КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри молекулярної біотехнології та біоінформатики
доц., к.б.н. Олексій Нипорко
Протокол №___засідання кафедри
від “___” _____20__р.

ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ ГЛІКОВАНОГО ГЕМОГЛОБІНУ У FTIR СПЕКТРАХ КРОВІ

Курсова робота
студента спеціальності 091 Біологія
ОП «Біологія (Високі технології)»
Мельниченка Миколи Валентиновича

Науковий керівник від кафедри
доцент кафедри молекулярної біотехнології та біоінформатики
к.ф.-м.н. **Войтешенко Іван Сергійович**

Робота виконана у відділі епігенетики
Інституту Геронтології імені Д. Ф. Чеботарьова
під керівництвом к.б.н. **Дмитра Краснєнкова**

Оцінка захисту роботи

_____

Київ – 2024 р.

Анотація

Мельниченко М. В. Сучасні методи діагностики діабету зазвичай базуються на вимірюванні миттєвого рівня глюкози за допомогою різних тестів. На відміну від цих методів, HbA1c вважається надійним показником, оскільки він відображає середньозважену концентрацію глюкози у крові за останні 120 днів. У даній роботі було створено дві регресійні моделі для прогнозування рівня глікованого гемоглобіну на основі FTIR спектрів крові. Перша модель була реалізована за допомогою PLSR, а друга використовувала поліномне наближення оригінальних спектрів з використанням псевдофойгтівської функції для параметризації піків спектру та побудови регресійних моделей на їх основі. За допомогою H2O AutoML була створена модель використовуючи площі розкладених піків для передбачення HbA1c та визначі найважливіші змінні.

Ключові слова: глікований гемоглобін, HbA1c, ATR-FTIR, PLSR, псевдофойгтівська функція, поліномне наближення;

Зміст

Умовні позначення

FPG – (fasting plasma glucuse) рівнь глюкози у крові натще.
OGTT – (Oral glucose tolerance test) оральний глюкозотолерантний тест.
Hb – гемоглобін.
HbA1c – глікований гемоглобін.
PLSR – (partial least square regression) регресія методом часткових квадратів.
PCA – (principal component analysis) аналіз головних компонент.
OLS – (ordinary least squares) метод найменших квадратів.
FTIR – Фур'є спектроскопія інфрачервоного випромінювання.
AGE – (Advanced glycation end-products) кінцеві продукти глікування.
RAGE – (Receptor for AGE) рецептори до AGE.
n – Число спостережень.
m – Число незалежних змінних.
k – Число залежних змінних.
l – число компонент PLSR моделі.
$\mathbf{X}$ – $(n \times m)$ матриця з незалежними змінними.
$\mathbf{Y}$ – $(n \times k)$ матриця з залежними змінними.
$\mathbf{V}$ – $(m \times l)$ матриця з лівими сингулярними векторами на кожній ітерації.
$\mathbf{W}$ – $(k \times l)$ матриця з правими сингулярними векторами на кожній ітерації.
$\mathbf{T}$ – $(n \times l)$ матриця оцінки $\mathbf{X}$.
$\mathbf{U}$ – $(n \times l)$ матриця оцінки $\mathbf{Y}$.
$\mathbf{P}$ – $(l \times m)$ Матриця навантаження на $\mathbf{X}$.
$\mathbf{C}$ – $(l \times k)$ Матриця навантаження на $\mathbf{Y}$.
$\mathbf{C_{XY}}$ – $(m \times l)$ Матриця крос-коваріації між $\mathbf{X}$ та $\mathbf{Y}$.
$\mathbf{E}_i$ – Серія різних матриць зі залишковими значеннями.
$\mathbf{x}, \mathbf{y}, \mathbf{t}$ – усі вектори-стовпці позначаються малими товстими літерами.
$\|\cdot\|$ – векторна нормалізація.
FWHM – (full width half maximum) ширина на піввисоті.

Вступ

Діабет є розповсюдженою та небезпечною групою метаболічних захворювань, пов'язаних з утворенням й використанням інсуліну. Оскільки основні існуючі методи для діагностування діабету відображають лише миттєву концентрацію глюкози у крові, вони можуть давати нестабільні результати через повсякденні чинники. Інший метод діагностування полягає у вимірюванні рівня глікованого гемоглобіну, (Розділ 1.1) який є довготривалим індексом глюкози у крові. Однак основна лабораторна методика вимірювання глікованого гемоглобіну є високоефективна рідинна хроматографія (HPLC), що потребує достатню кількість часу. Останнім часом Фур'є спектроскопія інфрачервоного випромінювання (FTIR) набула широкої популярності для діагностування різних хвороб, аналізуючи спектри поглинання біологічних матеріалів, зокрема крові. Оскільки більшість хвороб супроводжуються зміною складу чи структури тканин на молекулярному рівні, спектроскопія може бути гарним способом для детекції хвороб. Головною перевагою FTIR є те, що його модифікація за допомогою ATR аксесуара (Розділ 1.2) дозволяє досліджувати біологічні матеріали з мінімальною попередньою обробкою. Використання результатів FTIR спектроскопії для кількісного встановлення речовин не є новим і цей метод використовувався з самого початку розвитку хемометрики. Через мультиколінеарність спектральних даних використовують регресії методом часткових квадратів (PLSR), яка моделює зв'язок між незалежними та залежними змінними створюючи нові компоненти.

Об'єктом дослідження цієї роботи є FTIR спектри крові, які можна характеризувати основними піками, що відповідають структурним особливостям молекул у зразку. Кількісне встановлення досліджуваної речовини у зразку використовуючи поглинання світла можливо завдяки закону Ламберта-Бера: $A = \varepsilon lc$, який пов'язує абсорбанс (A) та концентрацію речовини (c) через молярний коефіцієнт абсорбції (ε) та оптичний шлях (l).

Поліномне наближення оригінальних спектрів може бути корисним для параметризації піків, завдяки чому ми можемо спостерігати їх площу, зсув, інтенсивність та ширину. Кожен з цих параметрів чи їх комбінація потенційно може

бути використана для числельного чи якісного передбачення речовин у досліджуваному зразку.

Метою даної роботи є створення регресійної моделі для передбачення рівня глікованого гемоглобіну з використанням FTIR спектрів крові та подальшого аналізу отриманих результатів. Окрім цього другою методю даної роботи є дослідити можливості використання поліномгого наближення FTIR спектру крові псевдовойтівською функцією для розкладання спектрів на окремі піки та прогнозування вмісту %HbA1c.

Розділ 1 Літогляд

1.1 Глікований гемоглобін

Діабет – це група хронічних захворювань, які характеризуються підвищеним рівнем глюкози через несправність у виробництві та/або утилізації інсуліну. Глюкоза відіграє важливі метаболічні функції, тому гіперглікемія асоціюється з підвищеними ризиком до розвитку життєво небезпечних ускладнень [1]. Станом на 2017 рік приблизна всесвітня кількість людей з діабетом сягає поняд 450 млн. та передбачено може досягти позначки у 693 млн. до 2045 року. Розповсюдження діабету станом на 2014 рік можна побачити на Рис. 1.1. Також було підраховано, що приблизно половина людей (49.7%) з діабетом не мають діагнозу [2, 3]. Діабет має значний вплив на суспільство через великі медичні витрати, втрачену продуктивність та передчасну смертність. Загальна вартість діабету в США у 2022 році становила 412,9 млрд доларів, з яких 306,6 млрд (74%) витрачались на прямі медичні витрати.

Зважаючи на економічний та соціальний ефект діабету його доступне діа-

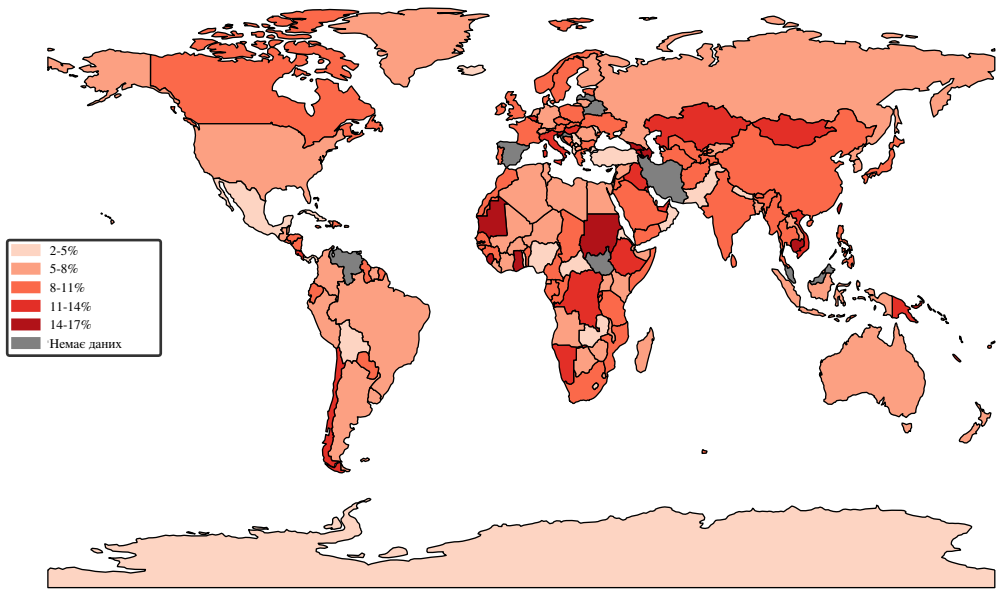


Рис. 1.1 Розповсюдження діабету у % від загальної популяції. Дані 2014 року отримані з ncdrisc.org.

гностування є надзвичайно важливим. У клінічній практиці виявлення діабету базується на вимірюванні рівня глюкози у крові. Можна виділити три основні методики [4]:

1. Визначення рівня глюкози в крові натще (FPG). Не рекомендується вживати їжу за 8 годин до проведення тесту. Значення більші за 126 мг/дЛ (7ммоль/Л) є критерієм для діагностування діабету.
2. Глюкозотолерантний тест (OGTT) використовується для визначення того, наскільки добре організм може переробляти більшу кількість цукру. Цей метод передбачає прийом 75г. розчиненої глюкози та фіксування значення глюкози у крові через 2 години. Критерієм для діагностування є значення більші за 200 мг/дЛ (11.1 ммоль/Л).
3. HbA1c критерій, що відображає кількість глікованого гемоглобіну. Для його визначення зазвичай використовують рідинну хромотографію високого тиску (HPLC).

Глікований гемоглобін є ефективним та довготривалим показником рівня глюкози, який часто використовується при діагностуванні та моніторингу діабету [4, 5]. Він складається з інших аддуктів та HbA, який, в свою чергу, включає в себе HbA1a, HbA1b та HbA1c. HbA1c є основним компонентом HbA та формується у результаті неензиматичної реакції альдегідної групи глюкози з аміно групами гемоглобіну з утворенням основи Шиффа та з подальшим перегрупуванням Амадорі, утворючи кетозамін (Рис. 1.2). Цей процес також відомий як реакції Майяра [6]. HbA1c є основним біохімічним індикатором довготривалого рівня глюкози у крові. Він обчислюється як відношення концентрації глікованого гемоглобіну до загальної кількості гемоглобіну у крові і вимірюється у відсотках чи ммоль/моль. Враховуючи, що еритроцит має тривалість життя приблизно 120 діб, HbA1c відображає середньозважену концентрацію глюкози за цей проміжок часу [7]. Це відбувається тому, що в будь-який проміжок часу значення HbA1c утворюється усіма циркулюючими у крові еритроцитами, причому недавні рівні глюкози (3-4 тижні) мають більший вплив, аніж старіші (3-4 місяці). Здорові особини мають HbA1c на рівні від 4% до 5.6%, тоді як прediабет характеризується значеннями в діапазоні 5.6-6.4%. Визначення відсотку глікованого гемоглобіну на рівні від 6.5% є підставою для діагнозування діабету [4]. Хоча HbA1c є корисним показником рівня цукру у крові, важливо усвідомлюва-

ти, що він є непрямим відображенням цього рівня, адже критерій в 6.5% здатен діагтостувати лише 30% від результатів, які були встановлені використовучи комбінацію методик FPG, OGTT та/або HbA1c разом [8]. Також відомо, що

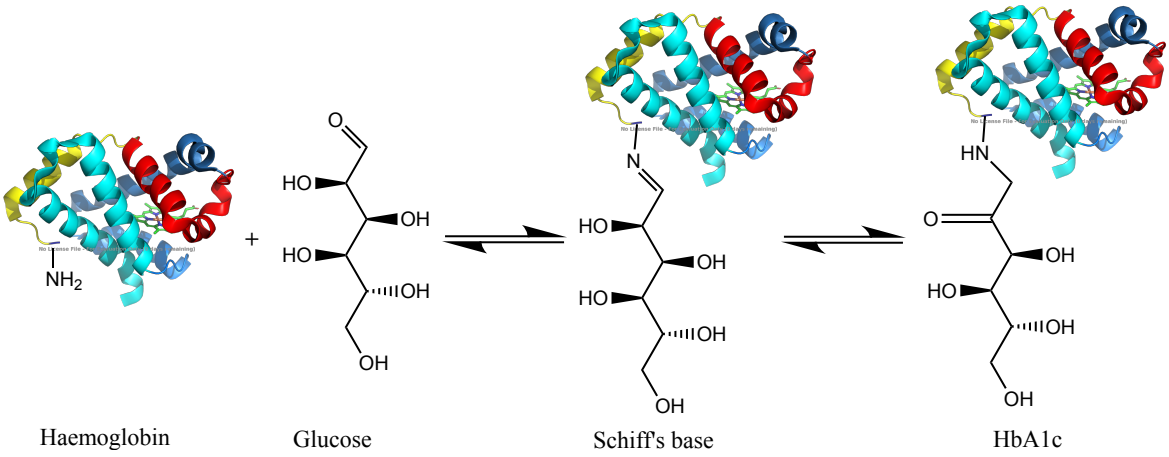


Рис. 1.2 Процес утворення HbA1c в результаті реакцій Майяра.

HbA1c є важливим біомаркером старіння [9]. Аналіз даних з досліджень SIGT та NHANES показали, що кожні 10 років життя значення HbA1c збільшується на 0.93 ммоль/моль (0.085%) та 0.87 ммоль/моль (0.08%) відповідно [10].

HbA1c відноситься до групи кінцевих продуктів глікування (англ. AGEs), які утворюються внаслідок глікування протеїнів чи ліпідів після їх тривалого контакту з цукрами. Все частіше з'являються докази того, що AGEs впливають на розвиток серцевосудинних захворювань за допомогою двох механізмів: утворення поперечних зв'язків у внутрішьо- та/або позаклітинних білків або приєднанню до RAGE рецепторів [11]. Надмірне утворення поперечних з'язків унаслідок підвищеного вмісту AGEs може призвести до втрати еластичності серцевих судин та розвитку діастолічної дисфункції. Інший механізм розвитку цих ускладнень відбувається через приєднання до AGE рецепторів, найважливішим з яких є RAGE. Одним із ефектів активації цих рецепторів є фіброз [12], порушення метаболізму кальцію [13] та гіперчутливості до запальних цитокінів в макрофагах й клітинах ендотелію [14]. Отже, AGEs та його рецептори грають ключову роль у розвитку серцевосудинних ускладнень у пацієнтів з діабетом [15].

1.2 Фур'є-спектроскопія

Коли світло проходить через речовину, частина фотонів, енергія яких дорівнює різниці двох енергетичних рівнів, поглинаються електронами атомів, завдяки чому останні переходять на новий енергетичний рівень та збуджуються. За допомогою цих фізичних характеристик ми можемо отримати спектри поглинання або емісії, які зазвичай використовуються для якісного чи кількісного аналізу досліджуваної речовини.

Інфрачервона (ІЧ) спектроскопія займається вивченням спектрів взаємодії речовин з електромагнітним випромінюванням в ІЧ діапазоні (400 $см^{-1}$ - 4000 $см^{-1}$). ІЧ спектри виникають через коливальні чи обертальні рухи молекули внаслідок поглинання енергії світла. Інфрачервона Фур'є-спектроскопія (далі FTIR) є технікою для отримання ІЧ спектру з твердих речовин, рідин та газів. Концептуальна ідея FTIR полягає у тому, щоб замість того, щоб світити монохромним світлом на кожну довжину хвилі і вимірювати поглинання, для отримання кінцевого результату ми використовуємо різнобарвний пучкок світла. Для інтерпретації отриманих даних нам потрібно використати Фур'є трансформацію, завдяки чому цей метод і отримав свою назву.

Інтерферометр Майкельсона, який використовується у FTIR складається з джерела поліхромного світла, дільника променя, рухомого та стаціонарного дзеркала (Рис. 1.2). Після випромінювання пучок світла проходить крізь дільник та ділиться навпіл (ідеальний випадок), прямуючи до двох дзеркал. Звідти він віддзеркалюється назад і частина світла проходить через досліджуваний зразок. Використовуючи різницю довжин пройдених оптичних шляхів через різні дзеркала та варіюючи відстань рухомого дзеркала ми отримуємо інтерферограму, яка надалі аналізується комп'ютерними застосунками з використанням трансформації Фур'є.

Одним із найважливіших переваг FTIR це в його здібності вимірювати спектри з більшим SNR (signal to noise ratio), що є метрикою, яка описує відношення корисної інформації до фонових шумів. Іншими перевагами FTIR є швидкість у приготуванні зразків та отриманням спектрів. Хоча FTIR має декілька важливих переваг над іншими методами, він стикається з недоліком у вигляді артефактів, які можуть виникати через зміну концентрації CO_2 чи водяної пари у зразку [16].

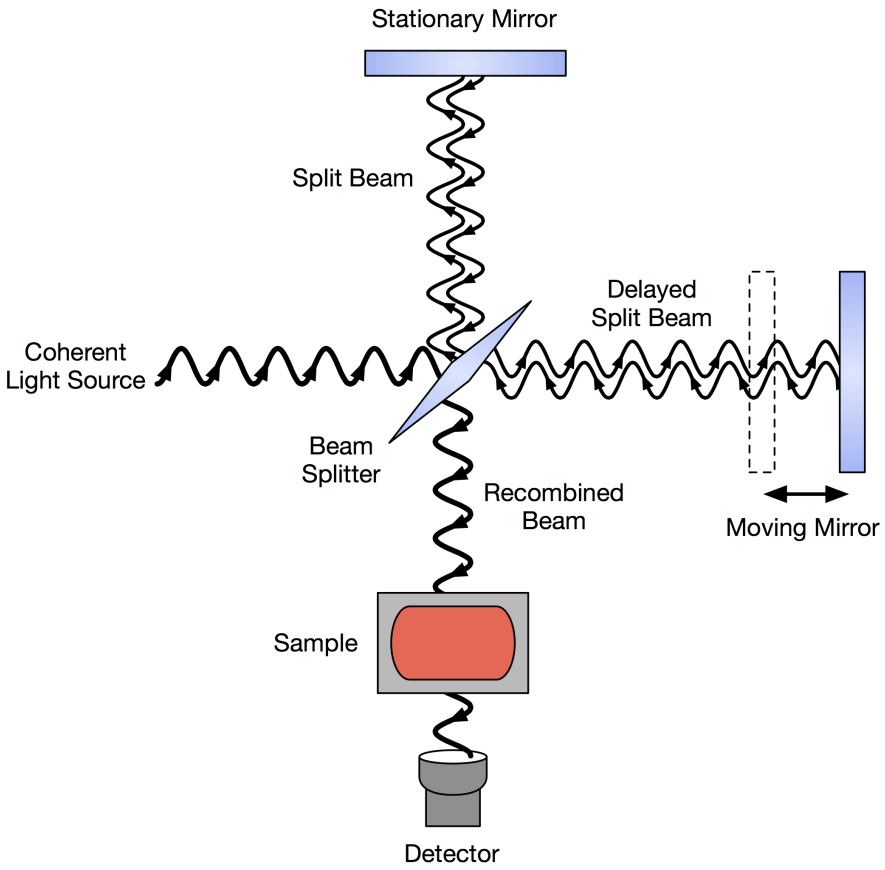


Рис. 1.3 Будова інтерферометра Майкельсона. Взято з **wikipedia.org**, автор - Sanchonx.

Зараз майже завжди разом з FTIR використовують метод порушення повного внутрішнього відбиття (англ. ATR), який є допоміжним оптичним методом для вивчення поверхневих шарів речовини в умовах повного внутрішнього відбиття. Повне внутрішнє відбиття виникає на межі двох середовищ, коли хвилі повністю

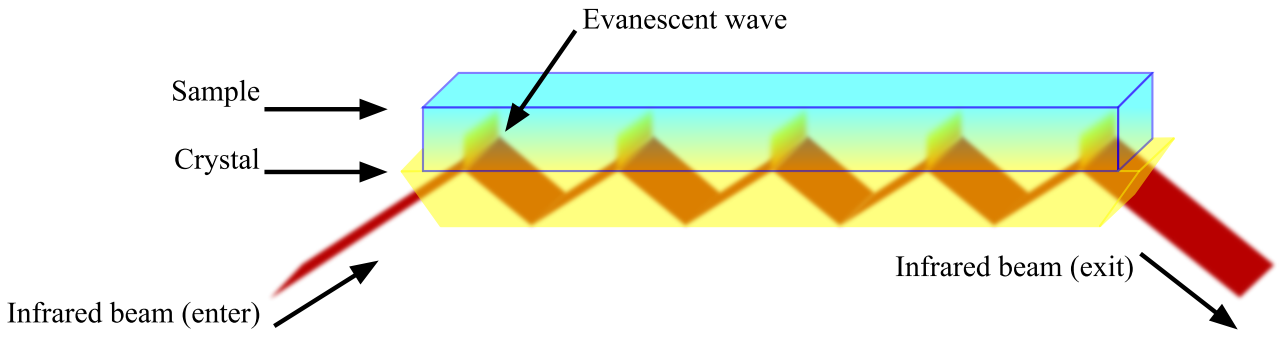


Рис. 1.4 Демонстрація ATR. Світло багаторазово відбивається від границі розділу кристалу з великим показником заломлення (жовтий) та зразком. Взято з **wikipedia.org**, автор - Fulvio314.

відбиваються назад у середовище без рефракції. Це відбувається у випадку, коли перше середовище (з якого йде хвиля) має більший показник заломлення ніж друге та кут, під яким проходить хвиля через розподіл середовищ, є достатньо великим. Вищезгаданий кут має назву критичного і визначається як найменший кут, за якого не відбувається рефракція. Згідно із законом Снеліуса:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{1.1}$$

Враховуючи, що рефракція припиняє відбуватись за $\theta_2 = 90°$:

$$\theta_c = \arcsin\left(n_2/n_1\right) \tag{1.2}$$

де n_1 - показник заломлення першого середовища, n_2 - другого, θ_1 - кут, під яким приходить хвиля, θ_2 - кут рефракції, θ_c - критичний кут. Результуючим ефектом повного внутрішнього відбиття є відбиття хвилі як мінімум один раз та утворення еванесцентної хвилі, яка потрапляє у зразок. Глибина проникнення цієї хвилі досить невелика, приблизно 0.5-2 мікрометри. Пучок світла потім виходить зі зразка та фіксується детектором. ATR дає можливість вивчати велику кількість біологічних матеріалів без потреби в їх попередній обробці. Наприклад, у цій роботі зразки крові були оброблені лише шляхом їх висушування (Розділ 2.1).

Типовий FTIR спектр можна характеризувати наявнюсті основних піків, які відповідають за коливання чи розтягування різних функціональних груп, наявних у молекулах зразку (Рис. 1.5). На Таблиці 1.1 наведено значення деяких піків, які можна знайти на FTIR спектрах крові.

1.3 Застосування FTIR для дослідження крові

За останні декілька десятиліть FTIR зазнало швидкого розвитку у напрямку простішого, швидкого та точнішого діагностування та моніторингу пацієнтів [цитату]. Оскільки кожна молекула має свій унікальний "слід"на спектрограмі, ефективний аналіз останньої дає змогу досліджувати стани різних органів чи тканин завдяки чисельному та/або якісному складу характерних сполук.

Одним із таких прикладів є потенційне застосування FTIR для ранньої детекції дитячої лейкемії [18]. У цьому дослідженні була виявлена різниця між спектрограмами здорових людей та хворих на лейкемію. Розділення здорових

та хворих спектрів відбувалося за допомогою диференціації спектрів з подальшим використанням методів для зменшення вимірності даних (PCA, ICA, і т. п.). Основна різниця спостерігалась на 2 похідній проміжку 1600-1700 см$^{-1}$, який відповідає за вторинну структуру білка з характерними піками для β-листів, α-спіралей та β-поворотів. Відповідний спектр був реконструйований з використанням поліномного наближення. З отриманих площ піків зробили висновок про відмінність у кількості бета структур (β-лист + β-поворот) у хворій й здорових групах і використали цю знахідку для створення класифікаційної моделі. Іншим прикладом використання FTIR є раннє діагнозування ВІЛ [19]. Головні відмінності були помітні на 900-1800 см$^{-1}$ (основний регіон для аналізу біомаркерів). Для вирішення класифікаційної задачі було використано багато моделей, але найкращий результат спостерігався у PCA-SVM, SPA-LDA та GA-LDA.

Одним із перспективних застосувань FTIR є створення моделей для швидкого діагностування діабету 2 типу [20]. Через вміст "шумів"у спектрах крові у цьому досліджені був використаний фільтр Савітського-Голая з подальшою екстракцією компонент методом PCA. Далі на цих компонентах будувалась кла-

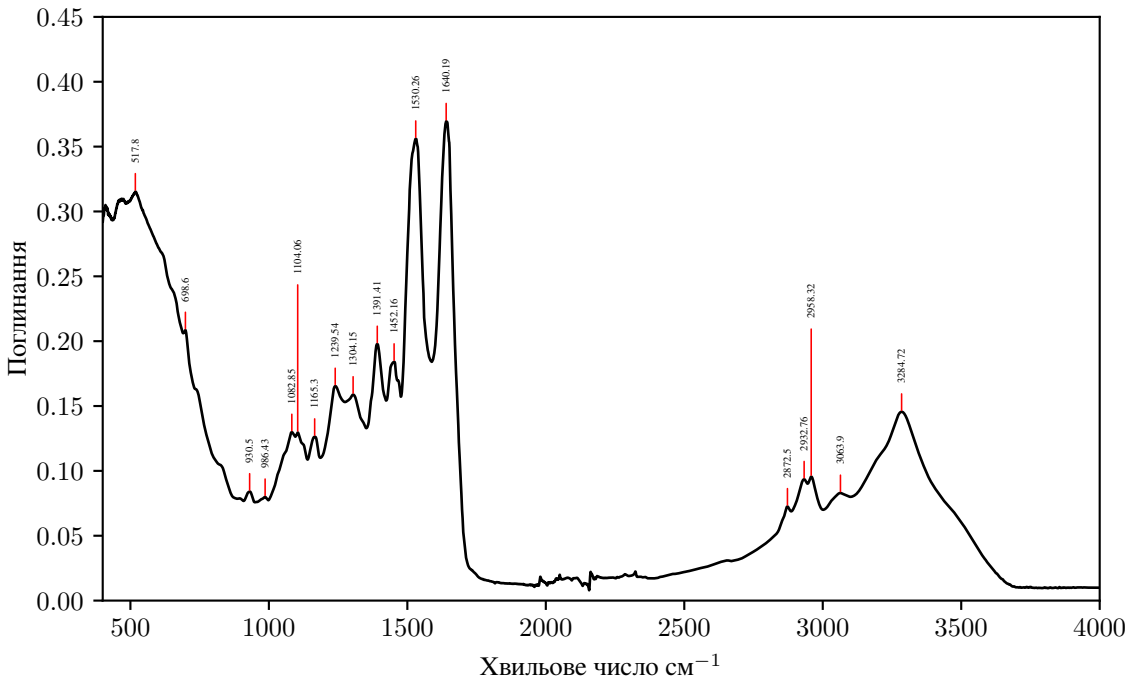


Рис. 1.5 Середній FTIR спектр крові.

Табл. 1.1 Біохімічне значення деяких піків FTIR спектру крові. Таблиця була адаптована з [17].

Хвильове число $см^{-1}$	Значення піку
930	ДНК з лівосторонньою спіраллю (Z-форма).
986	Розтягування (C3N) порфіринових конфігурацій.
1023	C-C розтягування у фосфоліпідах, глікоген.
1055	Ліпіди, холестерол, C-O розтягування глюкози.
1084	C-C розтягнення (PO_2) у фосфоліпідах та білках.
1104	Розтягнення глікогену CO у поєднанні з OH деформацією глюкози.
1122-1128	C-C розтягування у ліпідах, C-N розтягування у білках, D-маноза, глюкоза, розтягування C-метилу в гемі.
1165	C-O групи з серину, тріоніну та тирозину протеїнів.
1212	C-H деформації оксигемоглобіну, деоксигемоглобін, асиметричні розтягування PO_2^-.
1223-1225	Оксигемоглобін, нуклеїнові кислоти.
1239	Асиметричні та симетричні розтягування PO_2^-.
1304	Амід III (ліпіди, аденін).
1390	Згинання CH_3.
1452	CH_2-деформація в ліпідах, згинання C-H у білках.
1539-1546	Амід II, CN і CNH згинання у триптофані, деоксигемоглобін, метгемоглобін.
1560–1590	Деформація C=C, триптофан, фенілаланін, ацетоацетат, рибофлавін, гемоглобін, деоксигемоглобін, карбогемоглобін.
1641-1658	Амід I, C=O розтягування водневих зв'язків, оксигемоглобін.
1660	Білки Аміду I, розтягування C=C у ненасичених ЖК.
2872	Симетричне розтягування C-H зв'язку в CH_3.
2930	Асиметричне розтягування C-H зв'язку в CH_2 (жирні кислоти/холестерол).
2953	Асиметричне розтягування C-H зв'язку в CH_3. (жирні кислоти/ліпіди).
3294	Амід A, H-зв'язане N-H розтягування.

сифікаційна модель з використанням алгоритму XGBoost. Отримана точність (95.65%), чутливість (95.24%) та специфічність (96%) демонструють можливість використання даного методу для швидкого діагностування діабету 2 типу. Ва-

жливим напрямком діагностування діабету є розвиток неенвазійних методів, які не потребують руйнування біологічних оболонок людини. Було продемонстровано, що встановлення діабету можливе за допомогою FTIR зразків слини [21], губ [22] та нігтів [23].

Іншим потенційно важливим застосуванням FTIR є створення регресійних моделей для передбачення рівня HbA1c [24]. У вищезгаданній статті рівень HbA1c встановлювався завдяки двом PLSR моделям, які прогнозують загальний рівень звичайного та глікованого гемоглобіну і в результаті ми отримуємо їх відношення. Окрім діагностування діабету рівень HbA1c потенційно може бути використаний для передбачення біологічного віку, оскільки HbA1c є гарним біомаркером старіння [9] та існують відмінності між FTIR спектрами крові різних вікових груп [17].

Розділ 2 Матеріали та методи

2.1 Матеріали

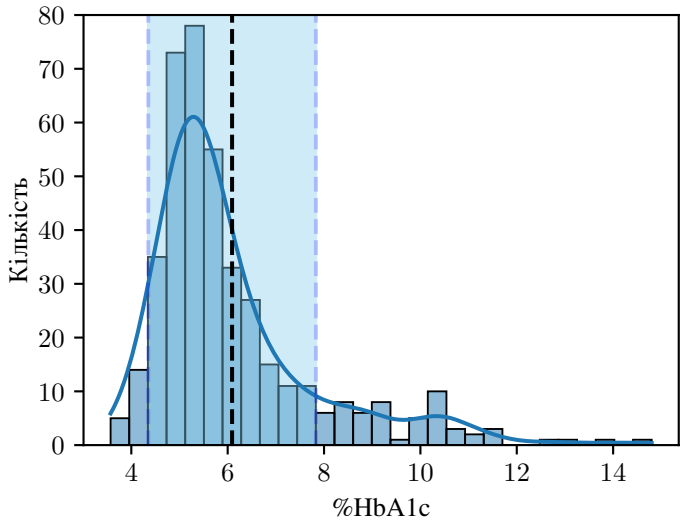


Рис. 2.1 Розподіл глікованого гемоглобіну пацієнтів з візуально наведеним середнім значенням та одним квадратичним відхиленням.

Усі учасники експерименту дали інформовану згоду. Середній вік учасників складав 59 ± 16 років. Було зібрано 294 зразки крові за допомогою вакутайнерів об'ємом 4 мл, які містили EDTA. 100 мЛ кожного зразка були висушені за кімнатної температури щоб позбутися сильного фону води на майбутніх спектрах. Для отримання спектрів на проміжку 4000 см$^{-1}$ - 400 см$^{-1}$ поглинання використовувався Nicolet iS50 FTIR спектрометр з оснащеним приладом для ATR. Останній в свою чергу був діамантовим та дозволяв одноразове заломлення світла. Усі спектри були отримані з роздільною здатністю в 4 см$^{-1}$ та для кожного результату було зроблено 32 скани. Вміст глікованого гемоглобіну був визначений за допомогою високоефективної рідинної хроматографії. Розподіл значень глікованого гемоглобіну наведений на Рис. 2.1.

2.2 Регресійна модель

Головна мета регресії це встановити вагову матрицю $\hat{\beta}$, що мінімізує похибку передбачення моделі. Формально це знаходження такого аргументу квадратичної функції втрат, який мінімізує її значення, а отже, і похибку передбачення

регресійної моделі:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \tag{2.1}$$

Для лінійної регресії вирішити цю оптимізаційну проблему можна за допомогою методу найменших квадратів (англ. OLS), рішення якого має наступний вигляд у матричній формі:

$$\hat{\beta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} \tag{2.2}$$

Регресія методом часткових квадратів (англ. PLSR) це класичний та ефективний хемометричний регресійний метод, який широко використовується в спектроскопічному аналізі через мультиколінеарність незалежних змінних [25, 26, 27]. PLS регресія поєднує в собі риси аналізу головних компонент (англ. PCA) та множинної лінійної регресії та узагальнює їх. Головна ідея PLSR полягає у ітеративному пошуку нормалізованих векторів $\hat{\mathbf{v}}$ та $\hat{\mathbf{w}}$, які забезпечують найбільшу коваріацію між $\mathbf{X}$ та $\mathbf{Y}$ після проєкції на них.

$$\tilde{\mathbf{v}}, \tilde{\mathbf{w}} = \arg\max_{\mathbf{t},\mathbf{u}} \mathbb{E}[\mathbf{t} \cdot \mathbf{u}] = \arg\max_{\mathbf{v},\mathbf{w}} \mathbb{E}[(\mathbf{X}\mathbf{v}) \cdot (\mathbf{Y}\mathbf{w})] \tag{2.3}$$

Перегрупування минулого рівняння (3) дає наступну оптимізаційну задачу з двома обмеженнями, де $\mathbf{C_{XY}}$ – матриця крос-коваріації:

$$\tilde{\mathbf{v}}, \tilde{\mathbf{w}} = \arg\max_{\mathbf{v},\mathbf{w}} \mathbf{v}^\intercal \mathbf{C_{XY}} \mathbf{w}, \text{ де } \|\mathbf{v}\| = \|\mathbf{w}\| = 1 \tag{2.4}$$

Можна продемонструвати [28], що рішенням рівняння (2.4) будуть лівий та правий сингулярні вектори сингулярного розкладу матриці крос-коваріації $\mathbf{C_{XY}}$, що відповідають найбільшому сингулярному значенню.

$$\mathbf{C_{XY}} = \mathbf{A\Sigma Z}^\intercal = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{bmatrix} \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{k1} & z_{k2} & \cdots & z_{kk} \end{bmatrix}^\intercal \tag{2.5}$$

На нормалізований вектор $\tilde{\mathbf{v}}$ далі відбувається відображення $\mathbf{X}$ і результат використовується для прогнозування $\mathbf{Y}$.

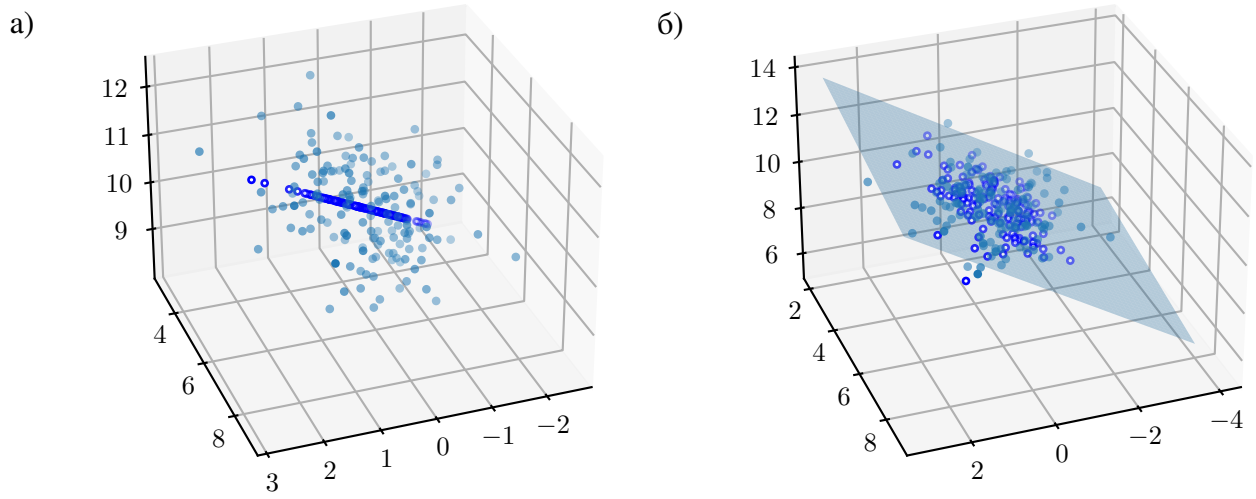


Рис. 2.2 Приклад проєкції $\mathbf{X}$ вимірністю $n \times 3$ на (а) одну та (б) дві компоненти.

PLS регресію геометрично можна інтерпретувати як проєкцію матриці $\mathbf{X}$ (n точок у m-вимірному просторі) на l-вимірну гіперплощину, де l - кількість компонентів моделі (Рис. 2.2). Така проєкція передбачає, що отримані компоненти гарно "описують"$\mathbf{Y}$. Саме тому була запронована інша назва для PLS, а саме проєкція на латентний простір (англ. **p**rojection on **l**atent **s**pace).

Пілся сингулярного розкладу матриці крос-коваріації (2.5) ми отримуємо два вектори $\tilde{\mathbf{v}} = (a_{11}, a_{21}, \ldots, a_{m1})$, $\tilde{\mathbf{w}} = (z_{11}, z_{21}, \ldots, z_{k1})$. Як з'ясувалось, правий та лівий сингулярні вектори, що відповідають другому сингулярному значенню не дають другу найбільшу коваріацію після проекції на них. Через це для обчислення кожної наступної компоненти $j = 1, 2, \ldots, l$ $(1 \leq l \leq m)$ ми повинні прибрати вплив минулих сингулярних векторів на матрицю $\mathbf{X}$ використовуючи дефляцію матриці (англ. matrix deflation).

$$\mathbf{X}_{j+1} = \mathbf{X}_j - \mathbf{t}_j \mathbf{p}_j^\mathsf{T} \tag{2.6}$$

Цей процес ітеративного пошуку та зберігання сингулярних векторів з подальшою дефляцією матриці повторюється в залежності від необхідної кількості компонент. Отримані матриці $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_l)$ та $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_l)$ називаються матрицями оцінки (англ. score matrix) $\mathbf{X}$ та $\mathbf{Y}$ відповідно.

$$\mathbf{T} = \mathbf{X}\mathbf{V} \tag{2.7}$$

$$\mathbf{U} = \mathbf{Y}\mathbf{W} \tag{2.8}$$

Вони мають властивість гарно "описувати"матриці $\mathbf{X}$ та $\mathbf{Y}$ після перемноження на навантаження (англ. loadings) $\mathbf{P}^\intercal$ й $\mathbf{C}^\intercal$. Тут і у подальшому $\mathbf{E}_i$ використовується для позначення матриці помилок передбачення.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\intercal + \mathbf{E}_1 \tag{2.9}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^\intercal + \mathbf{E}_2 \tag{2.10}$$

Матриця $\mathbf{T}$ надалі використовується для створення нової регресійної моделі на цих змінних і прогнозування $\mathbf{Y}$.

$$\mathbf{Y} = \mathbf{T}\mathbf{Q} + \mathbf{E}_3 \tag{2.11}$$

Обраховуючи $\mathbf{Q}$ за допомогою методу найменших квадратів та підставляючи (2.7):

$$\mathbf{Q} = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{Y} + \mathbf{E}_6 = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{X}^\intercal\mathbf{V}^\intercal\mathbf{Y} + \mathbf{E}_6 \tag{2.12}$$

Підставляючи (2.7) та (2.12) назад у (2.11):

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{V}(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{X}^\intercal\mathbf{V}^\intercal\mathbf{Y} = \mathbf{X}\mathbf{B} \tag{2.13}$$

Звідки отримуємо рівняння для коефіцієнтів PLSR:

$$\mathbf{B} = \mathbf{V}(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{X}^\intercal\mathbf{V}^\intercal\mathbf{Y} \tag{2.14}$$

Підсумовуючи, можна сказати що використовуючи лінійну комбінацію незалежних змінних з $\mathbf{X}$ PLS шукає нові латетні змінні які у подальшому використовуються для прогнозування $\mathbf{Y}$.

2.3 Обробка даних

Фільтр Савітського-Голая (далі SG) є одним із найпопулярніших фільтрів у хемометриці, який використовується для згладжування даних [29]. Він має два параметри: довжина вікна w та ступінь поліному o. Головна ідея цього фільтру полягає у створені поліномної регресії зі ступенем o для кожного вікна довжиною w (Рис. 2.3). Математично це можна виразити як конволюцію залежних змінних y з множиною w коефіцієнтів C_i:

$$Y_j = \sum_{i=\frac{1-w}{2}}^{\frac{w-1}{2}} C_i y_{j+i}, \quad \frac{w+1}{2} \leq j \leq n - \frac{w-1}{2} \tag{2.15}$$

Якщо дані розташовані на однаковій відстані то існує аналітичне рішення для встановлення коефіцієнтів поліномної регресії. Найпростішим прикладом фільтру Савітського-Голая є рухоме середнє, яке часто використовується для аналізу часових рядів та обраховується наступним чином (у даному випадку $o = 1$):

$$\text{MA}_w = \frac{1}{w} \sum_{i=n-w+1}^{n} p_i \tag{2.16}$$

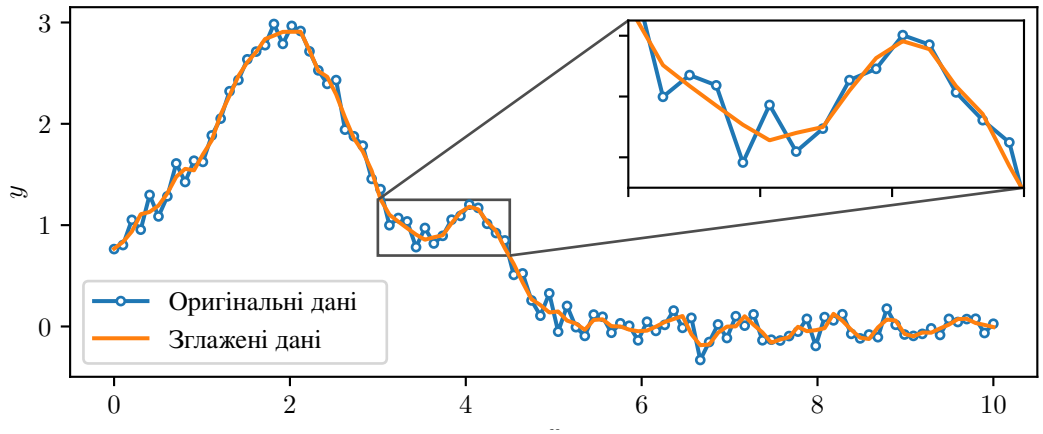


Рис. 2.3 Приклад використання фільтра Савітського-Голая для синтетичних даних ($w = 5, o = 2$).

Окрім згладжування, іншою важливою обробкою спектрів є їх нормалізація. Існує декілька основних методів:

1. **Векторна нормалізація**:

$$v_{\text{new}} = \frac{v_{\text{old}}}{\|v_{\text{old}}\|} \tag{2.17}$$

2. **Нормалізація за Амід I чи Амід II піком:**

$$v_{\text{new}} = \frac{v_{\text{old}}}{v_a} \tag{2.18}$$

де v_a - це значення інтенсивності поглинання Амід I чи Амід II піку.

3. **SNV (standard normal variate):**

$$v_{\text{new}} = \frac{v_{\text{old}} - v_\mu}{v_\sigma} \tag{2.19}$$

де v_μ - середнє значення спектру, v_σ - середньовадратичне відхилення спектру.

2.4 Поліномне наближення

Створення поліномного наближення має два основних етапи: 1) знаходження прихованих піків та 2) власне поліномне наближення. Через недостатню роздільну здатність спектрометра та/або специфічність будови досліджуваної речовини деякі піки накладаються один на одну, що ускладнює їх аналіз. Прикладом цього може бути Амід I, який складається з піків, що відповідають сигналам розтягування C=O різних вторинних структур білків у зразку (Рис. 2.4). Для знаходження прихованих піків використовують дві основні методики: деконволюцію та похідна спектроскопія. Останній метод передбачає взяття 2 (чи 4) похідної досліджуваного спектру. Перша похідна відображає швидкість зміни інтенсивності відносно хвильового числа, а друга містить інформацію про зміну цієї швидкості (нахилу) спектра. Наявність прихованих піків впливає на швидкісь зміни нахилу функції. Також мінімум на другій похідній відповідає локальному максимуму на оригінальному спектрі. Оскільки взятя похідної ампліфікує шуми, диференціювання спектрів було виконане з використанням SG фільтра, який забезпечує попереднє згладжування.

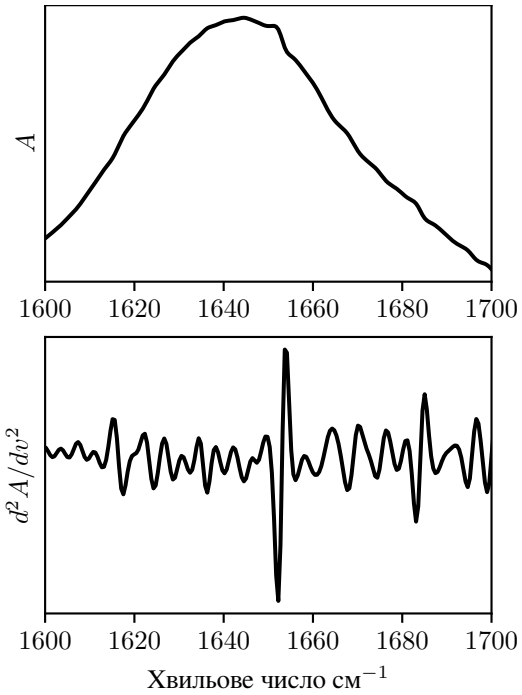


Рис. 2.4 (зверху) Амід I FTIR спектру крові та (знизу) його друга похідна. A - поглинання.

Для поліноміального наближення була використана псевдофойгтівська функція, яка є наближенням профілю Фойгта і представляє собою лінійну комбінацію кривих Гауса та Лоренца. Використання цієї функції було обумовлене тим, що у JAX (на основі якого написаний jaxfit) не існує прямої реалізації функції Фойгта, такої як у scipy.

Функція Гауса має наступну формулу:

$$G(x, x_0, \Gamma) = a_G e^{-b_G(x-x_0)^2} \tag{2.20}$$

де $a_G = \frac{2}{\Gamma}\sqrt{\frac{\ln 2}{\pi}}$, $b_G = \frac{4\ln 2}{\Gamma^2}$. Функція Лоренца:

$$L(x, x_0, \Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(x-x_0)^2 + (\Gamma/2)^2} \tag{2.21}$$

Тут x_0 - середнє значення розподілу, Γ - ширина на піввисоті (FWHM). Остання відображає ширину спектральної кривої, виміряною між двома точками, які лежать на половині максимальної амплітуди функції. Оригінальний спектр був апроксимований сумою псевдофойгтівських функцій (по одній функції на кожен знайдений пік):

$$V(x, x_0, A, \Gamma_G, \Gamma_L, \eta) = A\left[\eta G(x, x_0, \Gamma_G) + (1-\eta)L(x, x_0, \Gamma_L)\right] \tag{2.22}$$

Тут A - амплітуда, Γ_G - FWHM функції Гауса, а Γ_L - FWHM функції Лоренца, $0 < \eta < 1$ - параметр лінійної комбінації двох функцій.

```python
import jax.numpy as jnp

def combined_pseudo_voigt(x, *params):
    N = len(params) // 5
    result = jnp.zeros_like(x)
    for i in range(N):
        mu, gamma_gaussian, gamma_lorentzian, amplitude, eta =
        ↪  params[i*5:(i+1)*5]

        a_G = (2 / gamma_gaussian) * jnp.sqrt(jnp.log(2) / jnp.pi)
        b_G = (4 * jnp.log(2)) / (gamma_gaussian**2)
        gaussian_term = a_G * jnp.exp(-b_G * (x - mu)**2)

        lorentzian_term = (1 / jnp.pi) * ( (gamma_lorentzian / 2) / ( (x -
        ↪  mu)**2 + (gamma_lorentzian / 2)**2) )

        result += amplitude * (eta * gaussian_term + (1 - eta) *
        ↪  lorentzian_term)
    return result
```

Код 2.1 Python функція для псевдофойгтівського профілю використовуючи JAX.

Якщо η прямує до 1 то вихідна крива буде схожа на функцію Гауса, якщо до 0 – на Лоренца. Реалізована псевдофойгтівська функція для наближення спектру у Python має наступний вигляд (Код 2.1).

2.5 Програмне забезпечення

Робота була виконана у **Python** версії 3.11.4. Регресійна модель була реалізована за допомогою модуля **scikit-learn** [30], статистичний аналіз та зглажування була виконана за допомогою модуля **scipy** [31]. Поліномне наближення було створене за допомогою модуля **jaxfit** [32], який обчислює результат використовуючи прискорення за допомогою GPU/TPU. Також регресійна модель була створена за допомогою H2O AutoML [33]. Код цієї роботи є у відкритому доступі за наступним посиланням:
https://github.com/MelnychenkoM/hba1c_prediction.

Розділ 3 Результати

3.1 PLSR модель

Методи, описані у Розділі 2.3, були застосовані для препроцесингу спектрів. Використані параметри наведені в Таблиці 3.1. На Рис. 3.1 показано середній спектр до і після препроцесингу.

Для створення регресійної моделі усі дані спочатку були розділені на калібрувальний та валідаційний датасети, де останній містив 30% від загальної кількості зразків. Оскільки значення глікованого гемоглобіну не розподілені рівномірно, вони були розбиті на 5 категорій, які були використані для стратифікації під час розділення на датасети. Калібрувальний датасет використовувався для тренування моделі, а валідаційний – для перевірки результатів на даних, що не були використані під час тренування моделі. Визначення оптимальної кількості компонент є ключовим моментом реалізації PLSR моделі, оскільки під час додавання нових латентних змінних з'являється ризик перенавчання. Для перевірки моделі у випадку обмеженої кількості даних зазвичай використову-

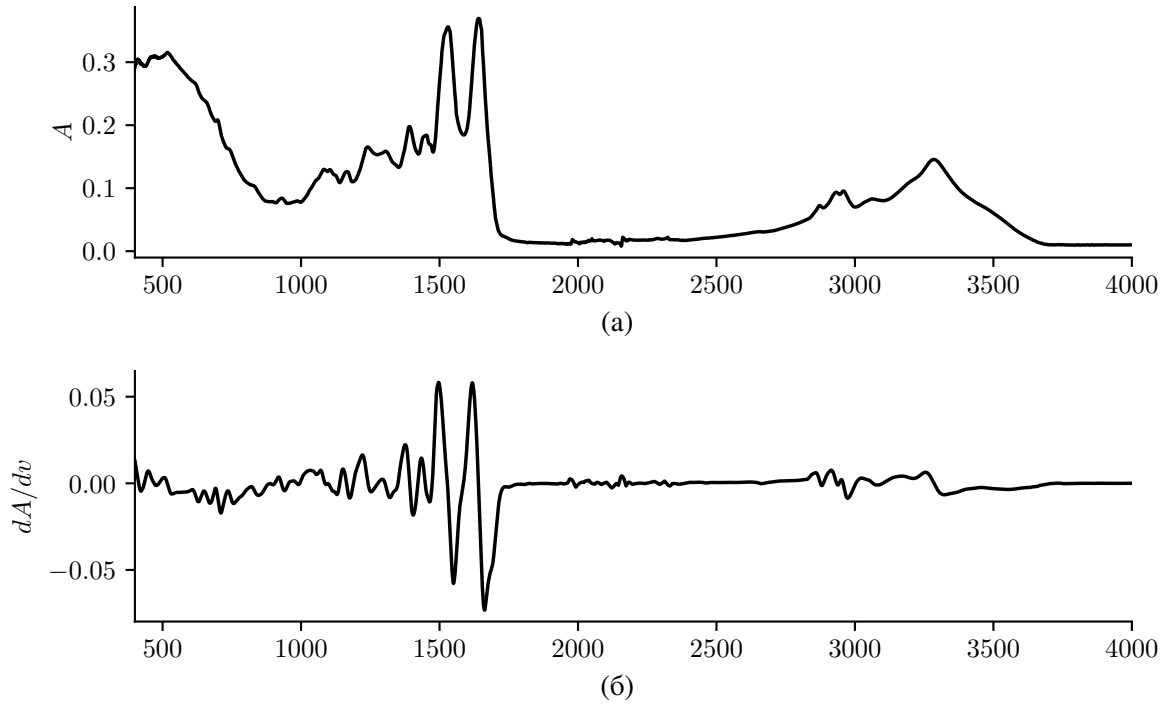


Рис. 3.1 Середній спектр до (а) та після (б) препроцесингу з використанням методів наведених у Табл. 3.1. Тут A - поглинання.

Табл. 3.1 Параметри препроцесингу

Препроцесинг	**Метод**	**Оптимальні параметри**
Зглажування	SG	$w = 55$, $o = 2$, deriv $= 1$
Нормалізація	Векторна	–

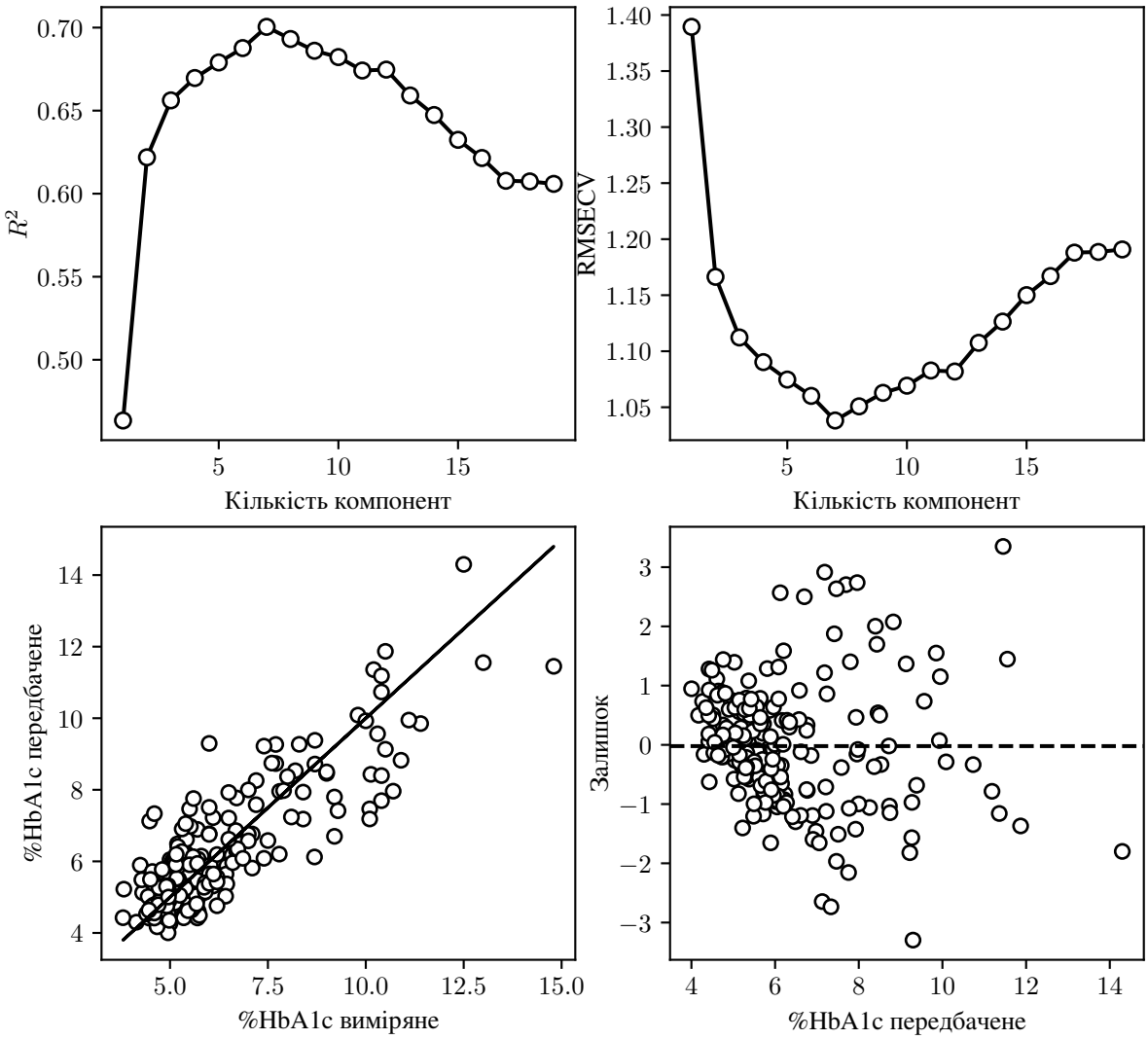


Рис. 3.2 Результат тренування калібраційних даних (CV_7). Верхій ряд - залежність кофіцієнту детермінації та RMSECV від кількості компонент, нижній - графік виміряного HbA1c від передбаченого (зліва) та залишки після передбачення (зправа).

ють перехресне затверджування (cross-validation, далі CV) CV_k, де калібраційні дані розбиваються на k частин, кожна з яких почергово використовується як тестувальні дані, а усі інші – тренувальні. Найчастіше k обирають у діапазоні від

5 до 10. Після перехресного затвердження передбачені дані для кожної частинин були зібрані та була розрахована середньоквадратична похибка перехресної перевірки RMSECV, яка обчислюється наступним чином:

$$\text{RMSECV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

де y_i – вміст HbA1c, що спостерігався, а $\hat{y}_i$ – передбачений вміст кожної розбитої групи.

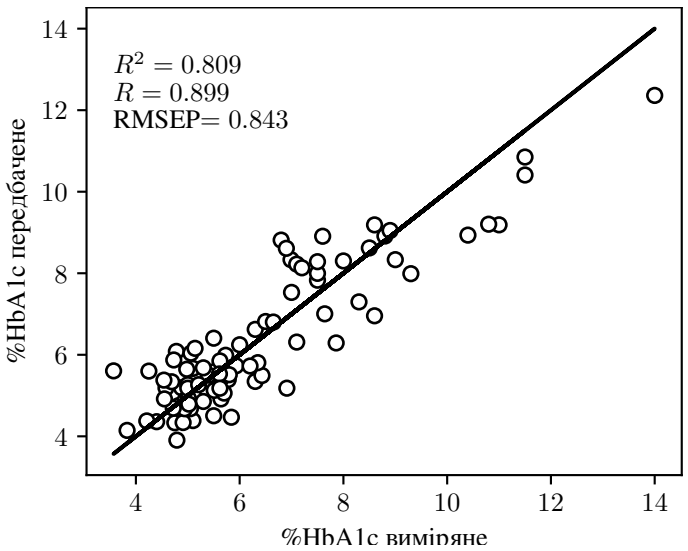


Рис. 3.3 Передбачення валідаційних даних. Кількість компонент: 6.

Менші значення RMSECV відповідають кращій моделі. Оскільки кожен спектр містить велику кількість незалежних змінних (приблизно 7000), багато з них не є інформативними і у деяких випадках вони можуть погіршувати роботу моделі. Враховуючи, що регіон від 800 до 1800 см$^{-1}$ містить більшість біологічної інформації (bio-finterprint region у літературі) він був використаний для створення цієї моделі. Результати тренування калібрувального датасету з 7 розділеннями (CV_7) наведені на Рис. 3.2. Згідно з перехресним схрещуванням оптимальна кількість компонент становила 6. Ці компоненти були використані для передбачення валідаційних даних. Отримані результати становили $R^2 = 0.809$, $R = 0.899$ та RMSEP $= 0.843\%$, де R^2 - коефіцієнт детермінації, R - коефіцієнт Пірсона, RMSEP - середньоквадратична похибка передбачення.

3.2 Поліномне наближення спектрів

Для попередньої обробки використовувався SG фільтр з параметрами ($w = 5$, $o = 3$), а також гумкова корекція базової лінії (rubberband baseline correction) з модуля BoxSERS (https://github.com/ALebrun-108/BoxSERS). Спектри були

нормалізовані відповідно до інтенсивності піку Аміда I. Приховані піки були знайдені на другій похідній спектрів, які були отримані за допомогою SG фільтра. Використовуючи псевдофогтівську функцію було виконане наближення двох регіонів: 1) 1000-1700 см$^{-1}$ та 2) 1600-1700 см$^{-1}$ (Амід I). Перший

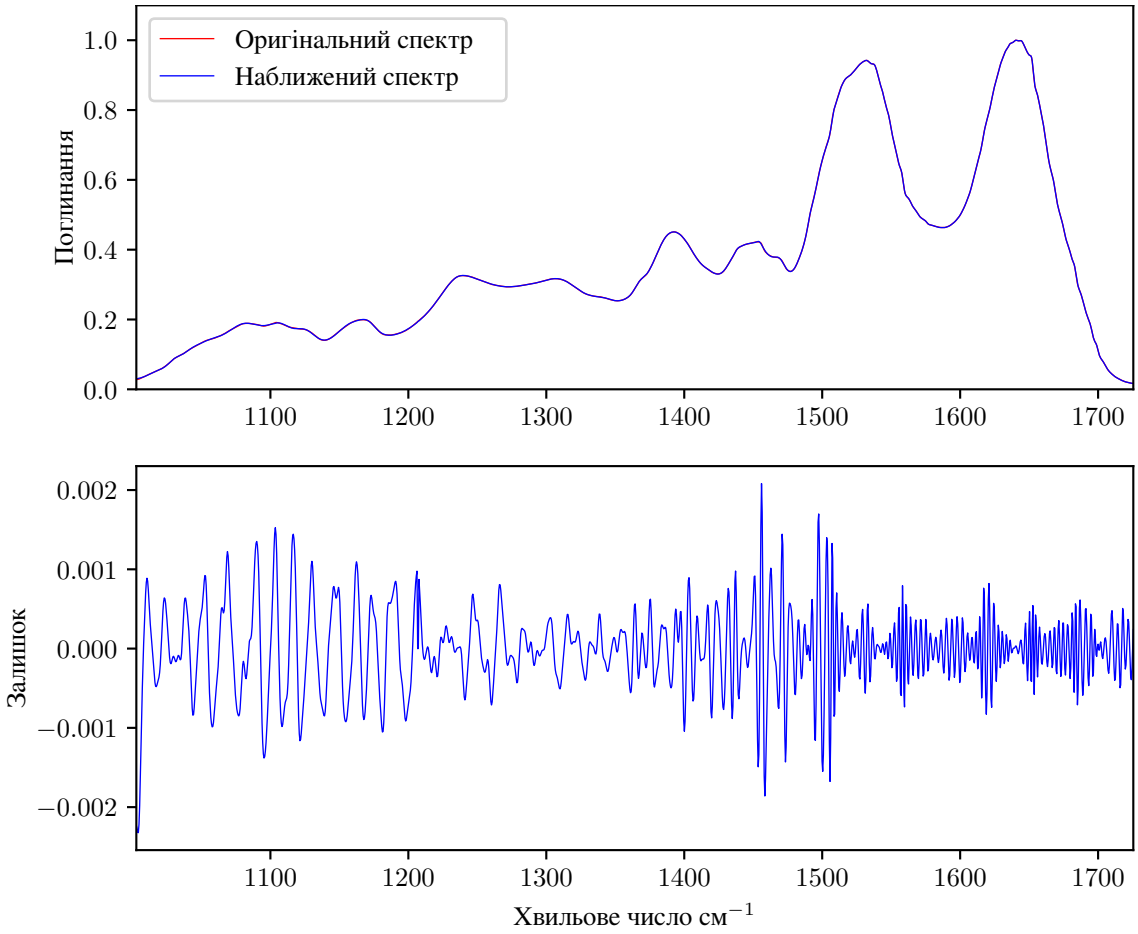


Рис. 3.4 Результати наближення одного спектру (зверху) та його залишки (знизу).

регіон був обраний, оскільки він містить інформацію про білкові структури та значну частину даних про вуглеводи. Крім того, у цьому діапазоні досить легко виявляти приховані піки за допомогою диференціювання. Усі наближені спектри мали коефіцієнт детермінації $R^2 > 0.999$ та середню залишкову суму квадратів приблизно $SS_1 < 0.001$ для першого регіону та $SS_2 < 0.0001$ для другого. На Рис. 3.4 наведено результати одного наближення та його залишки. FTIR дозволяє отримати доступ до інформації про вторинну структуру досліджуваного зразку і це можна побачити на розкладеному спектрів після 3.5).

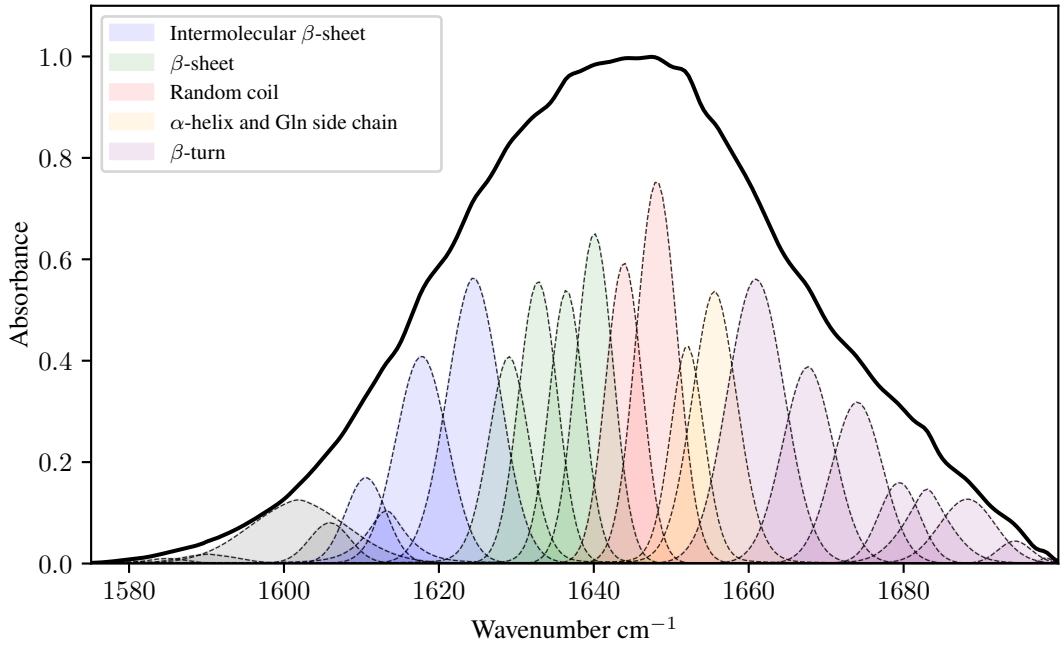
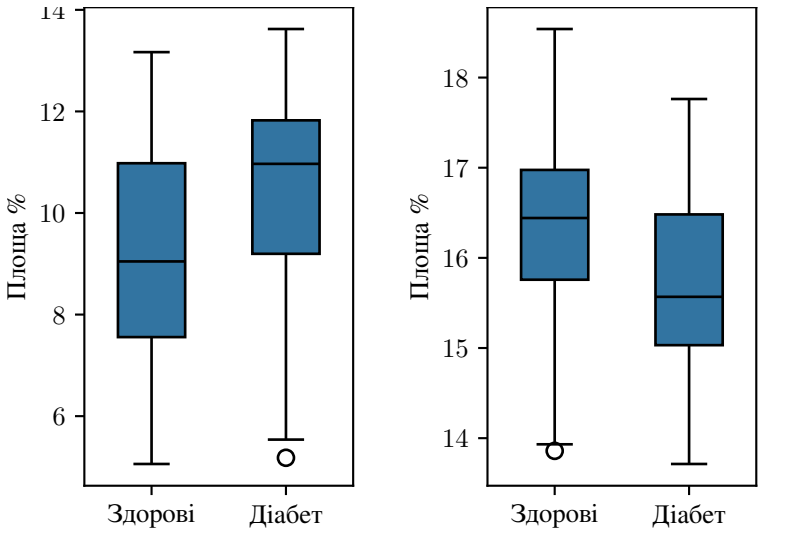


Рис. 3.5 Поліномне наближення Аміду I з наведеними вторинними структурами. Відповідність піків до вторинних струткур адаптовано з [34].

поліномного наближення (Рис. 3.5). Наприклад, піки на проміжку 1650-1659 см$^{-1}$ відповідають α-спіралі та бічному залишку гліцину, 1628-1642 см$^{-1}$ – β-листам, а 1610-1627 см$^{-1}$ – інтермолекулярним β-листам. Проведене наближення дозволяє знайти відносну площу під кривою кожного з піків та подивитись на її зміну в залежності від різних патологічних станів людини. Метод поліномного наближення для кількісного встановлення вторинних структур вже був описаний для потенційного діагностування лейкемії [18] з відмінністю, що у зазначеній статті використовувався розподіл Гауса. Якщо встановити діагностичний поріг



(а) α-спіраль (б) Інтермолекулярний β-лист

Рис. 3.6 Площа вторинних структур Амід I піку.

діабету на рівні HbA1c 6.5%, то при розподілі учасників на здорових та хворих можна спостерігати збільшення площі (і, відповідно, вмісту) α-спіралей з 9.25% до 10.42% ($p < 0.0001$) та зменшення вмісту інтермолекулярних β-листів з 16.34% до 15.71% ($p < 0.0001$) у здорових та хворих на діабет відповідно (Рис. 3.6). Статистичну значущість змін площ у двох групах було визначено за допомогою тесту Манна-Уітні. Попри виявлену різницю у середніх площах вторинних структур у FTIR спектрах крові між двома групами, спроба створити класифікаційну модель для передбачення діабету за допомогою даних про вторинну структуру не привела до точних результатів.

Для другого регіону (1000-1700 см$^{-1}$) на другій похідній було знайдено 67 різних піків, кожен з яких був апроксимований псевдофойгтівською функцією. Наближення другого регіону та розклад на утворюючі піки можна побачити на Рис. 3.7. Кожен пік мав 5 основних параметрів, які були описані в Розділі 2.4. Додатково, для кожного піку були визначені його FWHM та відносна площа відносно загальної площі після нормалізації. Для побудови регресійної моделі на основі визначених параметрів був використаний пакет H2O AutoML [33]. Цей інструмент

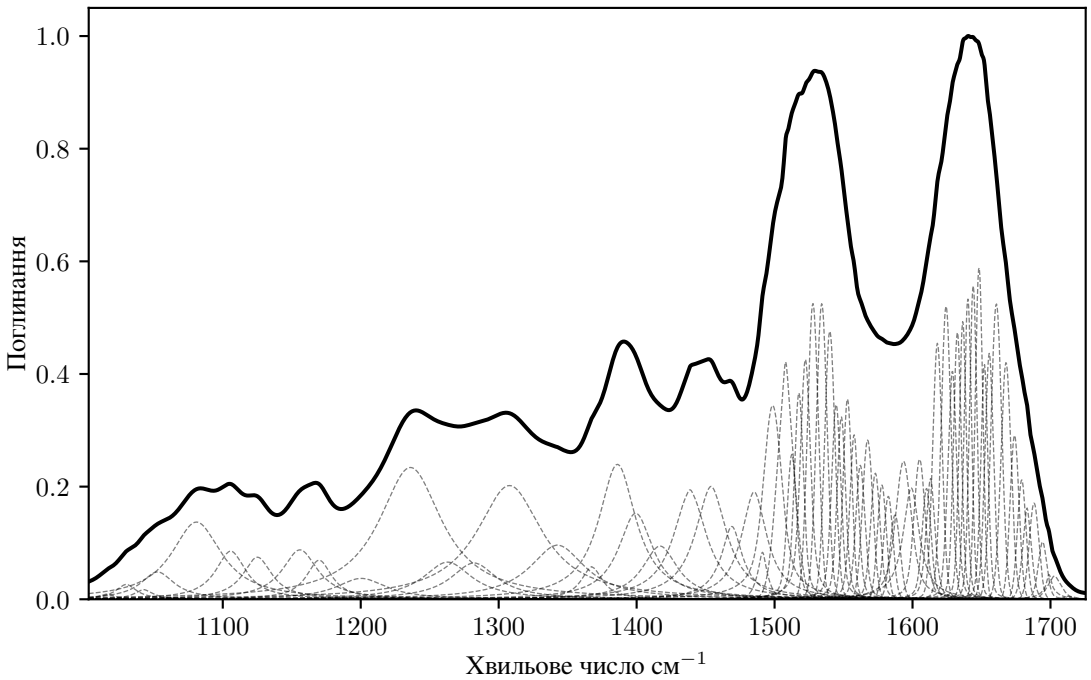


Рис. 3.7 Розклад 1000-1700 см$^{-1}$ регіону спектру на індивідуальні піки після поліномного наближення.

дозволяє автоматизувати вибір оптимальної моделі та параметрів для передбачення. Крім того, він надає можливість оцінити важливість змінних, що дозволяє зробити висновок про значущість різних піків для передбачення глікованого гемоглобіну. Найкращі моделі у H2O показали RMSE на рівні 1.25% і R^2 0.5. Оскільки це були ансамблеві моделі з різними регресорами, для визначення важливості змінних була використана узагальнена лінійна модель (GLM) під назвою **GLM 1 AutoML 7**. Прогнози цієї моделі на валідаційних даних можна побачити на Рис. 3.8. На Рис. 3.9 представлені значення важливості для десяти найкращих змінних. Виявилося, що найбільш важливими є піки, що відповідають за вміст вуглеводів. Наприклад, пік при 1012 см$^{-1}$ відповідає за 2-дезокси-D-рибозу, а 1029 см$^{-1}$ відповідає за полісахариди та, зокрема, глюкозу [35]. Третім за зна-

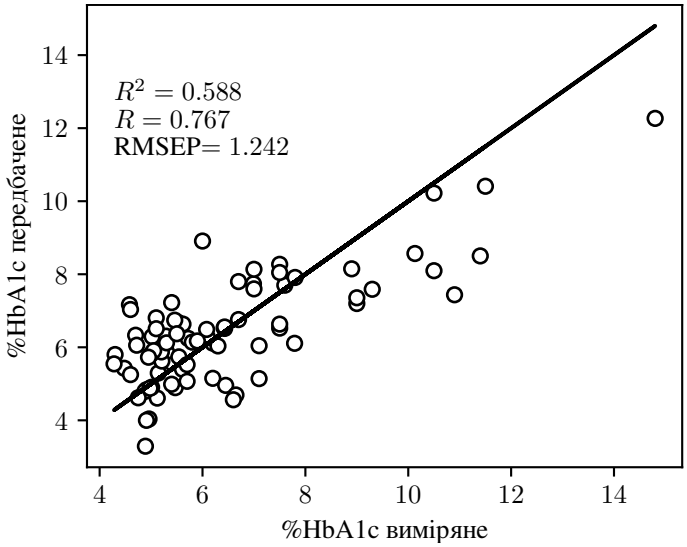


Рис. 3.8 Передбачення валідаційних даних за допомогою GLM моделі з H2O.

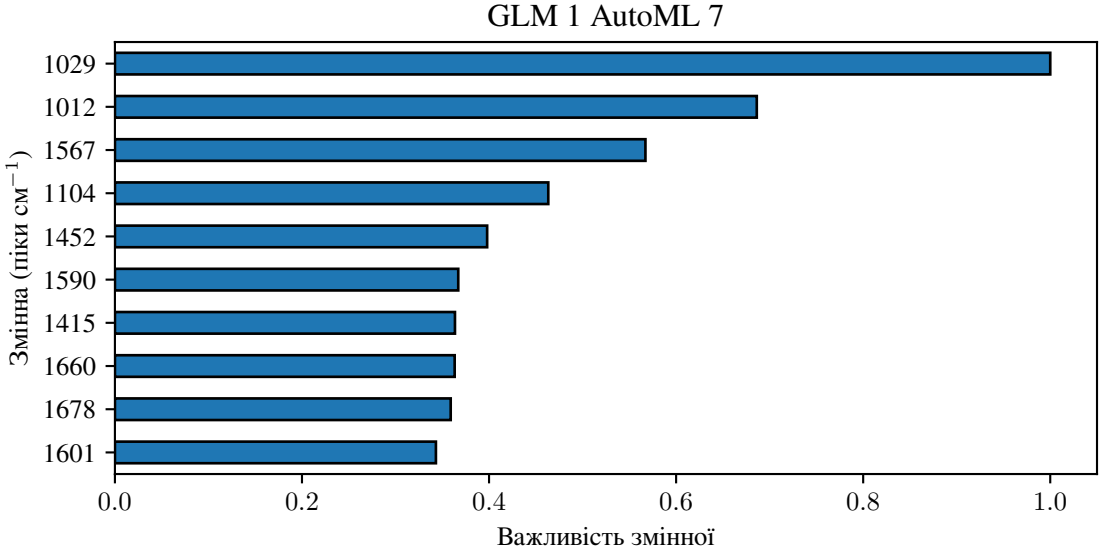
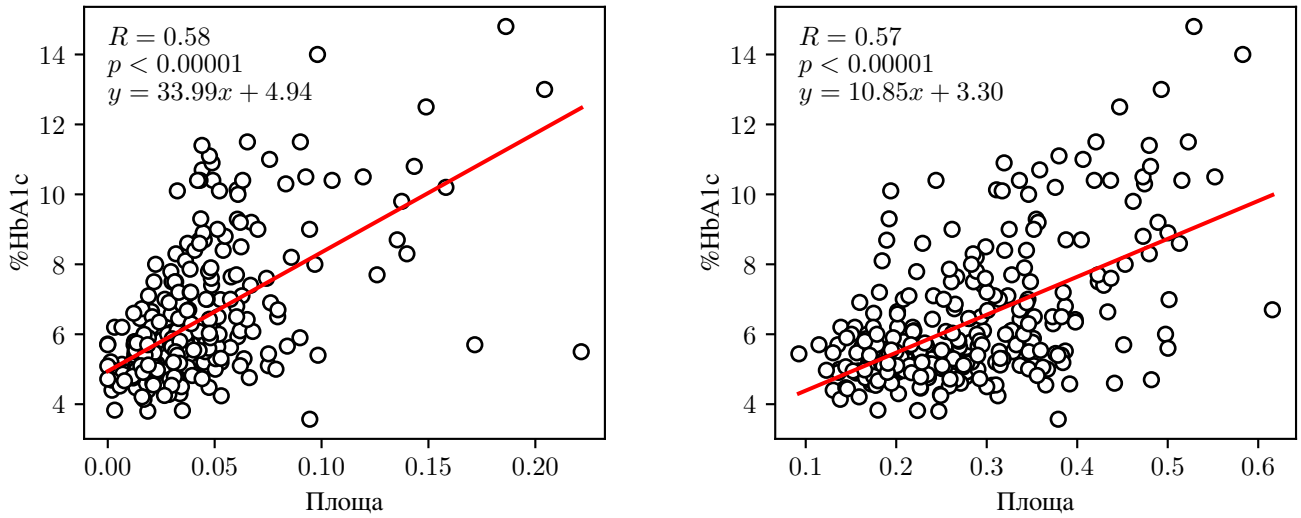


Рис. 3.9 Важливість змінних для передбачення HbA1c.



(а) Площа піку на 1012 см$^{-1}$ (б) Площа піку на 1029 см$^{-1}$

Рис. 3.10 Зв'язок між рівнем HbA1c та площею піків

чимістю виявився пік на 1567 см$^{-1}$, який характеризує різні стани гемоглобіну, у тому числі карбогемоглобіну і деоксигемоглобіну (Табл. 1.1). Пік на 1104 см$^{-1}$ вказує на полісахариди та глікоген, а пік на 1452 см$^{-1}$ - на ліпіди та білки [17]. Графіки залежності рівня глікованого гемоглобіну від площі найважливіших піків (згідно з Рис. 3.9) наведено на Рис. 3.10.

Висновки

У даній роботі була створена PLSR модель для передбачення рівня глікованого гемоглобіну використовуючи FTIR спектри крові. Результати прогнозування валідаційних даних після препроцесингу на проміжку 800-1800 см$^{-1}$ становили: $R^2 = 0.809$, $R = 0.899$, RMSEP $= 0.843\%$. Ця модель демонструє точність, яка наближається до раніше опублікованих результатів [24].

Була перевірена можливість використання поліномного наближення псевдо-фойгтівською функцією для параметризації піків, що утворюють ці спектри. Після препроцесингу спектрів за допомогою корекції базової лінії та SG фільтра було виконане наближення оригінального спектра. Аналіз площ вторинних структур після розділення датасету на здорових та хворих на діабет згідно з критерієм HbA1c 6.5% за допомогою тесту Манна-Уітні виявив зміни в площах піків, що відповідають різним вторинним структурам білків. Зокрема, площа піків, пов'язаних з α-спіралями, збільшилася з 9.25% у здорових осіб до 10.42% у пацієнтів з діабетом ($p < 0.0001$). Водночас, площа піків інтермолекулярних β-листів зменшилася з 16.34% до 15.71% ($p < 0.0001$). Для діагностування діабету використовували критерій рівня HbA1c 6.5%. Виявлена різниця не була достатньою для створення точної класифікаційної моделі використовуючи лише площу піків, що відповідають вторинним структурам білків.

Поліномне наближення 1000-1700 см$^{-1}$ регіону дало змогу параметризувати 67 піків та знайти їх площу.Ці дані були використані для побудови регресійної моделі за допомогою H2O AutoML. Для забезпечення інтерпретації результатів було обрано GLM модель, яка демонструвала дещо меншу точність у порівнянні з PLSR ($R^2 = 0.588$, $R = 0.767$, RMSEP $= 1.242\%$), проте було знайдено піки, що мали сильну асоціацію з рівнями глікованого гемоглобіну. Ця інформація може бути використана для подальшого дослідження молекулярних сигнатур діабету.

Менша точність GLM моделі для передбачення HbA1c з використанням площ може бути пов'язана з втратою інформації з інших діапазонів, наприклад, 600-1000 см$^{-1}$, який також містить піки, що відповідають різним станам гемоглобіну та іншим вуглеводам. Цей діапазон характеризується вищим рівнем

шуму, що ускладнює виявлення прихованих піків за допомогою диференціювання. Тому потрібно розглянути можливість використання альтернативних методів диференціювання для виявлення прихованих піків, окрім звичайного SG фільтру. Другою причиною недостатньої точності отриманої моделі може бути специфіка досліджуваного параметра крові, HbA1c, який є відношенням вмісту глікованого гемоглобіну до загального вмісту гемоглобіну. Це може ускладнювати передбачення, оскільки зразки з однаковим рівнем HbA1c можуть мати різний загальний вміст гемоглобіну.

Хоча передбачення з використанням площ піків не досягли бажаної високої точності, вдалося розкласти оригінальний спектр на окремі піки та визначити важливість змінних для прогнозування рівня HbA1c. Найзначущішими виявилися піки на 1012 $см^{-1}$ та 1029 $см^{-1}$, які характеризують вуглеводи, зокрема глюкозу. Інший ключовий пік розташований на 1567 $см^{-1}$, і він відповідає різним станам гемоглобіну, зокрема карбогемоглобіну та деоксигемоглобіну (Табл 1.1).

Джерела

[1] Jose Miguel Baena-Díez, Judit Peñafiel, Isaac Subirana, Rafel Ramos, Roberto Elosua, Alejandro Marín-Ibañez, María Jesús Guembe, Fernando Rigo, María José Tormo-Díaz, Conchi Moreno-Iribas, Joan Josep Cabré, Antonio Segura, Manel García-Lareo, Agustín Gómez de la Cámara, José Lapetra, Miquel Quesada, Jaume Marrugat, Maria José Medrano, Jesús Berjón, Guiem Frontera, Diana Gavrila, Aurelio Barricarte, Josep Basora, Jose María García, Natalia C. Pavone, David Lora-Pablos, Eduardo Mayoral, Josep Franch, Manel Mata, Conxa Castell, Albert Frances, María Grau, and on behalf of the FRESCO Investigators. Risk of Cause-Specific Death in Individuals With Diabetes: A Competing Risks Analysis. *Diabetes Care*, 39(11):1987–1995, 08 2016.

[2] Nam H Cho, Jonathan E Shaw, Shakoor Karuranga, Yutong Huang, Joao D da Rocha Fernandes, Andrew W Ohlrogge, and Belma Malanda. Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138:271–281, 2018.

[3] Parisa Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Shakoor Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, Jonathan E Shaw, David Bright, Rhys Williams, and IDF Diabetes Atlas Committee. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes research and clinical practice*, 157:107843, 2019.

[4] American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*, 44(Supplement_1):S15–S33, January 2021.

[5] Uma Krishnamurti and Michael W Steffes. Glycohemoglobin: a primary predictor of the development or reversal of complications of diabetes mellitus. *Clinical chemistry*, 47(7):1157–1165, 2001.

[6] Andri Ioannou and Constantinos Varotsis. Modifications of hemoglobin and myoglobin by maillard reaction products (mrps). *PloS one*, 12(11):e0188095, 2017.

[7] Curt L. Rohlfing, Hsiao-Mei Wiedmeyer, Randie R. Little, Jack D. England, Alethea Tennill, and David E. Goldstein. Defining the relationship between plasma glucose and hba1c: Analysis of glucose profiles and hba1c in the diabetes control and complications trial. *Diabetes Care*, 25(2):275–278, 2002.

[8] Catherine C. Cowie, Keith F. Rust, Danita D. Byrd-Holt, Edward W. Gregg, Earl S. Ford, Linda S. Geiss, Kathleen E. Bainbridge, and Judith E. Fradkin. Prevalence of Diabetes and High Risk for Diabetes Using A1C Criteria in the U.S. Population in 1988–2006. *Diabetes Care*, 33(3):562–568, 01 2010.

[9] Daniel W. Belsky, Avshalom Caspi, Renate Houts, Harvey J. Cohen, David L. Corcoran, Andrea Danese, HonaLee Harrington, Salomon Israel, Morgan E. Levine, Jonathan D. Schaefer, Karen Sugden, Ben Williams, Anatoli I. Yashin, Richie Poulton, and Terrie E. Moffitt. Quantification of biological aging in young adults. *Proceedings of the National Academy of Sciences*, 112(30):E4104–E4110, 2015.

[10] Nathaniel Dubowitz, Wei Xue, Qi Long, James G Ownby, Daniel E Olson, Diana Barb, Mary K Rhee, Anand V Mohan, Patricia I Watson-Williams, Sharon L Jackson, Anne M Tomolo, Tracy M 2nd Johnson, and Lawrence S Phillips. Aging is associated with increased HbA1c levels, independently of glucose levels and insulin resistance, and also with decreased HbA1c diagnostic specificity. *Diabetic medicine : a journal of the British Diabetic Association*, 31(8):927–935, 2014.

[11] Jasper W.L. Hartog, Adriaan A. Voors, Stephan J.L. Bakker, Andries J. Smit, and Dirk J. van Veldhuisen. Advanced glycation end-products (AGEs) and heart failure: Pathophysiology and clinical implications. *European Journal of Heart Failure*, 9(12):1146–1155, 2007.

[12] Liliane J. Striker and Gary E. Striker. Administration of AGEs in vivo induces extracellular matrix gene expression. *Nephrology Dialysis Transplantation*, 11(supp5):62–65, 01 1996.

[13] Ralica Petrova, Yasuhiko Yamamoto, Katsuhiko Muraki, Hideto Yonekura, Shigeru Sakurai, Takuo Watanabe, Hui Li, Masayoshi Takeuchi, Zenji Makita, Ichiro Kato, Shin Takasawa, Hiroshi Okamoto, Yuji Imaizumi, and Hiroshi Yamamoto. Advanced Glycation Endproduct-Induced Calcium Handling Impairment in Mouse Cardiac Myocytes. *Journal of Molecular and Cellular Cardiology*, 34(10):1425–1431, 2002.

[14] Zainab Hegab, Steven Gibbons, Ludwig Neyses, and Mamas A Mamas. Role of advanced glycation end products in cardiovascular disease. *World Journal of Cardiology*, 4(4):90–102, 2012.

[15] Sho ichi Yamagishi. Role of advanced glycation end products (AGEs) and receptor for AGEs (RAGE) in vascular damage in diabetes. *Experimental Gerontology*, 46(4):217–224, 2011.

[16] B. C. Smith. *Fundamentals of Fourier Transform Infrared Spectroscopy*. CRC Press, 2nd edition, 2011.

[17] T. Makhnii, O. Ilchenko, A. Reynt, Y. Pilgun, A. Kutsyk, D. Krasnenkov, M. Ivasyuk, and V. Kukharskyy. Age-Related Changes in FTIR and Raman Spectra of Human Blood. *Ukrainian Journal of Physics*, 61(10):853, Jan. 2019.

[18] Rados law Chaber, Aneta Kowal, Pawe l Jakubczyk, Christopher Arthur, Kornelia Lach, Renata Wojnarowska-Nowak, Krzysztof Kusz, Izabela Zawlik, Sylwia Paszek, and Józef Cebulski. A Preliminary Study of FTIR Spectroscopy as a Potential Non-Invasive Screening Tool for Pediatric Precursor B Lymphoblastic Leukemia. *Molecules*, 26(4), 2021.

[19] L. G. Silva, A. F. S. Péres, D. L. D. Freitas, et al. ATR-FTIR Spectroscopy in Blood Plasma Combined with Multivariate Analysis to Detect HIV Infection in Pregnant Women. *Sci Rep*, 10:20156, 2020.

[20] P. Guang, W. Huang, L. Guo, X. Yang, F. Huang, M. Yang, W. Wen, and L. Li. Blood-based ftir-atr spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: A stard compliant diagnosis research. *Medicine*, 99(15):e19657, 2020.

[21] Douglas Carvalho Caixeta, Murillo Guimarães Carneiro, Ricardo Rodrigues, Deborah Cristina Teixeira Alves, Luís Ricardo Goulart, Thúlio Marquez Cunha, Foued Salmen Espindola, Rui Vitorino, and Robinson Sabino-Silva. Salivary atr-ftir spectroscopy coupled with support vector machine classification for screening of type 2 diabetes mellitus. *Diagnostics*, 13(8), 2023.

[22] Satoshi Yoshida, Makoto Yoshida, Mayumi Yamamoto, and Jun Takeda. Optical screening of diabetes mellitus using non-invasive Fourier-transform infrared spectroscopy technique for human lip. *Journal of Pharmaceutical and Biomedical Analysis*, 76:169–176, 2013.

[23] Ieva Jurgeleviciene, Daiva Stanislovaitiene, Vacis Tatarunas, Marius Jurgelevicius, and Dalia Zaliuniene. Assessment of absorption of glycated nail proteins in patients with diabetes mellitus and diabetic retinopathy. *Medicina*, 56(12), 2020.

[24] Yun Han, Tao Pan, Huihui Zhou, and Rui Yuan. ATR-FTIR spectroscopy with equidistant combination PLS method applied for rapid determination of glycated hemoglobin. *Anal. Methods*, 10:3455–3461, 2018.

[25] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. PLS Methods.

[26] Hervé Abdi. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics*, 2(1):97–106, 2010.

[27] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In Craig Saunders, Marko Grobelnik, Steve Gunn, and John

Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, pages 34–51, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[28] Agnar Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[29] Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[32] Lucas R Hofer, Milan Krstajić, and Robert P Smith. Jaxfit: Trust region method for nonlinear least-squares curve fitting on the GPU. *arXiv preprint arXiv:2208.12187*, 2022.

[33] Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020.

[34] Aichun Dong, Ping Huang, and Winslow S. Caughey. Protein secondary structures in water from second-derivative amide i infrared spectra. *Biochemistry*, 29(13):3303–3308, 1990. PMID: 2159334.

[35] E. Wiercigroch, E. Szafraniec, K. Czamara, M. Z. Pacia, K. Majzner, K. Kochan, A. Kaczor, M. Baranska, and K. Malek. Raman and infrared spectroscopy of carbohydrates: A review. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 185:317–335, 2017.