaniu precyzyjniejszych pomiarów referencyjnych (R_{CV} w zakresie od 0.61 do 0.69). Dla trzech z opracowanych modeli nie udało się otrzymać zadowalających parametrów jakości. Jest to zbieżne z wynikami innych badań w tym zakresie [36,37] i najprawdopodobniej jest konsekwencją bardzo niskich stężeń tych aminokwasów w badanych próbkach oraz ich potencjalnymi interakcjami z innymi związkami zawartymi w pyłkach.

Jest to pierwsza tego typu praca, w której przeprowadzono analizę profilu aminokwasowego na podstawie widm ATR dla więcej niż kilku aminokwasów jednocześnie. Istnieje także szereg możliwości poprawy parametrów otrzymanych modeli, między innymi sprzężenie jej z pomiarami innych technik spektroskopowych, zwiększenie liczby próbek, poprawę rozdzielczości rejestrowanych widm czy oraz wyższą dokładność pomiarów referencyjnych.

Otrzymane parametry jakościowe opracowanych modeli kalibracyjnych zostały zebrane w tabeli 6.1. Z prezentowanego podsumowania wynika, że najwyższą zdolnością prognostyczną charakteryzowały się modele dla białka ogółem, sumy aminokwasów, alaniny i fenyloalaniny, dla których wartość parametru R_{CV} przekraczała 0.8.

Parametry modeli PLS dla oznaczeń zawartości białka i aminokwasów					
Nazwa substancji	R	R_{CV}	RMSEC	RMSEP	Liczba faktorów
Białko	0.966	0.892	0.78	0.62	3
Aminokwasy egzogenne	0.980	0.699	2.59	2.78	5
Suma aminokwasów	0.995	0.886	2.09	3.44	6
Alanina	0.987	0.880	0.61	0.47	4
Cysteina	0.974	0.614	0.14	0.14	9
Kwas Asparaginowy	0.925	0.712	0.74	0.76	3
Kwas Glutaminowy	0.941	0.583	0.74	0.63	4
Fenyloalanina	0.993	0.823	0.36	0.63	7
Glicyna	0.968	0.772	1.14	2.48	3
Histydyna	0.975	0.776	0.47	0.66	4
Izoleucyna	0.943	0.643	0.84	1.03	7
Lizyna	0.287	0.709	1.93	1.74	3
Leucyna	0.797	0.463	2.30	2.82	3
LEU + ILE	0.956	0.778	0.18	0.13	4
Metionina	0.991	0.688	0.19	0.31	8
Arginina	0.996	0.734	0.15	0.30	7
Treonina	0.857	0.442	0.69	0.83	3
Walina	0.918	0.790	1.19	1.01	3
Tyrozyna	0.945	0.702	0.72	1.06	3

Tablica 6.1: Zestawienie parametrów opracowanych modeli PLS dla analizowanych aminokwasów

Głównymi problemami w zastosowaniu modelowania PLS z użyciem danych spektroskopowych w analizie wieloskładnikowych układów złożonych są trudności w oznaczaniu substancji silnie skorelowanych stężeniowo oraz tych o niskiej zawartości w próbce. Możliwym sposobem korekty tych ograniczeń jest współmodelowanie skorelowanych związków

oraz dołączenie dodatkowych macierzy informatywnych pochodzących z innych metod analitycznych i uwypuklających różnice między badanymi związkami.

Otrzymane wyniki podkreślają także znaczenie doboru zakresów spektralnych w optymalizacji modelów ilościowych, a także potencjał techniki ATR w oznaczaniu składników próbek naturalnych. Zaproponowana metoda może stać się alternatywą dla standardowych metod analiz używanych w badaniach profilu aminokwasowego opartych na technikach ekstrakcji. Brak konieczności uprzedniego przygotowania próbek przed pomiarami ich widm sprawia, że metoda ta jest szybsza, bardziej ekologiczna, tańsza i podatna na automatyzację.

Rozdział 7

Streszczenie

We wstępie przybliżono zagadnienia związane z tematyką pracy, podano krótką charakterystykę analizowanych pyłków pszczelich, opisano proces budowy typowego modelu chemometrycznego oraz scharakteryzowano klasyczne metody selekcji danych. W części doświadczalnej przedstawiono proces tworzenia nowego algorytmu heurystycznego oraz użyte w pracy oprogramowanie. W kolejnym rozdziale porównano wyniki uzyskane przy użyciu nowej metody z tymi, które dały klasyczne metody selekcji danych. Otrzymane parametry modeli potwierdziły lepszą skuteczność opracowanego algorytmu w doborze zakresów widmowych. Następnie przedstawiono dziewiętnaście modeli PLS pozwalających przewidzieć zawartość aminokwasów w próbkach pyłku pszczelego. Dla dwunastu z nich współczynniki walidacji krzyżowej znajdowały się w zakresie od 0.7 do 0.89, a błędy względne wyniosły od kilku do kilkudziesięciu procent. W podsumowaniu stwierdzono, że uzyskane wyniki potwierdzają skuteczność zaproponowanej metody selekcji danych oraz potencjał technik spektroskopii oscylacyjnej, w tym ATR, w analizie ilościowej złożonego produktu naturalnego, jakim jest pyłek pszczeli.

Abstract

First part of the thesis presents a general overview of bee pollen, chemometric model creation and data selection methods used in the development of quantitative models. Experimental section covers the process of designing a new data selection algorithm and software used in the course of the studies. In the following chapter, a comparison of the new method in relation to the classic approach is presented. Obtained model parameters confirm the superiority of the new approach, which is then successively used to create nineteen predictive models allowing for determination of the amino acid profile. For the twelve of aforementioned models, crossvalidation parameters ranging from 0.7 to 0.89 were obtained, with relative errors ranging from few to several dozen percent. In the summary of the thesis the effectiveness of the proposed data selection method and the potential of vibrational spectroscopy in analysis of complex natural products, such as bee pollen was confirmed.

Bibliografia

[1] M. Tasumi. *Introduction to Experimental Infrared Spectroscopy: Fundamentals and Practical Methods*. Wiley, 2014.

[2] Zou Xiaobo, Zhao Jiewen, Malcolm J.W. Povey, Mel Holmes, and Mao Hanpin. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, 667(1):14–32, 2010.

[3] Mireia Farrés, Stefan Platikanov, Stefan Tsakovski, and Romà Tauler. Comparison of the variable importance in projection (vip) and of the selectivity ratio (sr) methods for variable selection and interpretation. *Journal of Chemometrics*, 29(10):528–536, 2015.

[4] Magdalena Węglińska, Roman Szostak, Agnieszka Kita, Agnieszka Nemś, and Sylwester Mazurek. Determination of nutritional parameters of bee pollen by raman and infrared spectroscopy. *Talanta*, 212:120790, 2020.

[5] Gilliam, Martha. Microbiology of pollen and bee bread : The yeasts. *Apidologie*, 10(1):43–53, 1979.

[6] Katarzyna Komosinska-Vassev, Pawel Olczyk, Justyna Kaźmierczak, Lukasz Mencner, and Krystyna Olczyk. Bee pollen: Chemical composition and therapeutic application. *Evidence-Based Complementary and Alternative Medicine*, 2015:1–6, 2015.

[7] Bilicki Antoni. Zastosowanie techniki ssnmr i metod chemometrycznych w analizie ilościowej wybranych parametrów pyłku pszczelego. 2020.

[8] T'ai Roulston, Jim Cane, and Stephen Buchmann. What governs protein content of pollen: Pollinator preferences, pollen–pistil interactions, or phylogeny? *Ecological Monographs*, 70:617–643, 2000.

[9] Kai Yang, Dan Wu, Xingqian Ye, Donghong Liu, Jianchu Chen, and Peilong Sun. Characterization of chemical composition of bee pollen in china. *Journal of agricultural and food chemistry*, 61, 2012.

[10] J. Mazerski. *Chemometria praktyczna.* Malamut, 2009.

[11] Willem Windig, Jeremy Shaver, and Rasmus Bro. Loopy msc: A simple way to improve multiplicative scatter correction. *Applied spectroscopy*, 62:1153–9, 2008.

[12] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[13] Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, and Huiwen Wang. *Editorial: Perspectives on Partial Least Squares.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[14] Marzban E. Sanchez, G. *All Models Are Wrong: Concepts of Statistical Learning.* 2020.

[15] Strona firmy eigenvector zawierająca funkcje pls toolbox. `wiki.eigenvector.com`.

[16] Marfran Santos, Camilo Morais, and Kassio Lima. Atr-ftir spectroscopy for virus identification: A powerful alternative. pages 1–16, 2020.

[17] Oliver Wahl and Ulrike Holzgrabe. Amino acid analysis for pharmacopoeial purposes. *Talanta*, 154:150–163, 2016.

[18] Weihua Xu, Congcong Zhong, Chunpu Zou, Bing Wang, and Ning Zhang. Analytical methods for amino acid determination in organisms. *Amino Acids*, 52(8):1071–1088, 2020.

[19] Tian Lu and Susan V. Olesik. Electrospun polyvinyl alcohol ultra-thin layer chromatography of amino acids. *Journal of Chromatography B*, 912:98–104, 2013.

[20] Jelena Kečkeš, Jelena Trifković, Filip Andrić, Milica Jovetić, Živoslav Tešić, and Dušanka Milojković-Opsenica. Amino acids profile of serbian unifloral honeys. *Journal of the Science of Food and Agriculture*, 93(13):3368–3376, 2013.

[21] Emma A. Wistaff, Silvia Beller, Anton Schmid, John J. Neville, and Thorben Nietner. Chemometric analysis of amino acid profiles for detection of fruit juice adulterations – application to verify authenticity of blood orange juice. *Food Chemistry*, 343:128452, 2021.

[22] Zheng Sun, Lingling Zhao, Ni Cheng, Xiaofeng Xue, Liming Wu, Jianbin Zheng, and Wei Cao. Identification of botanical origin of chinese unifloral honeys by free amino acid profiles and chemometric methods. *Journal of Pharmaceutical Analysis*, 7(5):317–323, 2017.

[23] Guisheng Zhou, Mengyue Wang, Yang Li, Ying Peng, and Xiaobo Li. Rapid and sensitive analysis of 27 underivatized free amino acids, dipeptides, and tripeptides in fruits of siraitia grosvenorii swingle using hilic-uhplc-qtrap®/ms2 combined with chemometrics methods. *Amino Acids*, 47(8):1589–1603, 2015.

[24] M.J. Reis Lima, Andréia O. Santos, Soraia Falcão, Luísa Fontes, Edite Teixeira-Lemos, Miguel Vilas-Boas, Ana C.A. Veloso, and António M. Peres. Serra da estrela cheese's free amino acids profiles by uplc-dad-ms/ms and their application for cheese origin assessment. *Food Research International*, 126:108729, 2019.

[25] Marcel de Puit, Mahado Ismail, and Xiaoma Xu. Lcms analysis of fingerprints, the amino acid profile of 20 donors. *Journal of Forensic Sciences*, 59(2):364–370, 2014.

[26] Maria José Nunes de Paiva, Helvécio Costa Menezes, Paulo Pereira Christo, Rodrigo Ribeiro Resende, and Zenilda de Lourdes Cardeal. An alternative derivatization method for the analysis of amino acids in cerebrospinal fluid by gas chromatography–mass spectrometry. *Journal of Chromatography B*, 931:97–102, 2013.

[27] Meng Liu, Xu Zhang, and Tianwei Tan. The effect of amino acids on lipid production and nutrient removal by rhodotorula glutinis cultivation in starch wastewater. *Bioresource Technology*, 218:712–717, 2016.

[28] Xueguang Shao, Lingni Miao, Zhichao Liu, Huili Liu, and Wensheng Cai. Simultaneous identification and quantitative determination of amino acids in mixture by nmr spectroscopy using chemometric resolution. *Spectroscopy Letters*, 44(4):244–250, 2011.

[29] Oana Romina Botoran, Roxana Elena Ionete, Marius Gheorghe Miricioiu, Diana Costinel, Gabriel Lucian Radu, and Raluca Popescu. Amino acid profile of fruits as potential fingerprints of varietal origin. *Molecules*, 24(24), 2019.

[30] Xiangyan Shi, Gregory P. Holland, and Jeffery L. Yarger. Amino acid analysis of spider dragline silk using 1h nmr. *Analytical Biochemistry*, 440(2):150–157, 2013.

[31] Yihang Zeng, Wensheng Cai, and Xueguang Shao. Quantitative analysis of 17 amino acids in tobacco leaves using an amino acid analyzer and chemometric resolution. *Journal of Separation Science*, 38(12):2053–2058, 2015.

[32] Qin Ouyang, Yongcun Yang, Jizhong Wu, Quansheng Chen, Zhiming Guo, and Huanhuan Li. Measurement of total free amino acids content in black tea using electronic tongue technology coupled with chemometrics. *LWT*, 118:108768, 2020.

[33] Thulya Chakkumpulakkal Puthan Veettil and Bayden R. Wood. A combined near-infrared and mid-infrared spectroscopic approach for the detection and quantification of glycine in human serum. *Sensors*, 22(12), 2022.

[34] Ahmad Asghari, Ashraf Haj Hosseini, and Peyman Ghajarbeygi. Fast and non-destructive determination of histamine in tuna fish by atr-ftir spectroscopy combined with pls calibration method. *Infrared Physics Technology*, 123:104093, 2022.

[35] Ji Zhang, Bing Li, Qi Wang, Xin Wei, Weibo Feng, Yijiu Chen, Ping Huang, and Zhenyuan Wang. Application of fourier transform infrared spectroscopy with chemometrics on postmortem interval estimation based on pericardial fluids. *Scientific Reports*, 7(1):18013, 2017.

[36] Igor V. Kovalenko, Glen R. Rippke, and Charles R. Hurburgh. Determination of amino acid composition of soybeans (glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, 54(10):3485–3491, 2006.

[37] P. C. WILLIAMS, K. R. PRESTON, K. H. NORRIS, and P. M. STARKEY. Determination of amino acids in wheat and barley by near-infrared reflectance spectroscopy. *Journal of Food Science*, 49(1):17–20, 1984.

[38] Xiu-Shi Yang and Li-Li Wang. Determination of protein, fat, starch, and amino acids in foxtail millet [setaria italica (l.) beauv.] by fourier transform near-infrared reflectance spectroscopy. *Food Science and Biotechnology*, 22(6):1495–1500, 2013.

[39] Siv Skeie, Guri Feten, Trygve Almøy, Hilde Østlie, and Tomas Isaksson. The use of near infrared spectroscopy to predict selected free amino acids during cheese ripening. *International Dairy Journal*, 16(3):236–242, 2006.

[40] Xing Liu, Xin Zhang, Yu-Zhi Rong, Jin-Hong Wu, Yong-Jian Yang, and Zheng-Wu Wang. Rapid determination of fat, protein and amino acid content in coix seed using near-infrared spectroscopy technique. *Food Analytical Methods*, 8(2):334–342, 2015.

[41] Li Wang, Qiang Wang, Hongzhi Liu, Li Liu, and Yin Du. Determining the contents of protein and amino acids in peanuts using near-infrared reflectance spectroscopy. *Journal of the Science of Food and Agriculture*, 93(1):118–124, 2013.

[42] Zhe Jiao, Xiao-xi Si, Gong-ke Li, Zhuo-min Zhang, and Xin-ping Xu. Unintended compositional changes in transgenic rice seeds (oryza sativa l.) studied by spectral and chromatographic analysis coupled with chemometrics methods. *Journal of Agricultural and Food Chemistry*, 58(3):1746–1754, 2010.

[43] Igor V. Kovalenko, Glen R. Rippke, and Charles R. Hurburgh. Determination of amino acid composition of soybeans (glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, 54(10):3485–3491, 2006.

[44] Zhengzong Wu, Enbo Xu, Jie Long, Fang Wang, Xueming Xu, Zhengyu Jin, and Aiquan Jiao. Use of attenuated total reflectance mid-infrared spectroscopy for rapid prediction of amino acids in chinese rice wine. *Journal of Food Science*, 80(8):C1670–C1679, 2015.

[45] Fei Liu, Zonglai L. Jin, Muhammad Shahbaz Naeem, Tian Tian, Fan Zhang, Yong He, Hui Fang, Qingfu F. Ye, and Weijun J. Zhou. Applying near-infrared spectroscopy and chemometrics to determine total amino acids in herbicide-stressed oilseed rape leaves. *Food and Bioprocess Technology*, 4(7):1314–1321, 2011.

[46] Muhammad Zareef, Quansheng Chen, Qin Ouyang, Felix Y. H. Kutsanedzie, Md. Mehedi Hassan, Annavaram Viswadevarayalu, and Ancheng Wang. Prediction of amino acids, caffeine, theaflavins and water extract in black tea using ft-nir spectroscopy coupled chemometrics algorithms. *Anal. Methods*, 10:3023–3031, 2018.

[47] Hemlata Bhatia, Hamidreza Mehdizadeh, Denis Drapeau, and Seongkyu Yoon. Inline monitoring of amino acids in mammalian cell cultures using raman spectroscopy and multivariate chemometrics models. *Engineering in Life Sciences*, 18(1):55–61, 2018.

[48] Xin Zhang, Shaohua Lu, Yi Liao, and Zhuoyong Zhang. Simultaneous determination of amino acid mixtures in cereal by using terahertz time domain spectroscopy and chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 164:8–15, 2017.

[49] Ofélia Anjos, António J A Santos, Teresa Dias, and Letícia M Estevinho. Application of ftir-atr spectroscopy on the bee pollen characterization. *Journal of Apicultural Research*, 56(3):210–218, 2017.

[50] Raluca Daniela Isopescu, Roxana Spulber, Ana Maria Josceanu, Dan Eduard Mihaiescu, and Ovidiu Popa. Romanian bee pollen classification and property modelling. *Journal of Apicultural Research*, 59(4):443–451, 2020.

[51] Maria Cristina A. Costa, Marcelo A. Morgano, Marcia Miguel C. Ferreira, and Raquel F. Milani. Analysis of bee pollen constituents from different brazilian regions: Quantification by nir spectroscopy and pls regression. *LWT*, 80:76–83, 2017.

[52] Nesrin Ecem Bayram, Yusuf Can Gercek, Saffet Çelik, Nazlı Mayda, Aleksandar Ž. Kostić, Aleksandra M. Dramićanin, and Aslı Özkök. Phenolic and free amino acid

profiles of bee bread and bee pollen with the same botanical origin – similarities and differences. *Arabian Journal of Chemistry*, 14(3):103004, 2021.

[53] Ana M. González Paramás, J. Alfonso Gómez Bárez, Carlos Cordón Marcos, Rafael J. García-Villanova, and José Sánchez Sánchez. Hplc-fluorimetric method for analysis of amino acids in products of the hive (honey and bee-pollen). *Food Chemistry*, 95(1):148–156, 2006.

[54] Joseph D Mancias and Jonathan Goldberg. Structural basis of cargo membrane protein discrimination by the human copii coat machinery. *The EMBO Journal*, 27(21):2918–2928, 2008.