UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Komplexní workflow pro analýzu omických dat v Pythonu



Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **prof. RNDr. Karel Hron, Ph.D., DSc.**
Vypracovala: **Bc. Nikola Urbánková**
Studijní program: 1103T004 Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2026

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Nikola Urbánková

Název práce: Komplexní workflow pro analýzu omických dat v Pythonu

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: prof. RNDr. Karel Hron, Ph.D., DSc.

Rok obhajoby práce: 2026

Abstrakt: Cílem práce bylo vytvořit komplexní analytické workflow v jazyce Python určené pro zpracování omických, především metabolomických dat. Nástroj vznikl pro Oddělení klinické biochemie FN Olomouc s cílem zautomatizovat každodenní analýzu vysoce dimenzionálních datových sad. Teoretická část práce detailně popsala pokročilé metody předzpracování dat, jako jsou LOWESS korekce, imputace chybějících hodnot či normalizace, a shrnula využívané postupy jednorozměrné i vícerozměrné statistické analýzy. Praktickým výstupem je uživatelsky přívětivý program, úspěšně otestovaný na reálných vzorcích včelí hemolymfy. Je možné takto datové soubory jsou spolehlivě zpracovat a automaticky vygenerovat kompletní textové i grafické reporty v HTML, PDF a MS Excel.

Klíčová slova: QC vzorky, vážená lokální regrese, korekce p-hodnot, testování hypotéz

Počet stran: 78

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Nikola Urbánková

Title: Comprehensive workflow for the analysis of omics data in Python

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: prof. RNDr. Karel Hron, Ph.D., DSc.

The year of presentation: 2026

Abstract: The aim of this thesis was to develop a comprehensive analytical workflow in Python designed for processing omics, primarily metabolomics, data. The tool was created for the Department of Clinical Biochemistry at the University Hospital Olomouc to automate the daily analysis of high-dimensional datasets. The theoretical part of the thesis detailed advanced data preprocessing methods, such as LOWESS correction, missing value imputation, and normalization, and summarized the applied procedures of univariate and multivariate statistical analysis. The practical outcome is a user-friendly program successfully tested on real samples of bee hemolymph. The script reliably processed the input data and automatically generated comprehensive textual and graphical reports in HTML, PDF, and MS Excel formats.

Key words: quality control samples, locally weighted regression, p-values adjustment, testing of hypothesis

Number of pages: 78

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana prof. RNDr. Karla Hrona, Ph.D., DSc. a všechny použité zdroje jsem uvedla v seznamu literatury. Pro stylistické úpravy textu a výpomoc s grafickými výstupy byl využit nástroj Google Gemini, a to nikoli jako náhrada tvůrčí činnosti.

V Olomouci dne ....

podpis

Obsah

Úvod **7**

1 Popis dat **9**
1.1 Omická data . . . 9
1.2 QC vzorky . . . 9
1.3 Reálná data . . . 11

2 Předzpracování dat **12**
2.1 LOWESS korekce . . . 13
2.1.1 Metoda robustní LOWESS . . . 14
2.1.2 Specifikace parametrů metody LOWESS v Pythonu . . . 18
2.2 Kritéria výběru metabolitů . . . 22
2.2.1 Variační koeficient . . . 22
2.2.2 Odchylka intenzit v QC vzorcích . . . 23
2.3 Imputace chybějících hodnot . . . 24
2.3.1 2/3 min . . . 24
2.3.2 Metoda k nejbližších sousedů . . . 25
2.3.3 Random Forest . . . 27
2.4 Logaritmická transformace . . . 31
2.5 Normalizace PQN . . . 32
2.6 Škálování dat . . . 34

3 Jednorozměrná statistika **37**
3.1 Test normality . . . 37
3.2 Odlehlé hodnoty . . . 38
3.2.1 Grubbsův test . . . 39
3.2.2 Mezikvartilové rozpětí . . . 41
3.2.3 Modifikované Z-skóre . . . 41
3.3 Charakteristiky . . . 43
3.4 Korekce . . . 43
3.4.1 Bonferroniho metoda . . . 45
3.4.2 Benjamini-Hochbergova metoda . . . 45
3.5 Parametrické testování hypotéz . . . 47
3.5.1 Dvouvýběrový t-test . . . 48
3.5.2 Welchův dvouvýběrový t-test . . . 48
3.5.3 Párový t-test . . . 48
3.5.4 ANOVA . . . 49
3.5.5 ANCOVA . . . 50
3.6 Neparametrické testování hypotéz . . . 55

3.6.1 Mann-Whitneyův U test 55
3.6.2 Wilcoxonův test 56
3.6.3 Kruskal-Wallisův test 57
3.7 Výsledky testování hypotéz 58
3.8 Korelační analýza 58
3.9 ROC analýza 61
3.10 Volcano plot 65

4 Vícerozměrná statistika **68**
4.1 PCA 68
4.2 Shlukování 70

5 Výsledky analýzy **73**

Závěr **74**

Literatura **75**

Poděkování

Ráda bych poděkovala svému vedoucímu práce panu prof. RNDr. Karlu Hronovi, Ph.D., DSc. za spolupráci, ochotu a čas, který mi věnoval při konzultacích. Také bych chtěla poděkovat panu prof. RNDr. Davidovi Friedeckému, Ph.D. a jeho kolegům z Fakultní nemocnice Olomouc za poskytnutí metabolomických dat a následnou pomoc při interpretaci výsledků.

Úvod

V současné době se v oblasti analýzy dat těší stále velké oblibě programovací jazyk Python, který nabízí robustní nástroje pro statistické zpracování. Cílem této práce je proto vytvoření komplexního analytického workflow v tomto jazyce, určeného primárně pro pracovníky Oddělení klinické biochemie Fakultní nemocnice Olomouc. Na oddělení pracují s omickými daty, zejména pak s daty metabolomickými a lipidomickými. Záměrem práce je tak usnadnit proces analýzy pro kterákoli zmíněná data a poskytnout intuitivní prostředí i pro uživatele s menšími zkušenostmi s programováním. Důraz je přitom kladen na samotný vývoj workflow a definování metod v kódu, nikoliv na biologickou interpretaci. Nejvyužívanějšími nástroji v Pythonu pro tuto práci jsou knihovny *Pandas* [25], *Statsmodels* [23] a *Scikit-learn* [21]. Ačkoliv jsme funkčnost testovali na více datasetech, pro názornou ukázku v textu slouží pouze jeden konkrétní.

Diplomová práce je strukturována do čtyř hlavních kapitol. První kapitola slouží jako úvod do problematiky omických dat. Druhá, stěžejní kapitola se podrobně věnuje předzpracování dat, které tvoří nezbytný základ pro všechny následné statistické výpočty. Třetí a čtvrtá kapitola se zabývají jednorozměrnou a vícerozměrnou statistickou analýzou. Zatímco standardní metody jsou popsány spíše stručně, detailnější pozornost je věnována pokročilejším přístupům, jako je např. imputace pomocí metody MissForest, analýza kovariance (ANCOVA), ROC analýza a dalším.

Práce není striktně rozdělena na teoretickou a praktickou část; obě roviny se v textu plynule prolínají. Způsob výkladu a použitá matematická notace se však přizpůsobují charakteru dané problematiky. Kapitola věnovaná předzpracování dat je pojata převážně aplikačně, a proto využívá značení

odpovídající realizacím. Naopak kapitoly věnované statistické analýze jsou ukotveny v teoretickém rámci s využitím standardního značení pro náhodné veličiny. Společným prvkem napříč kapitolami je doplnění výkladu o specifikaci funkcí a postupů aplikovaných v jazyce Python.

Reálná data jsou poskytnuta přímo z oddělení a jedná se o vzorky včelí hemolymfy (tělní tekutiny). Vzhledem k vysoké dimenzionalitě analyzovaných dat nejsou všechny výstupy součástí samotného textu. Kompletní výsledky jsou proto k dispozici ve formátu tabulek MS Excel a HTML či PDF reportů v elektronické příloze práce, společně se zdrojovými kódy, které jsou pro přehlednost rozděleny do tří Python skriptů odpovídajících jednotlivým kapitolám.

1 Popis dat

1.1 Omická data

Pojem omická data označuje specifický typ velkých dat („*Big Data*“) v molekulární biologii, který pokrývá různé úrovně od genomiky až po metabolomiku (Hasin et al. [9]). Získávají se komplexní analýzou celých souborů molekul v biologickém systému. Každý typ těchto dat dokáže sám o sobě odhalit odchylky spojené s konkrétním onemocněním. Získané informace slouží jako důležité ukazatele pro sledování vývoje nemoci a zároveň umožňují identifikovat, jaké biologické procesy se u pacientů liší ve srovnání se zdravou kontrolní skupinou.

V rámci biochemie se zaměřujeme primárně na metabolomická data. Tento obor současně kvantifikuje více typů malých molekul, jako jsou aminokyseliny, mastné kyseliny, sacharidy nebo jiné produkty buněčných procesů. Hladiny metabolitů a jejich relativní poměry odrážejí metabolickou funkci, přičemž odchylky od normálního rozmezí často naznačují nemoc (Patti et al [20]). Metabolomika úzce souvisí a navazuje na další typy dat, jako je lipidomika, která se specificky zaměřuje na analýzu lipidů.

1.2 QC vzorky

Kromě samotných vzorků s naměřenými hodnotami metabolitů jsou nedílnou součástí dat i tzv. *QC samples* neboli vzorky kontroly kvality. Tyto vzorky vznikají jako směs všech vzorků dané analýzy. V rámci analytické sekvence jsou QC vzorky zařazovány průběžně, přičemž jejich přesný počet není fixně dán a volí se individuálně dle uvážení analytika. Typicky se zařazují nejprve na prvních několika (např. čtyřech) pozicích pro stabilizaci systému a poté jsou vkládány v pravidelných intervalech, například jako každý desátý vzo-

rek. Vycházíme z předpokladu, že signál těchto QC vzorků musí být v čase stabilní; pokud vykazuje trend, je nutné jej revidovat. Pro více informací o QC vzorcích je čtenáři doporučena literatura [6] a [2].

Před začátkem analýzy je nutné ověřit, zda vzorky kontroly kvality neobsahují chybějící data. Pokud se takové hodnoty vyskytnou, jsou následně doplněny na základě pořadí nástřiku (z angl. *injection order*) s využitím informací z bezprostředně sousedících QC měření. Při doplňování chybějících hodnot se uplatňuje následující postup.

1. Pokud hodnota chybí uvnitř série měření, je nahrazena aritmetickým průměrem intenzit bezprostředně předcházejícího a následujícího QC vzorku. Zaveďme index $i = 1, \ldots, m$, kde m je celkový počet QC vzorků. Pro chybějící hodnotu uvnitř série (tj. pro $i \in \{2, \ldots, m-1\}$) je výpočet dán následujícím vzorcem

$$x_{QC,i} = \frac{x_{QC,i-1} + x_{QC,i+1}}{2},$$

   kde $x_{QC,i}$ je intenzita v aktuálním chybějícím bodě, $x_{QC,i-1}$ je intenzita předchozího a $x_{QC,i+1}$ následujícího QC vzorku.

2. V případě, že chybí hodnota u prvního nebo posledního QC vzorku v sérii, je hodnota nahrazena intenzitou nejbližšího sousedního měření.

Tento krok zajistí, že pro každý metabolit je k dispozici kompletní řada QC bodů, což umožní proložení korekční křivky v celém rozsahu, viz kapitola 2.1.

1.3 Reálná data

Pro demonstraci jednotlivých metod byla využita reálná metabolomická data. Jedná se o datový soubor poskytnutý Fakultní nemocnicí Olomouc (FNOL), který pochází z analýzy včelí hemolymfy. Včely byly chovány v klíckách, kdy každá klícka byla krmena jiným typem krmiva: 12 klícek pro každé krmivo a celkem tedy 48 klícek. Z každé klícky byly odebrány tři vzorky včel, resp. jejich hemolymfy (3 replikáty). V datasetu se nachází čtyři skupiny podle krmení včel:

- kontrolní skupina ($CTRL$), která byla krmena 50 % vodným roztokem sacharózy (včely nedostávají v krmivu protein),
- krmeny pylem (POL),
- krmeny řasou chlorelou (ALG),
- krmeny Ultrabee směsí (UB) - jedná se o komerční a běžně dostupnou krmící směs pro včely; směs rostlinných proteinů.

Všechny včely byly krmeny danou potravou po dobu 10 dní. Výchozí datový soubor obsahoval celkem 144 vzorků včelí hemolymfy, tj. 36 vzorků pro každou skupinu. Hlavním cílem bylo porovnání metabolomického profilu hemolymf včel na rozdílné výživě. Podle zvoleného analytického uspořádání byl na úvod série zařazen blok tří kontrolních vzorků, které slouží primárně ke stabilizaci měřicího systému. Následně byly QC vzorky do sekvence vkládány v pravidelných intervalech po každém devátém vzorku. Původní datový soubor obsahoval celkem 323 metabolitů. Zhodnocení úvodního bloku měření však prokázalo počáteční nestabilitu analytického systému. Z tohoto důvodu byly vyřazeny ty metabolity, které vykazovaly chybějící hodnotu (NA) alespoň v jednom z prvních tří kontrolních vzorků. Následně byly z datové ma-

tice odstraněny i počáteční stabilizační QC vzorky. Po těchto úpravách byl konečný počet metabolitů zredukován na 131, přičemž stabilita jejich měření je zaručena zbývajícími, pravidelně zařazovanými QC vzorky, kterých k další analýze zůstalo celkem 17.

Pro následující analýzu budeme předpokládat datovou matici $\mathbf{X}$, kde jednotlivé řádky (i) představují jednotlivé proměnné (metabolity) a sloupce (j) reprezentují jednotlivá pozorování (vzorky). Prvek x_{ij} tedy udává naměřenou intenzitu i-tého metabolitu v j-tém vzorku. Tomuto uspořádání jsou následně přizpůsobeny i veškeré výpočty a iterace v přiložených Python skriptech.

Vzhledem k tomu, že použitá datová sada bude součástí dalšího výzkumného projektu, nebudou zdrojová data v rámci této práce veřejně zpřístupněna.

2 Předzpracování dat

Předzpracování představuje klíčový krok před samotným statistickým vyhodnocením metabolomických dat. Důvodem je skutečnost, že měřicí přístroje mají svá technická omezení a data mohou být zatížena chybami měření, způsobenými například kolísáním tlaku v přístroji nebo změnami teploty v laboratoři. Ke korekci těchto chyb jsou nezbytné speciální statistické metody využívající vzorky kontroly kvality.

Po vyloučení nestabilních metabolitů na základě vyhodnocení vzorků kontroly kvality navazujeme na další nezbytné matematické úpravy, které zahrnují transformaci, normalizaci a finální škálování datového souboru. Provedení těchto kroků je kriticky důležité z důvodu obrovské přirozené variability biologického materiálu, kde se koncentrace jednotlivých metabolitů mohou vzájemně lišit o několik řádů. Cílem je sjednotit datové rozsahy tak,

aby se výsledné metabolické profily jednotlivých vzorků (typicky pacientů) staly objektivně a přímo srovnatelnými pro následnou statistickou analýzu. Kompletní výsledky z fáze předzpracování jsou k dispozici v excelové příloze *Preprocessing_result.xlsx* a příslušné zdrojové kódy se nacházejí v Python skriptu *Preprocessing.ipynb*.

Vzhledem k vysoce dimenzionální povaze dat je ještě před samotným předzpracováním nezbytné provést prvotní filtraci. Z datového souboru jsou vyřazeny ty metabolity, u nichž podíl chybějících hodnot přesahuje hranici 30 %. V našem případě však na základě tohoto pravidla k vyřazení žádného metabolitu nedošlo, jelikož se v celém datasetu vyskytují pouze dvě chybějící hodnoty.

Poznámka: Z hlediska matematické notace je v této kapitole využito značení odpovídající realizacím (konkrétním naměřeným hodnotám, tj. malá písmena), protože popisované algoritmy jsou přímo aplikovány na reálný datový soubor.

2.1 LOWESS korekce

Metodika v této podkapitole vychází z postupů popsaných v [6], [17] a [3]. Předpokladem použití QC vzorků je, že by jejich signál měl být v čase stabilní, jelikož se jedná o chemicky identický materiál. Jakýkoliv pozorovaný trend v intenzitě signálu QC vzorků je tedy považován za technickou chybu, kterou je nutné korigovat. K tomuto účelu se využívá metoda korekce signálu pomocí lokálně vážené regrese (zkrácené LOWESS, z angl. *locally weighted scatterplot smoothing*). LOWESS křivka je proložena body QC vzorků v závislosti na pořadí jejich nástřiku. Z této regrese je dále interpolována korekční křivka pro celou analytickou sérii. Pomocí ní je následně normalizován celý datový soubor pro každý metabolit, čímž se výsledné intenzity stabilních QC

vzorků transformují na ideální relativní hodnotu jedna.

LOWESS vyhlazování kombinuje jednoduchost klasické lineární regrese metodou nejmenších čtverců s flexibilitou nelineární regrese. Tato metoda nepřizpůsobuje datům jednu globální funkci, ale modeluje deterministickou část variability bod po bodu pomocí jednoduchých lokálních modelů. V každém bodě je proložena lokální polynomická regrese (typicky prvního nebo druhého stupně) na podmnožině dat, která je definována nejbližšími sousedy daného bodu. Polynomy jsou prokládány pomocí vážených nejmenších čtverců, kde body blíže odhadované hodnotě mají vyšší váhu než body vzdálenější.

Pro dosažení co nejpřesnějšího proložení regresní křivky je nezbytné optimalizovat tzv. parametr vyhlazování (z angl. *span*). Tento parametr určuje, jak velká část okolních datových bodů se využije pro výpočet každého lokálního polynomu. Parametr vyhlazování, značený α, je číslo mezi $(d+1)/n$ a 1, kde d představuje stupeň lokálního polynomu a n označuje celkový počet QC vzorků. Hodnota α udává podíl dat využitých pro každý lokální odhad. To znamená, že podmnožina dat použitá při každém proložení metodou vážených nejmenších čtverců obsahuje $n\alpha$ bodů (zaokrouhleno na nejbližší vyšší celé číslo). Pro příliš nízkou hodnotu tohoto parametru model začne zachytávat náhodný šum v datech a výsledná křivka bude vlnitá. Naopak při příliš vysoké hodnotě α se výsledná křivka bude blížit k prostému globálnímu proložení, čímž se ztratí schopnost zachytit lokální trendy.

2.1.1 Metoda robustní LOWESS

V této práci byla metoda aplikována pomocí funkce `lowess` z knihovny *statsmodels*, která vychází z robustního algoritmu popsaného Clevelandem [3].

Významným aspektem zvolené implementace je využití tzv. robustní regrese, která eliminuje vliv odlehlých hodnot v QC vzorcích. Standardní me-

toda LOWESS, založená pouze na vážení podle pozice v měřicí sekvenci, je citlivá na hrubé chyby měření. Pokud například jediný QC vzorek vykáže anomální signál vlivem chybného nástřiku, může to u standardního přístupu vést k výraznému lokálnímu zkreslení výsledné korekční křivky.

Robustní lokálně vážená regrese je metoda určená k vyhlazování množiny datových bodů (x_i, y_i), kde $i = 1, \ldots, n$. Odhadovaná hodnota pro libovolný bod x_k odpovídá hodnotě lokálního polynomu, který je daty proložen pomocí metody vážených nejmenších čtverců. Váha přiřazená konkrétnímu pozorování (x_i, y_i) je tím vyšší, čím blíže se bod x_i nachází k x_k, a se vzrůstající vzdáleností postupně klesá. Algoritmus navíc využívá robustní postup proložení, který efektivně zabraňuje tomu, aby odlehlá měření zkreslila výsledné vyhlazené hodnoty.

Představená vyhlazovací metoda vychází z předpokladu, že analyzovaná data lze popsat modelem ve tvaru

$$y_i = g(x_i) + \epsilon_i,$$

kde g je hladká funkce a ϵ_i jsou náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem. V tomto modelu hodnota $\hat{y}_i$ představuje lokální odhad funkce $g(x_i)$. Předpoklad hladkosti funkce g umožňuje použít body v okolí (x_i, y_i) pro výpočet hodnoty $\hat{y}_i$. Než přistoupíme k detailnímu výpočtu, nechť W je váhová funkce s následujícími vlastnostmi:

1. $W(x) > 0$ pro $|x| < 1$,
2. $W(-x) = W(x)$,
3. $W(x)$ je nerostoucí funkce pro $x \geq 0$,
4. $W(x) = 0$ pro $|x| \geq 1$.

Nechť $0 < \alpha \leq 1$ je parametr vyhlazování a nechť r označuje počet bodů v lokální podmnožině dat, přičemž jeho hodnota odpovídá součinu αn zaokrouhlenému na nejbližší celé číslo.

Vzhledem k tomu, že váhová funkce $W(x)$ pro nezáporná x monotónně klesá, snižují se i samotné váhy $w_k(x_i)$ s rostoucí vzdáleností bodu x_k od x_i. To znamená, že blízko sousedící body mají na výsledný odhad $\hat{y}_i$ zásadní vliv, zatímco váha vzdálenějších pozorování klesá. Zvyšování α navíc zvětšuje okolí vlivných bodů a tím má tendenci zvyšovat hladkost vyhlazených bodů.

Před samotným uvedením rovnic si nejprve stručně popíšeme hlavní myšlenku celého výpočetního postupu. Počáteční odhad $\hat{y}_i$ (tzv. vyrovnaná hodnota) v každém bodě x_i odpovídá hodnotě polynomu stupně d, který je daty proložen pomocí metody vážených nejmenších čtverců s využitím vzdálenostních vah $w_k(x_i)$. Tento prvotní výpočet se nazývá standardní lokálně vážená regrese. V dalším kroku je pro každé pozorování definována nová, doplňující robustní váha (označovaná jako δ) na základě velikosti jeho rezidua. Čím větší je absolutní hodnota rezidua (tedy čím více se bod odchyluje od prvotního modelu), tím menší robustní váha je mu přiřazena, a naopak. Následně se vypočítají nové odhady úplně stejně jako v prvním kroku, avšak původní vzdálenostní váhy $w_k(x_i)$ se nahradí jejich součinem s robustními vahami. Tento postup aktualizace vah a opětovného proložení se následně několikrát iterativně opakuje. Celý tento proces, zahrnující počáteční výpočet i následné iterace pro eliminaci odlehlých hodnot, se nazývá robustní lokálně vážená regrese.

Nyní přistoupíme k detailnímu matematickému popisu samotného algoritmu. Pro každý datový bod x_i nechť h_i označuje jeho vzdálenost k r-tému nejbližšímu sousedovi. Hodnota h_i tedy odpovídá r-té nejmenší hodnotě mezi

$|x_i - x_j|$ pro $j = 1, \ldots, n$. Pro $k = 1, \ldots, n$ definujme váhy

$$w_k(x_i) = W(h_i^{-1}(x_k - x_i)).$$

Algoritmus robustní lokálně vážené regrese je definován následujícími kroky.

1. Pro každé i se vypočítají odhady $\hat{\beta}_j(x_i)$, $j = 0, \ldots, d$, parametrů polynomiální regrese stupně d závislosti y_k na x_k. Ta je proložena pomocí metody vážených nejmenších čtverců s vahami $w_k(x_i)$ pro body (x_k, y_k). Tedy $\hat{\beta}_j(x_i)$ jsou hodnoty β_j, které minimalizují

$$\sum_{k=1}^{n} w_k(x_i)(y_k - \beta_0 - \beta_1 x_k - \ldots - \beta_d x_k^d)^2.$$

   Vyhlazený bod v x_i s použitím lokálně vážené regrese stupně d má souřadnice $(x_i, \hat{y}_i)$, přičemž $\hat{y}_i$ je predikovaná hodnota regrese v bodě x_i. Platí tedy

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(x_i) x_i^j.$$

2. Nechť B je bikvadratická váhová funkce (tzv. *bisquare weight function*) definovaná jako

$$B(x) = \begin{cases} (1 - x^2)^2, & \text{pro } |x| < 1 \\ 0, & \text{pro } |x| \geqslant 1. \end{cases}$$

   Nechť $e_i = y_i - \hat{y}_i$ představují rezidua z aktuálních predikovaných hodnot. Nechť s je medián hodnot těchto reziduí $|e_i|$. Robustní váhy jsou pak definovány vztahem

$$\delta_k = B(e_k/6s).$$

3. Pro každé i se následně vypočítají nové hodnoty $\hat{y}_i$ opětovným pro-

ložením polynomu d-tého stupně pomocí vážené metody nejmenších čtverců s upravenými vahami $\delta_k w_k(x_i)$ v bodech (x_k, y_k).

4. Kroky 2 a 3 se iterativně opakují celkem t-krát. Konečné hodnoty $\hat{y}_i$ získané v poslední iteraci představují výsledné hodnoty robustní lokálně vážené regrese.

Účelem iterativního postupu popsaného v krocích 2 až 4 je zajištění robustnosti výsledné regrese. Cílem je zaručit, aby ani přítomnost určitého podílu odlehlých hodnot nezkreslila výsledný odhad. Tyto odlehlé hodnoty, které typicky vznikají v situacích, kdy má náhodná chyba ϵ_i rozdělení s těžkými chvosty (z angl. *heavy-tailed distribution*), získávají v průběhu iterací velmi nízké robustní váhy δ_k. Díky tomuto mechanismu se jejich vliv na výsledný tvar regresní křivky efektivně minimalizuje.

Jakmile algoritmus zkonverguje a robustní váhy δ_k jsou pevně určeny, je možné vypočítat odhadovanou hodnotu pro jakýkoliv bod x (tedy i mimo původně naměřené pozice x_i). K výpočtu se opět využije lokální polynomiální proložení, tentokrát s aplikací výsledných vah $\delta_k w_k(x)$. Díky tomuto přístupu lze odhadované hodnoty snadno vyčíslit a graficky znázornit, například v ekvidistantní síti bodů na ose x.

2.1.2 Specifikace parametrů metody LOWESS v Pythonu

Závěr této podkapitoly je věnován specifikaci klíčových parametrů modelu `lowess` v jazyce Python, které byly následně aplikovány při zpracování reálných dat.

Parametr *it* definuje počet iterací robustního proložení. V našem případě používáme $it = 2$.

- Při hodnotě $it = 0$ je provedena pouze standardní lokální vážená regrese. Algoritmus proloží data vyhlazovací křivkou bez aplikace robustního kroku.

- Při nastavení hodnoty $it > 0$ se aktivuje iterativní robustní postup (v praxi se obvykle volí 1 až 2 iterace). Po prvotním proložení křivky jsou pro jednotlivá měření vyčíslena rezidua. Bodům vykazujícím značnou odchylku od modelu je v následujícím kroku výpočtu přiřazena adekvátně nižší váha, a regresní proložení se s těmito upravenými vahami provede znovu.

Dále byl parametr vyhlazování α (v implementaci jazyka Python označovaný argumentem *frac*) pro naši analýzu nastaven na hodnotu 0.7. Jak již bylo definováno teoreticky, tento parametr určuje podíl dat využitých pro proložení každého lokálního polynomu. V našem datovém souboru, kde celkový počet QC vzorků činí $n = 17$, to konkrétně znamená, že pro výpočet každého bodu korekční křivky bylo zohledněno 12 nejbližších QC vzorků (dle vztahu $n \cdot \alpha$, po zaokrouhlení nahoru).

Ačkoliv je regresní křivka prokládána výhradně na základě hodnot vzorků kontroly kvality, výsledný model umožňuje interpolaci vyhlazených hodnot pro naprosto všechny analyzované vzorky v závislosti na jejich indexu (pořadí měření). Způsob, jakým jsou vyhlazené hodnoty počítány pro mezilehlé indexy v sekvenci, lze optimalizovat pomocí parametru *delta*. Tento parametr slouží k úspoře výpočetního času u velmi rozsáhlých datových souborů a určuje, zda bude pro každý bod proveden plnohodnotný výpočet, nebo zda bude využita aproximace. V naší analýze využíváme výchozí nastavení $delta = 0$.

- Při hodnotě $delta > 0$: Parametr definuje vzdálenost na ose x, v je-

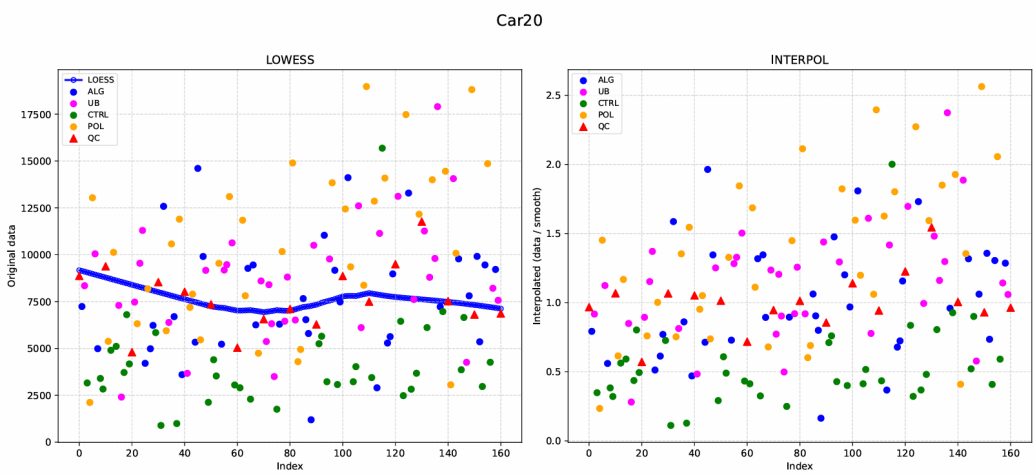
jímž rámci je výpočetně náročná lokální vážená regrese nahrazena zjednodušenou lineární interpolací. Pro daný bod se tak přeskočí exaktní výpočet regrese u všech okolních bodů, které jsou k němu blíže než stanovená vzdálenost *delta*. Další lokální regrese se proloží až pro bod ležící za touto hranicí a hodnoty všech bodů mezi nimi jsou pouze odhadnuty lineární interpolací (tedy spojením dvou vzdálenějších bodů rovnou čarou). Tento přístup tvůrci algoritmu doporučují u rozsáhlých dat ($N > 5000$).

- Při hodnotě $delta = 0$: Jedná se o výchozí nastavení funkce. Při této hodnotě nedochází k žádné aproximaci lineární interpolací. Místo toho je pro každý jednotlivý analyzovaný vzorek definováno vlastní vyhlazovací okno. Hodnota pro daný index je pak vypočítána přímo pomocí samostatné lokální regrese, a to pouze na základě QC vzorků nacházejících se uvnitř tohoto okna. Každý bod v sekvenci je tak získán nezávislým lokálním výpočtem.

Jakmile jsou tyto vyhlazené hodnoty získány pro celý datový soubor, přistupuje se k samotné korekci signálu. Ta spočívá v normalizaci dat, při které jsou původní naměřené intenzity vyděleny svými odpovídajícími vyhlazenými hodnotami.

Na obrázku 1 vidíme výsledek metody na reálných datech pro proměnnou *Car20*. Výsledné grafy pro další proměnné čtenář najde v příloze *Smooth.pdf*. Vlevo je modře znázorněna regresní křivka, která prokládá výhradně QC vzorky (vyznačené červenými trojúhelníky). Z průběhu křivky je zřejmá robustnost této metody, neboť nedochází k jejímu vychýlení vlivem přítomné odlehlé hodnoty. Vpravo máme interpolované hodnoty.

Z naměřených dat je patrné, že kontrolní skupina (*CTRL*) vykazuje nižší hodnoty metabolitu *Car20* ve srovnání s ostatními skupinami. Tento rozdíl



Obrázek 1: Aplikace robustní LOWESS na reálná data

indikuje, že sledovaná krmiva (kromě *CTRL*) aktivně ovlivňují metabolismus a podporují lepší zpracování energetických zásob.

Poznámka: Před samotnou aplikací korekce signálu metodou LOWESS je nezbytné ověřit, zda u některého z QC vzorků nedošlo k úplnému výpadku. Takové chybějící hodnoty by znemožnily konstrukci spojité korekční křivky, a proto musí být před dalším výpočtem detekovány a nahrazeny postupem imputace, který byl popsán v předchozí části textu 1.2.

Poznámka: V praxi může nastat situace, kdy je v sérii naměřen nedostatečný počet QC vzorků (méně než sedm), což může způsobit nestabilitu odhadu při použití metody LOWESS. Pokud v okolí určitého indexu není dostatečný počet QC bodů, může metoda vracet chybějící hodnoty a regresní křivka se stává nespojitou. Proto v těchto případech přidáváme na začátek a na konec série tři umělé QC body s intenzitou rovnou nejbližšímu skutečnému QC měření. Tyto body slouží pouze ke stabilizaci výpočtu vyhlazení a nejsou dále používány při samotné korekci dat.

2.2 Kritéria výběru metabolitů

Následným krokem po korekci dat je filtrace datového souboru, jejímž cílem je odstranit nespolehlivě měřené metabolity. Ke zhodnocení kvality měření se i v této fázi využívají QC vzorky. Pro posouzení stability a dostatečné intenzity signálu byl pro každý metabolit vypočítán variační koeficient (CV) a medián hodnot. Výpočet obou těchto kritérií je implementován v Pythonu pomocí vlastních funkcí dle příslušných vzorců.

2.2.1 Variační koeficient

Variační koeficient (CV) představuje relativní míru variability, která v procentech vyjadřuje míru kolísání naměřeného signálu vůči jeho průměrné hodnotě [2]. Vzhledem k tomu, že QC vzorky jsou chemicky zcela identické, měl by být jejich naměřený signál v ideálním případě konstantní. Jakákoliv pozorovaná odchylka tedy přímo odráží technickou variabilitu (chybu) měření.

Variační koeficient pro i-tý metabolit na základě QC vzorků spočítáme podle následujícího vztahu

$$CV_i = \frac{s_{QC,i}}{\bar{x}_{QC,i}} \cdot 100,$$

kde $s_{QC,i}$ je vypočtená směrodatná odchylka a $\bar{x}_{QC,i}$ je průměr intenzit daného metabolitu v QC vzorcích.

Jako hranice pro vyřazení metabolitů byla stanovena hodnota 30 %. Metabolity, jejichž signál v QC vzorcích vykazoval vyšší míru (tedy CV dosáhlo nebo přesáhlo 30 %), byly vyhodnoceny jako zatížené příliš velkou chybou měření a z další analýzy byly odstraněny. Tento limit vychází z běžně užívaných protokolů pro necílenou metabolomiku, kde se tato metrika často označuje také jako relativní směrodatná odchylka RSD [6, 2].

V tabulce 1 jsou shrnuty hodnoty variačního koeficientu (CV) pro QC vzorky před a po aplikaci metody LOWESS. Před provedením korekce mělo celkem 30 metabolitů příliš vysokou variabilitu s hodnotou CV nad 30 %. Díky úspěšné korekci se tento počet podařilo snížit na pouhé 2 nestabilní metabolity (konkrétně *Cr* a *Hx*). Jelikož tyto dva metabolity vykazovaly vysokou variabilitu i po korekci, z další analýzy jsme je vyloučili.

Tabulka 1: Varianční koeficient pro QC před a po aplikaci LOWESS korekce.

a) Počet nestabilních metabolitů

Fáze zpracování	CV(QC) $>$ 30 %
Před korekcí	30
Po korekci	2

b) Nestabilní metabolity

Metabolit	Výsledné CV [%]
Cr	49,0
Hx	66,8

2.2.2 Odchylka intenzit v QC vzorcích

V rámci předzpracování dat byla provedena filtrace metabolitů na základě porovnání jejich intenzit v QC vzorcích a v reálných biologických vzorcích. Vzorky kontroly kvality jsou v necílené metabolomice standardně připravovány slitím stejných alikvotů všech analyzovaných vzorků. Z principu své přípravy by měl QC vzorek reprezentovat průměrné složení celého datového souboru. Očekává se tedy, že medián intenzity každého stabilního metabolitu v QC vzorcích bude odpovídat mediánu intenzity v reálných vzorcích.

Pokud se však medián intenzity metabolitu v QC vzorcích liší o více než 150 % od mediánu biologických vzorků, je takový metabolit vyřazen. Tuto podmínku lze matematicky zapsat jako

$$|\tilde{x}_{\mathrm{QC},i} - \tilde{x}_{\mathrm{vzorky},i}| > 1{,}5 \cdot \tilde{x}_{\mathrm{vzorky},i},$$

kde $\tilde{x}_{\mathrm{QC},i}$ je medián intenzit i-tého metabolitu v QC vzorcích a $\tilde{x}_{\mathrm{vzorky},i}$ je medián jeho intenzit v reálných biologických vzorcích. Tato extrémní odchylka indikuje, že QC vzorek pro daný metabolit nereflektuje reálné složení. Příčinou této nestability může být technická chyba během přípravy směsného kontrolního vzorku, např. nechtěná kontaminace jiným materiálem.

V rámci tohoto kroku filtrování bylo z datové sady celkem odstraněno 11 metabolitů. Pro další analýzu tak zůstává k dispozici 118 metabolitů.

2.3 Imputace chybějících hodnot

Datové soubory běžně obsahují chybějící hodnoty, které jsou typicky reprezentovány nulami nebo jako NA. Takové hodnoty mohou vznikat z různých příčin, jako jsou problémy s detekcí píků, chyby při zarovnávání (tzv. *alignment*) nebo vlivem matricových efektů [15]. Přítomnost těchto chybějících údajů však komplikuje následné statistické zpracování, neboť většina vícerozměrných metod (např. PCA) vyžaduje kompletní datovou matici.

K řešení problému chybějících dat se v literatuře uvádí celá řada přístupů. Představíme si nejčastěji aplikované metody, mezi které patří jednoduchá imputace konstantní hodnotou a pokročilejší, vícerozměrné algoritmy založené na strojovém učení, mezi něž patří imputace metodou k nejbližších sousedů (zkráceně kNN, z angl. *k-Nearest Neighbors*) nebo využití algoritmu náhodných lesů (zkráceně RF, z angl. *Random Forest*). Imputace probíhá odděleně pro každou skupinu, aby nedošlo ke zkreslení rozdílů mezi skupinami.

2.3.1 2/3 min

Při imputaci konstantou lze chybějící data nahradit například průměrem, mediánem či hodnotou odpovídající 2/3 minimální naměřené intenzity daného metabolitu.

V naší implementaci funguje metoda nahrazení konstantou tak, že je nejprve identifikována minimální naměřená intenzita pro každý jednotlivý metabolit. Veškeré chybějící hodnoty u dané látky jsou následně nahrazeny hodnotou odpovídající 2/3 tohoto minima. Tento způsob použijeme, pokud máme důvodné podezření, že chybějící hodnoty (nuly) nejsou náhodné chyby, ale tzv. zleva cenzorovaná data (z angl. *left-censored values*). To znamená, že metabolit ve vzorku pravděpodobně je, ale jeho koncentrace je pod limitem detekce přístroje v biologickém materiálu při použití konkrétního nastavení chromatografie a/nebo hmotnostní spektrometrie [15].

2.3.2 Metoda k nejbližších sousedů

Metoda k nejbližších sousedů (kNN) patří mezi základní neparametrické metody statistického učení. Pro specifický problém chybějících hodnot v metabolomických datech využíváme v Pythonu variantu `KNNimpute`, formálně definovanou v práci Troyanskaya [26]. Na rozdíl od klasifikačních úloh je zde princip založen na váženém průměrování spojitých hodnot nejbližších sousedů. Jelikož algoritmus vyhledává sousedy na základě euklidovské vzdálenosti, klíčovým předpokladem pro správné fungování této metody je provedení logaritmické transformace dat ještě před samotnou imputací.

V praxi algoritmus postupuje tak, že pro každý metabolit s neúplnými daty určí k nejbližších sousedních metabolitů pomocí euklidovské vzdálenosti. Chybějící pozice v datové matici jsou poté nahrazeny váženým průměrem, jenž je vypočítán výhradně z dostupných měření u těchto nalezených sousedů.

Algoritmus pro imputaci vektoru $\mathbf{x}$ s chybějící j-tou složkou probíhá následovně.

1. **Nalezení sousedů:** V prostoru o dimenzi p (kde p je počet proměnných bez chybějících hodnot) nalezneme množinu k nejbližších vektorů

označenou $N_k(\mathbf{x})$ na základě eukleidovské vzdálenosti $d(\mathbf{x}, \mathbf{y})$.

2. **Výpočet vah:** K potlačení vlivu vzdálenějších sousedů je každému sousedu $i \in N_k(\mathbf{x})$ přiřazena váha w_i, jež je nepřímo úměrná jeho vzdálenosti
$$w_i = \frac{1}{d(\mathbf{x}, \mathbf{y}_i)}.$$

3. **Imputace:** Chybějící hodnota $\hat{x}_j$ je následně odhadnuta jako vážený průměr hodnot j-tých složek všech nalezených sousedů
$$\hat{x}_j = \frac{\sum_{i=1}^{k} w_i y_{i,j}}{\sum_{i=1}^{k} w_i}.$$

Klíčovým parametrem je hodnota k, která určuje počet nejbližších sousedů. Abychom se vyhnuli subjektivní volbě pevné konstanty, je ideální velikost parametru k určit pomocí křížové validace. Samotná logika tohoto optimalizačního postupu byla převzata z publikace [24]. Algoritmus postupuje tak, že v rámci každé skupiny vzorků nejprve identifikuje platná měření a 5 % z nich náhodně skryje. Pomocí zbývajících platných měření pak algoritmus zkouší tyto skryté pozice predikovat pro různá k (v rozsahu od 2 do 20). Své odhady následně porovná se skutečnými skrytými hodnotami a pro každé k spočítá odmocninu střední čtvercové chyby (RMSE). Nastavení, u kterého nastala nejmenší chyba, se nakonec použije pro finální doplnění opravdových chybějících dat.

Imputaci pomocí kNN využijeme i na naše reálná data, která obsahují jen 2 chybějící hodnoty.

2.3.3 Random Forest

Alternativním a vysoce robustním přístupem pro imputaci chybějících dat je využití algoritmu náhodných lesů (RF), konkrétně jeho iterativní varianty označované jako MissForest. Jak uvádějí Stekhoven a Bühlmann v [24], hlavní výhodou této neparametrické metody je její schopnost zpracovat data smíšeného typu s minimem předpokladů o jejich struktuře. Algoritmus Random Forest se vyznačuje vysokou přesností i v náročných podmínkách, jakými jsou vysokodimenzionální data, složité interakce proměnných či nelineární vztahy. Samotný problém chybějících hodnot je zde řešen pomocí iterativního schématu: v prvním kroku je model natrénován na dostupných (pozorovaných) datech, následně jsou predikovány chybějící hodnoty a celý proces se iterativně opakuje a zpřesňuje.

Základním stavebním kamenem pro metodu Random Forest je binární rozhodovací strom (Hastie [10], kap. 9.2.), konkrétně algoritmus CART (z angl. *Classification and Regression Trees*). Vzhledem k tomu, že cílem je imputace spojitých veličin (intenzit či koncentrací metabolitů), zaměříme se na regresní stromy. Matematicky lze regresní strom definovat jako model, který rozděluje prostor prediktorů (vstupních proměnných) X na M vzájemně disjunktních oblastí $R_1, R_2, \ldots, R_M$. V každé z těchto oblastí je pak odezva modelována konstantou c_m. Predikční funkce $f(\mathbf{x})$ má tvar

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m I(\mathbf{x} \in R_m),$$

kde $I(\cdot)$ je indikátorová funkce. Pro minimalizaci součtu čtverců chyb je optimální konstantou $\hat{c}_m$ aritmetický průměr hodnot y_i v dané oblasti R_m

$$\hat{c}_m = \frac{1}{n_m} \sum_{\mathbf{x}_i \in R_m} y_i,$$

kde n_m představuje celkový počet pozorování v oblasti R_m.

Hledání optimálního rozdělení prostoru je výpočetně náročný kombinatorický problém, proto se využívá tzv. *greedy algorithm*. Proces začíná s celým datovým souborem a hledá optimální dělící proměnnou j a dělící bod s, které rozdělí data do dvou polorovin

$$R_1(j,s) = \{\mathbf{x} \mid x_j \leq s\} \quad \text{a} \quad R_2(j,s) = \{\mathbf{x} \mid x_j > s\}.$$

Dvojice (j,s) je volena tak, aby minimalizovala chybu v obou vzniklých uzlech

$$\min_{j,s}\left[\min_{c_1}\sum_{\mathbf{x}_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{\mathbf{x}_i\in R_2(j,s)}(y_i-c_2)^2\right].$$

Tento proces se rekurzivně opakuje pro každou vzniklou oblast, dokud není splněna zastavovací podmínka (např. minimální počet vzorků v uzlu).

Jedním z hlavních nedostatků regresních stromů je jejich vysoký rozptyl. Malá změna v trénovacích datech může vést k vytvoření zcela odlišného stromu. Pro snížení tohoto rozptylu se využívá metoda zvaná Bagging (z angl. *Bootstrap Aggregation*). Princip metody spočívá v trénování mnoha modelů na různých podmnožinách dat a následném zprůměrování jejich predikcí. Nechť $\mathbf{Z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ představuje trénovací data. Bootstrap je metoda opakovaného výběru vzorků, kdy z množiny $\mathbf{Z}$ opakovaně vybíráme n vzorků s vracením. Tím získáme B nových bootstrapových množin $\mathbf{Z}^{*1}, \mathbf{Z}^{*2}, \ldots, \mathbf{Z}^{*B}$. Pro každý bootstrapový výběr $\mathbf{Z}^{*b}$, kde $b = 1, \ldots, B$, natrénujeme regresní strom, čímž získáme predikční funkci $\hat{f}^{*b}(\mathbf{x})$.

Finální predikce metody Bagging v bodě $\mathbf{x}$ je definována jako průměr predikcí všech jednotlivých modelů

$$\hat{f}_{\text{bag}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(\mathbf{x}).$$

Metoda Random Forest neboli náhodný les (Hastie [10], kap. 15.) představuje podstatnou modifikaci metody Bagging. Ačkoliv Bagging efektivně snižuje rozptyl modelu, trpí jedním nedostatkem: pokud existuje v datech jeden nebo několik silných prediktorů, budou si stromy v souboru velmi podobné (budou vysoce korelované), což limituje efektivitu průměrování. Random Forest tento problém řeší zavedením dodatečné náhodnosti do procesu konstrukce stromů, čímž zvyšuje diverzitu jednotlivých stromů.

Algoritmus Random Forest pro regresi (imputaci spojitých hodnot) postupuje následovně.

1. Pro $b = 1$ až B (počet stromů v lese):
   (a) Z trénovacích dat $\mathbf{Z}$ provedeme bootstrapový výběr $\mathbf{Z}^{*b}$ o rozsahu n.
   (b) Na datech $\mathbf{Z}^{*b}$ vybudujeme regresní strom T_b. Při každém dělení uzlu však neprohledáváme všech p proměnných, ale náhodně vybereme pouze podmnožinu m proměnných (kde $m < p$).
   (c) Z této náhodné podmnožiny vybereme nejlepší proměnnou a dělící bod, který minimalizuje součet čtverců chyb.
   (d) Strom necháme růst do maximální hloubky bez prořezávání (tzv. *pruning*), což je technika odstraňování nevýznamných koncových větví jako ochrana proti přeučení modelu na trénovacích datech.
2. Výsledná predikce v bodě $\mathbf{x}$ je průměrem predikcí všech stromů

$$\hat{f}_{\mathrm{rf}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}).$$

Klíčovým parametrem je velikost náhodné podmnožiny m. Pro regresní úlohy se doporučuje volit $m \approx p/3$ nebo $m \approx \sqrt{p}$. Tato strategie zajišťuje,

že i méně dominantní prediktory mají šanci ovlivnit strukturu stromu, což vede k rozmanitějšímu (a tedy robustnějšímu) souboru modelů.

Ačkoliv má standardní algoritmus Random Forest schopnost pracovat s chybějícími hodnotami v prediktorech, pro trénování vyžaduje kompletní vektor pozorování závisle proměnné. V kontextu imputace, kdy se každá proměnná s chybějícími daty stává postupně odezvou, představuje tento požadavek komplikaci. Algoritmus MissForest [24] řeší tento problém zavedením inicializačního kroku a iterativního schématu. Pro libovolnou proměnnou X_s (kde $s = 1, \ldots, p$) obsahující chybějící hodnoty můžeme datovou matici rozdělit do čtyř částí

- $\mathbf{y}_{\text{obs}}^{(s)}$ představuje vektor pozorovaných hodnot proměnné X_s,
- $\mathbf{y}_{\text{mis}}^{(s)}$ představuje vektor chybějících hodnot proměnné X_s,
- $\mathbf{X}_{\text{obs}}^{(s)}$ označuje matici hodnot všech ostatních proměnných (mimo X_s) v řádcích, kde je hodnota X_s pozorována,
- $\mathbf{X}_{\text{mis}}^{(s)}$ označuje matici hodnot všech ostatních proměnných (mimo X_s) v řádcích, kde hodnota X_s chybí.

Samotná imputace pak probíhá v následujících krocích.

1. Nejprve se provede počáteční hrubý odhad chybějících hodnot v celé matici, například pomocí průměru nebo mediánu. Následně se proměnné X_s seřadí vzestupně podle množství chybějících hodnot (od proměnné s nejmenším počtem chybějících hodnot po tu s největším).

2. Pro každou proměnnou X_s (v daném pořadí) se nejprve natrénuje model náhodného lesa, kde závislou proměnnou tvoří pozorované hodnoty $\mathbf{y}_{\text{obs}}^{(s)}$ a prediktory tvoří matice $\mathbf{X}_{\text{obs}}^{(s)}$. Natrénovaný model se následně aplikuje na matici $\mathbf{X}_{\text{mis}}^{(s)}$, čímž se získají predikce pro chybějící hodnoty $\mathbf{y}_{\text{mis}}^{(s)}$.

3. Imputovaná datová matice se o tyto nově predikované hodnoty aktualizuje. Další proměnná v pořadí tak již pro svou vlastní predikci využívá tyto zpřesněné hodnoty.

4. Kroky 2 a 3 se iterativně opakují pro všechny proměnné, dokud není splněno ukončovací kritérium.

Ačkoliv byl tento algoritmus původně implementován ve stejnojmenném balíčku *MissForest* [24], v prostředí Python se dnes jedná o zastaralou a neudržovanou knihovnu. Z toho důvodu se v současné praxi využívá stabilní knihovna *scikit-learn*, kde je původní iterativní algoritmus MissForest spolehlivě nahrazen použitím modulu `IterativeImputer` v kombinaci s bázovým modelem `RandomForestRegressor`.

2.4 Logaritmická transformace

Metabolomická data často vykazují silně pravostranně zešikmené rozdělení a heteroskedasticitu, kdy rozptyl roste úměrně s průměrnou hodnotou signálu. Pro účely této práce byla zvolena transformace přirozeným logaritmem [27]. Její primární úlohou je převedení multiplikativních vztahů na aditivní, stabilizace rozptylu a úprava asymetrických rozdělení do podoby blížící se normálnímu rozdělení. Splnění těchto podmínek je klíčovým předpokladem pro mnohé následné statistické analýzy.

Vzhledem k tomu, že se koncentrace různých metabolitů mohou ve vzorcích lišit o několik řádů, je transformace nezbytná i pro správné fungování předchozích kroků předzpracování. Algoritmy založené na euklidovské vzdálenosti, jako je metoda kNN využívaná pro imputaci chybějících hodnot, jsou extrémně citlivé na absolutní velikost hodnot. Bez transformace by ve výpočtu podobnosti vzorků dominovaly metabolity s nejvyššími absolutními

hodnotami. Logaritmická transformace však snižuje vysoké hodnoty výrazněji než ty nízké, čímž přirozeně stlačuje široký dynamický rozsah dat. Tím zajišťuje, že do výpočtu vzdálenosti smysluplně přispějí i metabolity s nižšími koncentracemi.

Tímto jevem však vytváří tzv. efekt pseudoškálování. Tento efekt však není určen násobením škálovacím faktorem, jako u „skutečného“ efektu škálování, ale účinkem, který mají tyto transformace na původní hodnoty. Tento efekt pseudoškálování proto zřídka stačí k úplnému přizpůsobení rozdílů v rozsahu hodnot. Proto může být užitečné po transformaci použít škálovací metodu.

V Pythonu stačí využít standardní funkci `numpy.log`.

2.5 Normalizace PQN

Každý vzorek může vykazovat různou celkovou intenzitu v důsledku technických či fyziologických faktorů, jako je například odlišný objem moči nebo odebrané kožní tekutiny. K odstranění těchto rozdílů v naředění se využívá normalizace pomocí PQN (z anglického *Probabilistic Quotient Normalization*). V této podkapitole čerpáme z článku [5].

Tato metoda vychází z předpokladu, že biologicky významné změny koncentrací jednotlivých analytů ovlivňují pouze část naměřených metabolitů, zatímco fyzikální změny v celkovém naředění vzorku ovlivňují plošně všechny proměnné. Intenzita většiny signálů je tedy primárně funkcí ředění. Pro zjištění míry tohoto ředění se každý vzorek porovnává s tzv. referenčním spektrem (nejčastěji se využívá medián všech vzorků). Výsledným normalizačním faktorem pro daný vzorek je pak medián podílů (kvocientů) mezi ním a referencí, což představuje velmi robustní odhad celkového posunu.

Proces normalizace lze formálně popsat v následujících čtyřech krocích.

1. **Výpočet referenčního spektra:** Pro každý metabolit $i \in \{1, \ldots, p\}$ je stanoven referenční bod $x_{ref,i}$ jako medián jeho intenzit napříč všemi n vzorky

$$x_{\text{ref},i} = \text{median}(x_{i,1}, x_{i,2}, \ldots, x_{i,n}).$$

   Výsledkem je referenční vektor reprezentující typický metabolický profil celého datového souboru.

2. **Výpočet matice kvocientů (fold-change):** Každá hodnota $x_{i,j}$ v původní datové matici je vydělena příslušnou referenční hodnotou daného metabolitu, čímž vzniká matice kvocientů $\mathbf{Q}$

$$q_{i,j} = \frac{x_{i,j}}{x_{ref,i}}.$$

3. **Odhad normalizačních koeficientů:** Pro každý vzorek $j \in \{1, \ldots, n\}$ je extrahován PQN koeficient c_j určující míru jeho globálního naředění. Tento koeficient je vypočten jako medián kvocientů daného vzorku přes všechny metabolity

$$c_j = \text{median}(q_{1,j}, q_{2,j}, \ldots, q_{p,j}).$$

4. **Normalizace dat:** Finální normalizovaná matice $\tilde{\mathbf{X}}$ vznikne vydělením každé původní hodnoty $x_{i,j}$ příslušným PQN koeficientem daného vzorku

$$\tilde{x}_{i,j} = \frac{x_{i,j}}{c_j}.$$

Poznámka: Vzhledem k logaritmické transformaci dat je v Python skriptu klasický výpočet PQN modifikován. Z vlastnosti logaritmů v kroku 2 a 4 využíváme místo dělení odčítání: $\log(\frac{A}{B}) = \log(A) - \log(B)$.

Přestože logaritmickou transformaci dat provádíme standardně, normalizace pomocí PQN je volitelným krokem, který závisí na typu dat. Je užitečná zejména v případech, kdy analyzujeme vzorky o různých výchozích objemech, např. když u staršího pacienta odebereme jiné množství kožní tekutiny než u mladšího. PQN dokáže tyto rozdíly v naředění vzorků efektivně kompenzovat. Tento efekt vyrovnání je dobře patrný na volcano grafu (viz kapitola 3.10). Pokud je vzorek celkově více naředěný, velká část metabolitů se uměle posune doleva do záporných hodnot, což by vedlo k falešně pozitivním výsledkům. PQN tento posun koriguje a celou distribuci vycentruje zpět k nule.

2.6 Škálování dat

Na závěr této kapitoly se podíváme na škálování dat, které se používá pro odstranění závislosti pořadí metabolitů na průměrné koncentraci a velikosti, viz článek [27]. Škálovaná data dále používáme např. do PCA. Zaměříme se na tyto čtyři typy škálování: centrování, UV, range a Pareto škálování. Jednotlivé metody jsou popsány v následujících odstavcích, ze kterých přímo vychází i jejich implementace v jazyce Python.

Centrování

Cílem centrování je odstranit z dat absolutní posun (tzv. *offset*). Matematicky lze tuto úpravu zapsat vztahem

$$\tilde{x}_{i,j} = x_{i,j} - \overline{x}_i,$$

kde $\overline{x}_i$ je průměrná hodnota i-tého metabolitu napříč všemi vzorky. Model se tak může soustředit pouze na relativní rozdíly mezi daty, nikoliv na jejich

absolutní podobnosti.

UV škálování

Metoda UV škálování (také *Autoscaling*) představuje jeden z nejběžnějších statistických přístupů pro úpravu dat. Matematicky lze tuto transformaci vyjádřit vztahem

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \overline{x}_i}{s_i},$$

kde s_i je směrodatná odchylka i-tého metabolitu. Díky tomuto škálování získají všechna data jednotný rozptyl. Všechny metabolity se tak stávají pro následný model stejně důležitými, a to zcela bez ohledu na to, v jakých absolutních koncentracích se v biologických vzorcích původně nacházely.

Range scaling

Další využívanou metodou je tzv. *Range scaling*. Matematicky lze tento výpočet zapsat jako

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \overline{x}_i}{x_{i_{\max}} - x_{i_{\min}}},$$

kde $x_{i_{\max}}$ je největší naměřená hodnota i-tého metabolitu napříč všemi vzorky a $x_{i_{\min}}$ je naopak jeho nejmenší naměřená hodnota. Podobně jako u UV škálování se všechny metabolity stávají stejně důležitými, nyní ale s větším důrazem na jejich biologický rozsah.

Pareto škálování

Jako poslední metodu uvedeme tzv. Pareto škálování

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \overline{x}_i}{\sqrt{s_i}},$$

které snižuje relativní důležitost velkých hodnot, ale zároveň částečně zachovává původní strukturu dat. Výsledek tak zůstává blíže původnímu měření než v případě UV škálování. Nevýhodou však je, že zůstává citlivější na velké změny v datech (tzv. fold-changes).

Tento typ škálování jsme využili i v naší analýze, jelikož se jedná o standardní metodiku kliniky, z níž data pocházejí. Následně budou tyto hodnoty vstupovat do PCA (viz kapitola 4.1).

3 Jednorozměrná statistika

Jakmile jsou data předzpracována, můžeme přistoupit k samotné analýze. Prvním krokem je jednorozměrná statistika, v rámci které se postupně zaměříme na popisnou část, testování hypotéz, korelační analýzu a ROC analýzu.

Úvodní tři podkapitoly se zabývají popisnou statistikou, kam spadá testování normality, detekce odlehlých hodnot a souhrnné charakteristiky doplněné o grafické výstupy. Následující část se zaměřuje na testování hypotéz prostřednictvím parametrických i neparametrických testů, a to včetně nezbytné korekce p-hodnot. Závěr této kapitoly je pak věnován korelaci, ROC analýze a vizualizaci pomocí Volcano plotu.

Veškeré výsledky daných metod jsou uvedeny v přiloženém excelovém souboru *Univariate_results.xlsx*, přičemž kompletní implementace výpočtů se nachází v Python skriptu *Univariate_analysis.ipynb*.

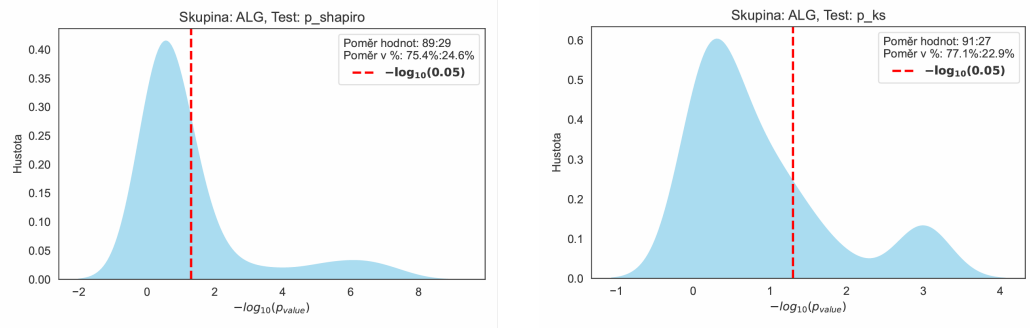
Poznámka: Na rozdíl od kapitoly věnované předzpracování dat, kde bylo využito značení pro realizace, se v této kapitole pro teoretický popis statistických modelů a testů využívá standardní značení pomocí náhodných veličin (velká písmena).

3.1 Test normality

Před samotným testováním hypotéz je nutné ověřit předpoklad normality dat, na kterém závisí volba dalších statistických metod. K tomu je v práci využit Shapiro-Wilkův a Lillieforsův test (modifikace Kolmogorov-Smirnovova testu pro neznámé parametry). Normalita je testována pro každou skupinu vzorků (typicky jedna nebo více skupin pacientů a kontroly) odděleně.

Na obrázku 2 jsou znázorněny výsledky testování normality pro skupinu ALG pomocí obou testů. Na ose x se nacházejí transformované hodnoty

$-\log_{10}(p)$ a osa y zobrazuje hustotu těchto hodnot. Červená přerušovaná čára vyznačuje hraniční hodnotu $-\log_{10}(0{,}05)$. V levé části od této vertikály leží transformované p-hodnoty, pro které v původním měřítku platí $p \geq 0{,}05$. U těchto proměnných nebyl nalezen dostatek důkazů pro zamítnutí normality, lze je tedy považovat za normálně rozdělené. Naopak v pravé části klesá p-hodnota pod 0,05, což znamená, že u těchto proměnných byla normalita striktně zamítnuta. Předpoklad normality podle Shapiro-Wilkova testu nebyl zamítnut u 89 proměnných (75,4 %), v případě Lillieforsova testu to bylo u 91 proměnných (77,1 %). U většiny proměnných ve skupině *ALG* tedy nezamítáme nulovou hypotézu o normálním rozdělení. Výsledky pro ostatní skupiny vykazují podobný trend (viz příloha *Normality.pdf*).



(a) Shapiro-Wilk test

(b) Lillieforsův test

Obrázek 2: Porovnání rozložení p-hodnot pro skupinu *ALG* podle zvoleného testu normality

3.2 Odlehlé hodnoty

Pro detekci odlehlých hodnot uvažujeme tři metody: Grubbsův test, metoda mezikvartilového rozpětí (IQR) a modifikované Z-skóre. Testování probíhá po skupinách, tj. hledáme odlehlé hodnoty v rámci jedné skupiny. Po označení složky za odlehlou ji z datového souboru odstraníme. V příloze *Univariate_results.xlsx* jsou tyto buňky vyznačeny barevně.

3.2.1 Grubbsův test

Grubbsův test, popsaný v publikacích [1] a [19], je parametrická metoda pro identifikaci odlehlých hodnot v jednorozměrném datovém souboru, přičemž předpokládá normální rozdělení dat. Vzhledem k vícerozměrné povaze metabolomických dat není možné aplikovat tento test na celou matici současně. Z toho důvodu byl navržen algoritmus, který zpracovává data po jednotlivých proměnných. Pro každý zkoumaný metabolit byl z matice extrahován jednorozměrný vektor hodnot odpovídajících intenzitám dané látky napříč všemi vzorky ve zkoumané skupině.

Grubbsův test je definován pro následující hypotézy:

H_0: V datovém souboru nejsou žádné odlehlé hodnoty.

H_A: V datovém souboru je právě jedna odlehlá hodnota.

Testová statistika G je pak definována jako

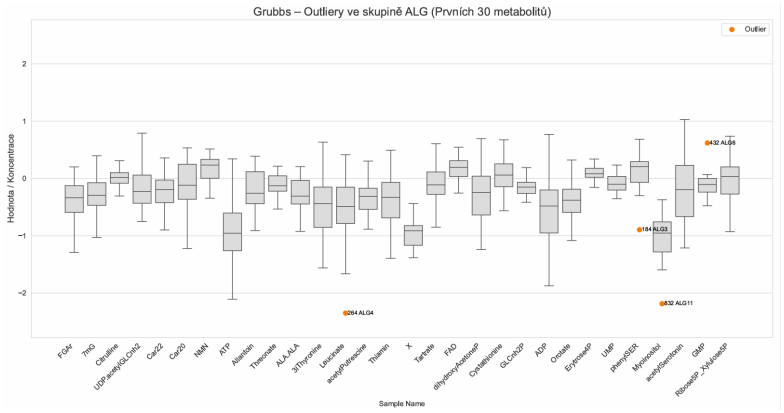
$$G = \frac{\max_{j=1,\dots,n} |X_j - \bar{X}|}{S},$$

kde X_j je náhodná veličina, jejíž realizace x_j odpovídá naměřené hodnotě daného metabolitu v j-tém vzorku, $\bar{X}$ je výběrový průměr a S je příslušná výběrová směrodatná odchylka.

Jedná se o oboustranný test, u kterého je hypotéza o neexistenci odlehlých hodnot zamítnuta na hladině významnosti α, pokud

$$g > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{(t_{\alpha/(2n),n-2})^2}{n-2+(t_{\alpha/(2n),n-2})^2}},$$

kde $t_{\alpha/(2n),n-2}$ označuje $\left(1-\frac{\alpha}{2n}\right)$-kvantil Studentova t-rozdělení s $(n-2)$ stupni volnosti. Korekce hladiny významnosti α na $\frac{\alpha}{2n}$ zohledňuje velikost souboru a oboustrannost testu (tzv. Bonferroniho korekce, viz kapitola 3.4).



Obrázek 3: Výsledek Grubbsova testu pro skupinu ALG zobrazené pro prvních 30 metabolitů

Jelikož Grubbsův test standardně identifikuje v jednom kroku pouze jedinou (tu nejvíce odlehlou) hodnotu, byl algoritmus implementován iterativně, aby bylo možné odhalit i vícenásobné extrémy. Nalezená odlehlá hodnota je odstraněna a test se opakuje. Vícenásobné opakování však mění pravděpodobnost detekce; test by proto neměl být používán pro vzorky o velikosti šest nebo méně ($n \leq 6$), protože v takovém případě často chybně označuje většinu bodů jako odlehlé hodnoty.

V jazyce Python využíváme test `Grubbs` z knihovny *outliers*. Pokud byla identifikována odlehlá hodnota, došlo k jejímu dočasnému vyřazení z datového vektoru a test byl na takto zredukovaný soubor aplikován znovu. Tento cyklus se opakoval, dokud testovací kritérium na hladině významnosti $\alpha = 0.05$ neodhalilo žádné další odlehlé pozorování.

Na obrázku 3 máme výsledek pro prvních 30 metabolitů ve skupině ALG. Oranžové body značí odlehlé hodnoty nalezené Grubbsovou metodou. Celkové výsledky čtenář najde v příloze *Grubbs_ outliers.pdf*.

3.2.2 Mezikvartilové rozpětí

Pro identifikaci odlehlých hodnot bez striktního předpokladu normálního rozdělení dat byla využita neparametrická metoda mezikvartilového rozpětí (IQR), jež představuje standardní nástroj popisné statistiky [12]. Tato metoda je vysoce robustní vůči extrémním hodnotám, neboť k výpočtu hranic nevyužívá průměr ani směrodatnou odchylku, ale kvartily.

Pro každou proměnnou jsou nejprve definovány teoretické kvartily: první kvartil Q_1 (25. percentil) a třetí kvartil Q_3 (75. percentil). Mezikvartilové rozpětí IQR je pak definováno jako

$$IQR = Q_3 - Q_1.$$

Za odlehlé hodnoty označíme všechna měření, jejichž realizace x_j se nacházejí mimo interval vymezený dolní a horní mezí

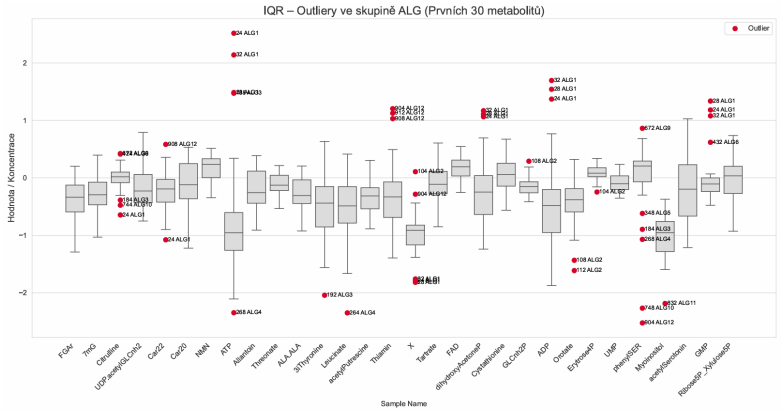
$$[\tilde{x}_{0,25} - 1{,}5 \cdot iqr, \quad \tilde{x}_{0,75} + 1{,}5 \cdot iqr],$$

kde $\tilde{x}_{0,25}$ a $\tilde{x}_{0,75}$ jsou hodnoty kvartilů vypočtené z naměřených dat a iqr je výsledné mezikvartilové rozpětí pro daný metabolit.

Odlehlé hodnoty detekované metodou IQR lze vyčíst i z krabicového grafu, kde jsou graficky znázorněny jako jednotlivé body ležící mimo rozsah tzv. vousů, viz obrázek 4 a dále v příloze *IQR_ outliers.pdf*.

3.2.3 Modifikované Z-skóre

Posledním přístupem detekce je robustní metoda založená na modifikovaném Z-skóre [16]. Výhodou tohoto přístupu je jeho nezávislost na rozsahu výběru (počtu pozorování), přičemž se předpokládá přibližně normální rozdělení dat.



Obrázek 4: Výsledek IQR testu pro skupinu ALG zobrazené pro prvních 30 proměnných

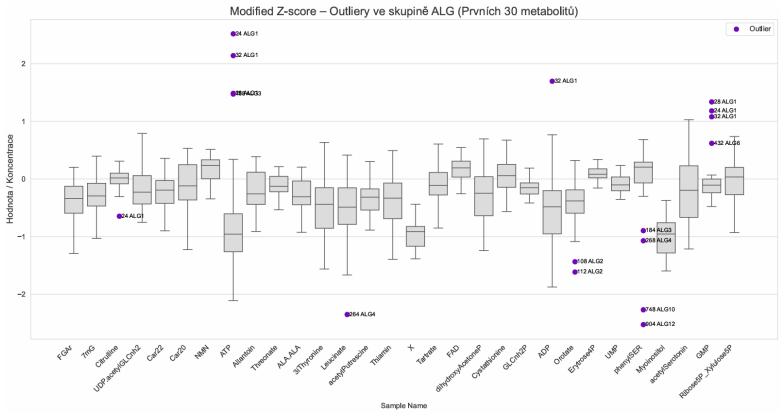
Pro detekci odlehlých hodnot analyzujeme každý metabolit samostatně. Statistika modifikovaného Z-skóre je pro daný metabolit v j-tém vzorku definována vztahem

$$M_j = \frac{0{,}6745 \cdot (X_j - \tilde{X})}{MAD},$$

kde X_j je náhodná veličina, jejíž realizace x_j odpovídá naměřené intenzitě daného metabolitu, $\tilde{X}$ představuje medián a MAD je mediánová absolutní odchylka (z angl. *Median Absolute Deviation*). Ta je definována jako $MAD = \text{median}(|X_j - \tilde{X}|)$. Konstanta 0,6745 odpovídá třetímu kvartilu (75 % kvantilu) standardního normálního rozdělení, ke kterému MAD asymptoticky konverguje.

Identifikace samotných odlehlých hodnot je následně založena na porovnání vypočtené realizace testové statistiky $|m_j|$ s doporučenou prahovou hodnotou 3,5. Konkrétní měření, která tuto hranici překročí, jsou klasifikována jako odlehlá.

Na obrázku 5 máme opět zobrazeno prvních 30 metabolitů s barevně označenými odlehlými hodnotami detekovanými metodou modifikovaného Z-skóre. Celkové výsledky čtenář najde v příloze *Modified_Z_score_outliers.pdf*.



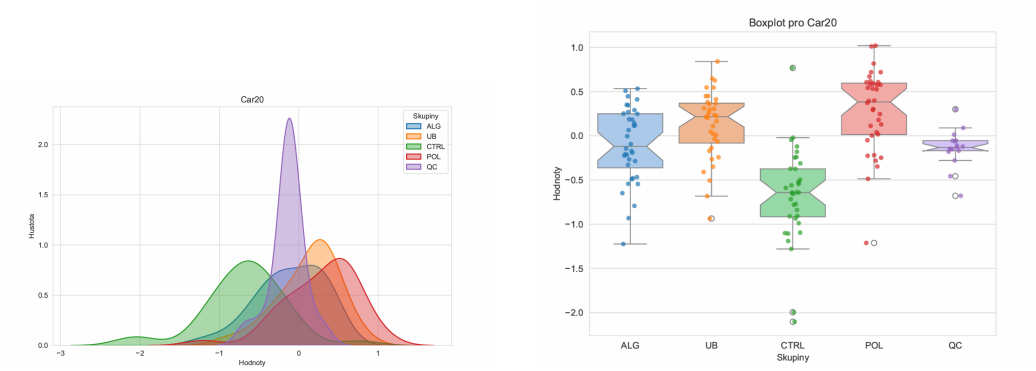
Obrázek 5: Výsledek modifikovaného Z-skóre testu pro skupinu *ALG* zobrazené pro prvních 30 proměnných

3.3 Charakteristiky

V popisné statistice by neměly chybět souhrnné charakteristiky. V datech se zaměřujeme na průměr, medián, směrodatnou odchylku, střední chybu průměru (počítanou jako směrodatná odchylka dělená druhou odmocninou z počtu měření, tedy $\frac{s}{\sqrt{n}}$), dolní a horní interval spolehlivosti a dolní a horní kvartil. Všechny tyto vypočítané hodnoty jsou k nahlédnutí v příloze. K těmto výpočtům jsou připojeny také grafické výstupy, jako jsou histogramy (levý obrázek 6a, příloha *Histograms.pdf*) a boxploty (pravý obrázek 6b, příloha *Boxplots.pdf*).

3.4 Korekce

Než se pustíme do samotného testování hypotéz, zavedeme si pojem korekce p-hodnot. Pokud nesprávně zamítneme nulovou hypotézu H_0, když ve skutečnosti platí, dopustíme se chyby I. druhu. Naopak, pokud nulovou hypotézu nezamítneme v situaci, kdy H_0 ve skutečnosti neplatí, dochází k chybě II. druhu. V tabulce 2 máme shrnutí obou scénářů. Tyto chyby se v praxi často označují také jako falešně pozitivní výsledky (chyba I. druhu) a falešně nega-



(a) Histogram (b) Boxploty

Obrázek 6: Grafické výstupy charakteristik

tivní výsledky (chyba II. druhu). Míra chyby I. druhu je pak definována jako pravděpodobnost, že se této chyby dopustíme. V ideálním případě bychom chtěli minimalizovat pravděpodobnosti obou chyb současně. Pokud ale zmenšíme pravděpodobnost chyby I. druhu, obvykle nám naroste pravděpodobnost chyby II. druhu a naopak. V praxi bereme chybu I. druhu jako závažnější, proto se zaměříme na její minimalizaci. Zároveň se ale snažíme, aby i pravděpodobnost chyby II. druhu zůstala malá. Více informací lze nalézt v [17].

		Skutečnost	
		H_0 platí	H_0 neplatí (H_a)
Rozhodnutí	Zamítnout H_0	Chyba I. druhu	Správný závěr
	Nezamítnout H_0	Správný závěr	Chyba II. druhu

Tabulka 2: Přehled možných scénářů souvisejících s testováním H_0

Při vícenásobném testování zamítáme mnoho platných hypotéz a dopouštíme se tak mnoha chyb I. druhu. Z tohoto důvodu zavádíme korekci p-hodnoty, která zohledňuje počet testů. Uvedeme si zde dvě metody. V Pythonu využíváme funkci `multipletests`, s parametrem `method="bonferroni"` nebo `method="fdr_bh"`.

3.4.1 Bonferroniho metoda

Při vícenásobném testování se využívá koncept FWER (z angl. *Family-wise error rate*), který zobecňuje pojem chyby I. druhu. FWER je formálně definována jako pravděpodobnost, že se při testování souboru hypotéz dopustíme alespoň jedné chyby I. druhu. Nejznámějším zástupcem metod pro kontrolu FWER je Bonferroniho korekce. Předpokládejme, že chceme souběžně otestovat m nulových hypotéz $H_{01}, \ldots, H_{0m}$ na celkové hladině významnosti α. Bonferroniho metoda upravuje práh pro zamítnutí každé dílčí hypotézy tak, že individuální testy jsou vyhodnocovány na přísnější hladině významnosti α/m. Tím je zaručeno, že celková pravděpodobnost chyby I. druhu (FWER) nepřekročí hodnotu α. Tato metoda je však velmi konzervativní; zpřísněný práh způsobuje, že test zamítá jen velmi málo nulových hypotéz. To vede k nárůstu pravděpodobnosti chyby II. druhu.

3.4.2 Benjamini-Hochbergova metoda

Další přístup k řešení problému vícenásobného testování představuje kontrola *False Discovery Rate* (FDR), tedy kontrola očekávaného podílu chybně zamítnutých nulových hypotéz mezi všemi zamítnutými hypotézami. Na rozdíl od FWER neexistuje pro FDR univerzálně stanovený práh (jako je hladina α), ale využívá se parametr q, který vyjadřuje maximální přípustnou míru falešných objevů (falešně signifikantních metabolitů). Při kontrole FDR na hladině q = 20 % algoritmus maximalizuje počet zamítnutých nulových hypotéz a zároveň garantuje, že v průměru nebude více než 20 % z těchto odhalených objevů falešně pozitivních. Volba konkrétní hodnoty q pak závisí na povaze zkoumaného problému a požadované přísnosti testu.

K efektivní kontrole FDR při testování m hypotéz se nejčastěji využívá Benjamini-Hochbergova metoda. Ta funguje následovně.

1. Určíme úroveň q, na které chceme kontrolovat FDR.

2. Napočítáme p-hodnoty, $p_1, \ldots, p_m$, pro m nulových hypotéz.

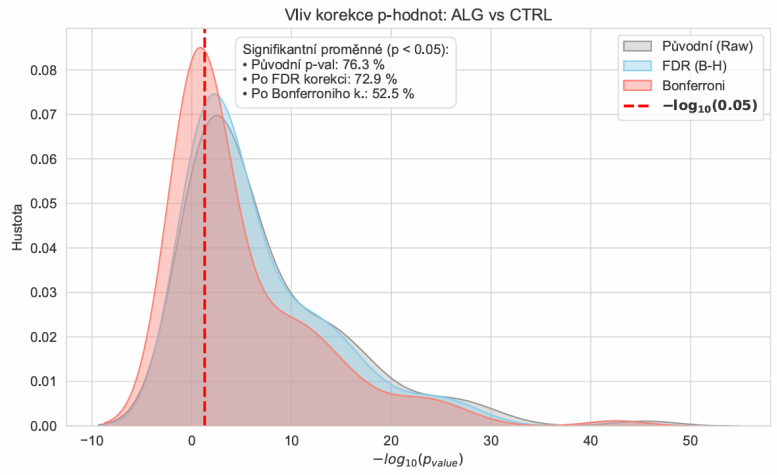
3. Vypočtené p-hodnoty seřadíme vzestupně: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

4. Definujme L jako
$$L = \max\{j : p_{(j)} < qj/m\}.$$

5. Zamítáme všechny H_{0j}, pro které platí $p_j < p_{(L)}$.

Na obrázku 7 můžeme vidět srovnání klasických p-hodnot s hodnotami po aplikaci korekcí na vícenásobné testování v případě dvouvýběrového t-testu. Osa x reprezentuje transformované p-hodnoty ve tvaru $-\log_{10}(p)$ a osa y zobrazuje jejich hustotu. Červená čárkovaná vertikála vyznačuje hraniční hodnotu odpovídající hladině významnosti $p = 0{,}05$ v původním měřítku. Proměnné, jejichž hodnoty leží vpravo od této čáry, vykazují prokazatelný rozdíl mezi testovanými skupinami ALG a $CTRL$, zatímco proměnné vlevo jsou statisticky nevýznamné. To znamená, že čím více vpravo data jsou, tím více jsou příslušné markery statisticky signifikantní.

Graf se skládá ze tří překrývajících se křivek znázorňujících rozdělení hodnot v logaritmickém měřítku: šedá křivka odpovídá původním nekorigovaným p-hodnotám, modrá křivka představuje mírnější Benjamini-Hochbergovu korekci (FDR) a světle červená křivka ukazuje nejstriktnější Bonferroniho korekci. Z vizualizace je patrné, že aplikace korekcí posouvá rozdělení hodnot směrem doleva, čímž se snižuje počet proměnných, které překonají hranici signifikance. Přesný procentuální podíl proměnných, které pro danou metodu zůstaly statisticky významné, je shrnut v informačním boxu v levé horní části grafu.



Obrázek 7: Porovnání p-hodnot

3.5 Parametrické testování hypotéz

V této sekci se zaměříme na testování shody středních hodnot ve dvou nebo více skupinách (uvažujeme přitom pouze oboustranné testy). Budeme se zabývat parametrickými testy, které lze využít za předpokladu normálního rozdělení dat. V implementaci v Pythonu vycházíme z knihovny *scipy.stats*. Jelikož se u čtenáře předpokládá základní znalost statistiky, je zde uveden pouze stručný přehled těchto testů. Podrobnější informace lze nalézt v literatuře [12].

V tabulce 3 se nachází přehled daných testů.

TEST	**PŘEDPOKLAD**		
	normalita	shoda rozptylů	nezávislost výběrů
Dvouvýběrový t-test	✓	✓	✓
Welchův dvouvýběrový t-test	✓	×	✓
Párový t-test	✓	×	×

Tabulka 3: Přehled parametrických testů

3.5.1 Dvouvýběrový t-test

U dvouvýběrového t-testu předpokládáme, že obě nezávislé náhodné veličiny mají normální rozdělení a shodný konečný rozptyl. Shodu rozptylů můžeme ověřit pomocí Leveneova testu. Testová statistika je založena na odhadu rozdílu středních hodnot, který se provádí pomocí rozdílu výběrových průměrů. Za platnosti nulové hypotézy se testová statistika řídí Studentovým t-rozdělením. Testování probíhá pro každou proměnnou mezi všemi možnými dvojicemi skupin. K výpočtům v jazyce Python využíváme funkci `ttest_ind` s parametrem `equal_var=True`.

3.5.2 Welchův dvouvýběrový t-test

Welchův t-test je modifikací klasického t-testu a používá se v případě, kdy není splněn předpoklad shody rozptylů. Ostatní předpoklady a základní princip testování zůstávají stejné. V Pythonu opět využíváme funkci `ttest_ind`, tentokrát ale s parametrem nastaveným na `equal_var=False`.

3.5.3 Párový t-test

Jak vyplývá z názvu metody, párový t-test se používá pro dva závislé výběry se spojitými veličinami. Základním předpokladem je, že rozdíly těchto párových hodnot mají normální rozdělení. Typickým příkladem závislých výběrů je měření krevního tlaku před a po podáním určitého léku. Nulová hypotéza předpokládá, že skutečný rozdíl středních hodnot je roven nule. Uvedenou hypotézu ověřujeme jednovýběrovým t-testem aplikovaným na zmíněné diference. V Pythonu využíváme pro tento výpočet funkci `ttest_rel`.

3.5.4 ANOVA

V předešlých testech jsme uvažovali pouze dvě skupiny. Pro testování shody středních hodnot u více než dvou skupin se využívá jednofaktorová analýza rozptylu, zkráceně ANOVA. Tato metoda předpokládá normalitu dat v jednotlivých skupinách a shodu jejich rozptylů. Jedná se v podstatě o zobecnění dvouvýběrového t-testu.

Model ANOVY můžeme napsat následovně

$$Y_{i,j} = \mu + \alpha_j + \epsilon_{i,j}, \tag{1}$$

kde $Y_{i,j}$ je náhodná veličina představující naměřenou intenzitu v i-tém pozorování j-té skupiny ($i = 1, \ldots, n_j$; $j = 1, \ldots, k$), μ je celková střední hodnota a α_j je efekt příslušnosti k j-té skupině. Člen $\epsilon_{i,j}$ představuje náhodnou chybu, u které se předpokládá normální rozdělení $N(0, \sigma^2)$.

Testová statistika je založena na porovnání dvou různých odhadů rozptylu σ^2. Odhad vycházející z variability mezi skupinami má podobu $\frac{S_A}{k-1}$ a odhad založený na variabilitě uvnitř skupin má podobu $\frac{S_e}{n-k}$. Za platnosti nulové hypotézy dávají oba odhady přibližně stejné hodnoty a jejich podíl se blíží k jedné. Testová statistika

$$F_A = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{S_A}{S_e}\frac{n-k}{k-1}$$

má za platnosti H_0 Fisherovo rozdělení $F_{k-1,n-k}$. Nulovou hypotézu H_0 zamítáme na hladině významnosti α pro dostatečně velké hodnoty této statistiky. V jazyce Python se pro tento výpočet používá funkce `f_oneway`.

V případě zamítnutí nulové hypotézy o shodě středních hodnot na hladině významnosti α je třeba rozhodnout, které konkrétní skupiny se od sebe liší.

Za tímto účelem použijeme Tukeyho metodu pro mnohonásobné porovnávání. Touto metodou zamítáme hypotézu o rovnosti středních hodnot pro dvojici skupin i a j, jestliže pro jejich vypočtené průměry $\bar{y}_i$ a $\bar{y}_j$ platí:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > s \cdot q_{k,n-k;1-\alpha} \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

kde $q_{k,n-k;1-\alpha}$ je příslušný kvantil studentizovaného rozpětí a s představuje odmocninu nestranného odhadu rozptylu (vypočteného v rámci ANOVY). V případě stejně velkého rozsahu všech výběrů $r = n_1 = \cdots = n_k$ se výše uvedený vztah zjednoduší na tvar

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > q_{k,n-k;1-\alpha} \frac{s}{\sqrt{r}}.$$

V jazyce Python se pro tento výpočet používá funkce `tukey_hsd`, která vychází z knihovny *statsmodels*.

3.5.5 ANCOVA

Analýza kovariance (ANCOVA) je statistická metoda, která kombinuje vlastnosti a principy analýzy rozptylu a lineární regresní analýzy. Používá se k testování nulové hypotézy o shodě středních hodnot ve třech nebo více skupinách. ANCOVA vždy zahrnuje alespoň tři proměnné: nezávislou proměnnou (faktor), závislou proměnnou a kovariát. Kovariát je proměnná, která přímo nesouvisí s hlavním cílem analýzy, ale je pravděpodobně korelovaná se závislou proměnnou. Může jít o spojitou nebo kategoriální proměnnou, jako je například věk, krevní tlak, BMI či pohlaví. V této podkapitole čerpáme z literatury Hebáka [11] a Huitemy [13].

Hlavní motivací pro použití ANCOVY je snížení zkreslení mezi nezávis-

lými a závislými proměnnými. Cílem metody je spravedlivě porovnat skupiny, které do analýzy vstupují s odlišnými výchozími parametry (mají například různý věk). Přidáním kovariátů do modelu dochází k vysvětlení a následnému odstranění části rozdílů mezi pacienty ve stejné skupině (reziduální variability). Výsledkem je menší nevysvětlená chyba a zvýšení statistické síly testu, což umožňuje prokázat statisticky významné rozdíly i tam, kde by standardní ANOVA selhala.

Analýza kovariance vysvětluje celkovou variabilitu v datech pomocí dvou hlavních složek. První složku tvoří rozdíly mezi samotnými skupinami (α_j), druhou pak lineární vliv kovariátu ($\beta(X_{i,j} - \overline{X}_{..})$. Statistický model lze zapsat ve tvaru

$$Y_{i,j} = \mu + \alpha_j + \beta(X_{i,j} - \overline{X}_{..}) + \epsilon_{i,j}, \tag{2}$$

kde

- $Y_{i,j}$ je hodnota závislé proměnné i-tého jedince v j-té skupině,
- μ je celkový populační průměr závislé proměnné,
- α_j je specifický účinek j-té skupiny (faktoru),
- β je regresní koeficient vyjadřující lineární závislost Y na kovariátu X,
- $X_{i,j}$ je hodnota kovariátu pro i-tého jedince v j-té skupině,
- $\overline{X}_{..}$ je celkový průměr hodnot kovariátů napříč všemi vzorky,
- $\epsilon_{i,j}$ představuje náhodnou chybu s rozdělením $N(0, \sigma^2)$.

Lze si všimnout, že rozšířením standardního modelu ANOVA (1) o regresní člen vzniká model ANCOVA (2).

Pro úspěšnou aplikaci tohoto modelu musí zkoumaná data splňovat následující podmínky: nezávislost výběrů, normální rozdělení Y, homoskedasticitu, lineární závislost mezi X a Y a shodné regresní koeficienty ve všech skupinách.

Než přistoupíme k hodnocení samotného vlivu faktoru, algoritmus ANCOVY postupně ověřuje tři hypotézy v následujícím pořadí.

1. Ověříme $H_{01} : \beta_1 = \beta_2 = \cdots = \beta_k = \beta$, tedy shodu regresních koeficientů ve skupinách. Pokud by se významně lišily, standardní model ANCOVY není adekvátní.

2. Ověříme $H_{02} : \beta = 0$, tedy nulovou hodnotu celkového regresního koeficientu. Neprokážeme-li přítomnost lineární závislosti Y na X, nemá provedení ANCOVY význam a závislost můžeme posoudit podle standardní ANOVY.

3. Ověříme $H_{03} : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, neboli shodu skupinových průměrů vysvětlované proměnné Y. Očištěné rozdíly v závislé proměnné (tedy ty, ze kterých byl odečten vliv předpovídaný pomocí kovariátu X) se porovnají mezi skupinami, což je hlavní náplní analýzy kovariance. Pokud tuto nulovou hypotézu zamítneme, pokládáme vliv faktoru (příslušnosti ke skupině) na proměnnou Y za prokázaný na zvolené hladině významnosti α.

Pro samotné otestování hlavní hypotézy H_{03} (zda se skupiny liší po očištění o vliv kovariátu) je nutné nejprve spočítat tzv. korigované (opravené) součty čtverců. Tyto hodnoty představují variabilitu dat očištěnou o vliv doprovodné proměnné X:

- celkový opravený součet čtverců: $S_T(Y^*) = S_T(Y) - \frac{[S_T(X,Y)]^2}{S_T(X)}$,

- reziduální opravený součet čtverců: $S_e(Y^*) = S_e(Y) - \frac{[S_e(X,Y)]^2}{S_e(X)}$,
- meziskupinový opravený součet čtverců: $S_A(Y^*) = S_T(Y^*) - S_e(Y^*)$.

Na základě těchto opravených součtů čtverců je následně sestavena testová statistika F, která je základem výstupu analýzy kovariance (a ze které se odvozuje výsledná p-hodnota). Vypočítá se jako podíl opraveného meziskupinového a reziduálního rozptylu

$$F = \frac{S_A(Y^*)/(k-1)}{S_e(Y^*)/(n-k-1)},$$

kde k je počet porovnávaných skupin a n je celkový počet subjektů. V případě, že je tato F-statistika významná a dojde k zamítnutí nulové hypotézy H_{03}, nelze pro interpretaci výsledků porovnávat běžné (naměřené) průměry skupin, neboť ty jsou stále zatíženy zkreslením ze strany kovariátů. Model proto počítá tzv. korigované (upravené) průměry.

Korigovaný průměr je predikovaná hodnota. Představuje průměrnou hodnotu závislé proměnné, která by se u dané skupiny očekávala v případě, že by se její průměr u zvoleného kovariátu přesně shodoval s celkovým průměrem (tzv. *grand mean*) celého zkoumaného vzorku. Vzorec pro výpočet korigovaného průměru je

$$\overline{Y}_{j,\text{adj}} = \overline{Y}_j - \beta(\overline{X}_j - \overline{X}_{..}), \tag{3}$$

kde $\overline{Y}_{j,\text{adj}}$ je korigovaný průměr j-té skupiny, $\overline{Y}_j$ je původní (neupravený) průměr, β je regresní koeficient (totožný s koeficientem v modelu (2)),
a $(\overline{X}_j - \overline{X}_{..})$ je rozdíl průměru kovariátů dané skupiny od celkového průměru kovariátů. Pokud předpokládáme rovnoběžnost regresních přímek, pak se zjištěný rozdíl mezi parametry u absolutního členu regrese obou skupin rovná přesně rozdílu mezi těmito korigovanými průměry.

Protože fyziologické parametry (např. věk) korelují se samotnou hladinou metabolitů, vzniká zásadní otázka: je průměrná hladina metabolitu u skupiny pacientů odlišná čistě v důsledku samotného onemocnění, nebo je tento rozdíl způsoben rozdílnými výchozími charakteristikami (např. vyšším věkem) pacientů?

Běžné (naměřené) průměry tuto otázku nedokážou zodpovědět, neboť jsou zatíženy zkreslením ze strany těchto kovariátů. Z tohoto důvodu model analýzy kovariance (ANCOVA) počítá tzv. korigované průměry, které matematicky vyrovnají vliv věku, BMI a pohlaví tak, jako by obě skupiny měly naprosto shodné výchozí podmínky. Následné post-hoc testování a výpočet statistické významnosti (p-hodnot) probíhá výhradně na těchto korigovaných průměrech. Tím je zajištěno, že zjištěné změny v metabolickém profilu souvisejí přímo se samotným onemocněním a nejsou jen důsledkem odlišného věku či tělesných parametrů pacientů.

V jazyce Python využíváme funkci `anova_lm` z modulu *statsmodels*. Samotný výpočet modelu probíhá ve dvou navazujících krocích. Nejprve je pomocí metody nejmenších čtverců (funkce `ols`) sestaven vícenásobný lineární regresní model. Tento model v sobě kombinuje jak hlavní sledovaný kategoriální faktor (příslušnost ke skupině nemocných nebo kontrol), tak všechny zvolené spojité i kategoriální kovariáty (věk, BMI a pohlaví). Na tento sestavený model je následně aplikována funkce `anova_lm`, která vygeneruje klasickou tabulku analýzy rozptylu s výslednými F-statistikami a p-hodnotami. Zásadním parametrem tohoto kroku je volba výpočtu součtů čtverců typu III (`typ=3`, tzv. *Type III Sum of Squares* neboli marginální součty čtverců). Tento typ výpočtu je nezbytný vzhledem k rozdílnému počtu subjektů v jednotlivých skupinách. Výpočet typu III zaručuje, že výsledná statistická významnost hlavního faktoru (diagnózy) je testována a vyhodnocena až poté,

co je z modelu matematicky odečten vliv všech ostatních přítomných kovariátů. Toto matematické očištění probíhá prostřednictvím výpočtu korigovaných průměrů definovaných výše ve vzorci (3). Tím je zajištěno, že výsledné p-hodnoty nejsou zkresleny pořadím, ve kterém byly proměnné do modelu vloženy. Přestože aktuální vstupní data doprovodné proměnné neobsahují, výpočet je připraven pro statistické zpracování libovolných metabolomických dat s kovariáty.

3.6 Neparametrické testování hypotéz

V případě, kdy nelze předpokládat normalitu rozdělení dat, testujeme shodu spojitých rozdělení mezi dvěma nebo více skupinami (uvažujeme oboustranný test). Teorii k následujícím metodám čerpáme z literatury [12] a [4] a v Pythonu používáme modul *scipy.stats*.

3.6.1 Mann-Whitneyův U test

Mann-Whitneyův U test (zkráceně MW-U) ověřuje nulovou hypotézu, že oba nezávislé výběry pocházejí ze stejného rozdělení. Nechť m a n jsou rozsahy těchto dvou výběrů. Všech $m + n$ pozorovaných hodnot nejprve sloučíme a uspořádáme vzestupně podle velikosti. Označme T_1 součet pořadí hodnot z prvního výběru a T_2 součet pořadí hodnot z druhého výběru. Poté vypočítáme hodnoty testových statistik

$$U_1 = mn + \frac{m(m+1)}{2} - T_1 \quad \text{a} \quad U_2 = mn + \frac{n(n+1)}{2} - T_2.$$

Nulovou hypotézu zamítáme na hladině α, je-li hodnota $\min(U_1, U_2)$ menší nebo rovna příslušné kritické hodnotě. Pro výpočty v jazyce Python využíváme funkci `mannwhitneyu`.

3.6.2 Wilcoxonův test

Wilcoxonův jednovýběrový test (z angl. *Wilcoxon signed-rank test*) představuje neparametrickou alternativu k párovému t-testu (Conover [4]). Využíváme jej k testování závislých výběrů v případech, kdy diference nesplňují předpoklad normality.

Nejprve se pro každý z n párů vypočítají diference $D_i = Y_i - X_i$ a jejich absolutní hodnoty $|D_i|$. Z dalšího postupu vyřadíme všechny páry s nulovou diferencí (tj. $X_i = Y_i$, respektive $D_i = 0$). Počet zbývajících nenulových diferencí označme n', kde $n' \leq n$. Těmto n' absolutním diferencím $|D_i|$ přiřadíme vzestupně pořadí od 1 do n'. Pokud se některé absolutní diference rovnají, přiřadí se každé z nich průměrné pořadí, které by jim jinak bylo přiděleno.

Předpoklady testu jsou následující.

1. Rozdělení každé diference D_i je symetrické.
2. Veličiny D_i jsou vzájemně nezávislé.
3. Všechny veličiny D_i mají stejnou střední hodnotu.
4. Měřítko veličin D_i je alespoň intervalové.

Následně definujeme pořadí R_i tak, že původnímu pořadí absolutní diference vrátíme její původní znaménko. Získáme tak kladná a záporná pořadí. Testová statistika T^+, resp. T^- je pak definována jako součet všech kladných, resp. záporných pořadí

$$T^+ = \sum_{D_i>0} R_i, \qquad T^- = \sum_{D_i<0} R_i.$$

Nulovou hypotézu zamítáme na hladině významnosti α, pokud je statistika $T = \min(T^+, T^-)$ menší nebo rovna příslušné kritické tabulkové hodnotě. V jazyce Python využíváme k těmto výpočtům funkci `wilcoxon`.

3.6.3 Kruskal-Wallisův test

Kruskal-Wallisův test slouží k testování shody rozdělení ve třech a více skupinách. Jedná se o neparametrickou alternativu k jednofaktorové analýze rozptylu (ANOVA) a zároveň o zobecnění Mann-Whitneyho U testu. Testujeme hypotézu, že všechny nezávislé výběry pocházejí ze stejného rozdělení.

Při výpočtu nejprve sloučíme a seřadíme všech n pozorování do neklesající posloupnosti a určíme pořadí každé hodnoty. V případě, že se vyskytují shodná pozorování, přiřadíme jim průměrné pořadí. Označme T_i jako součet pořadí těch čísel, která patří do i-tého výběru pro $i = 1, \ldots, k$. Za platnosti nulové hypotézy H_0 má testová statistika

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

asymptoticky (při $n_i \to \infty, \forall i = 1, \ldots, k$) rozdělení χ^2 s $k-1$ stupni volnosti. Kritický obor je tedy vymezen intervalem $W = (\chi^2_{1-\alpha}(k-1), +\infty)$. Nulovou hypotézu zamítáme na hladině významnosti α, pokud realizace statistiky Q spadá do tohoto kritického oboru. V jazyce Python využíváme pro tento výpočet funkci `kruskal`.

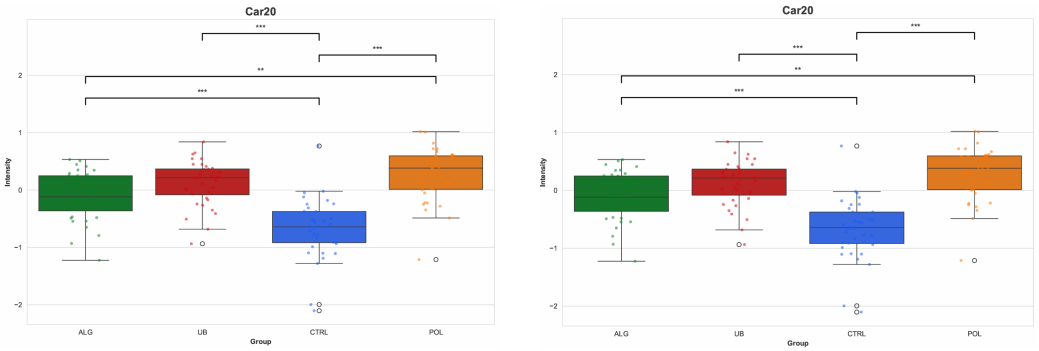
V případě zamítnutí nulové hypotézy Kruskal-Wallisova testu nelze pro následné mnohonásobné porovnávání využít standardní Tukeyho test, neboť zkoumaná data nesplňují předpoklad normálního rozdělení. K identifikaci konkrétních dvojic skupin, které se od sebe statisticky významně liší, se proto využívá neparametrický Dunnův post-hoc test. Ten vyhodnocuje rozdíly v průměrném pořadí naměřených hodnot. Tím umožňuje určit, zda je celkové rozložení dat mezi dvěma konkrétními skupinami statisticky odlišné.

3.7 Výsledky testování hypotéz

Kompletní výsledky všech provedených testů jsou přiloženy v excelovské tabulce *Univariate_ results.xlsx*. V analyzovaných datech se nenacházejí žádné závislé (párové) výběry, proto tabulka neobsahuje výsledky párových testů. Vytvořený skript je však na zpracování závislých dat připraven a v případě odpovídající datové sady by byly výsledky uloženy totožným způsobem. Výsledná tabulka obsahuje názvy porovnávaných skupin, hodnoty testových statistik, původní p-hodnoty a také korigované p-hodnoty podle Benjamini-Hochbergovy a Bonferroniho metody. V případě ANOVA testu a Kruskal-Wallisova testu jsou navíc specifikovány statisticky významné dvojice skupin identifikované na základě Tukeyho a Dunnova post-hoc testu. Výsledky všech předchozích testů, ať už parametrických (kromě ANCOVY) či neparametrických, lze přehledně vizualizovat pomocí boxplotů s vyznačením statistické významnosti rozdílů mezi jednotlivými skupinami. Jako ukázka je zde prezentován výsledek analýzy rozptylu (levý obrázek 8a) a Kruskal-Wallis test (pravý obrázek 8b) s následnými post-hoc testy. Grafy pro ostatní testy a proměnné jsou uvedeny v PDF přílohách. Na obrázku 8 vidíme boxploty pro proměnnou *Car20*. Hvězdičky nad spojnicemi indikují, mezi kterými konkrétními dvojicemi skupin byl nalezen statisticky významný rozdíl. Počet těchto hvězdiček pak vyjadřuje dosaženou hladinu významnosti: jedna hvězdička značí $p < 0{,}05$, dvě hvězdičky $p < 0{,}01$ a tři hvězdičky představují vysoce statisticky významný rozdíl s $p < 0{,}001$.

3.8 Korelační analýza

Korelační analýza se zabývá zkoumáním vztahů mezi proměnnými. V této práci uvádíme dvě nejčastěji využívané metody, přičemž teoreticky vycházíme



(a) Výsledek analýzy rozptylu (b) Výsledek Kruskal-Wallise

Obrázek 8: Výsledky testů pro proměnnou *Car20*

z publikace [12]. Výběrový Pearsonův korelační koeficient R měří sílu a směr lineární závislosti.

Pro testování nulové hypotézy, že se skutečný korelační koeficient rovná nule ($H_0 : \rho = 0$), převádíme výběrový koeficient R na testovou statistiku T. Tu získáme ze vzorce

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}},$$

kde n je rozsah výběru. Získaná hodnota t se následně porovnává s kritickými hodnotami Studentova t-rozdělení o $n-2$ stupních volnosti.

Naopak, výběrový Spearmanův korelační koeficient R_s hodnotí těsnost závislosti na základě pořadí jednotlivých hodnot, čímž se stává robustnějším vůči odlehlým pozorováním a nevyžaduje striktní předpoklad linearity, nýbrž pouze obecnější předpoklad monotónního vztahu. Pro dostatečně rozsáhlé výběry (zpravidla pro $n > 30$) lze k testování statistické významnosti využít větu o asymptotické normalitě. Testová statistika Z má v takovém případě tvar

$$Z = R_S\sqrt{n-1}.$$

Realizace testové statistiky z se následně porovnává s kritickými hodnotami

standardního normálního rozdělení $N(0,1)$.

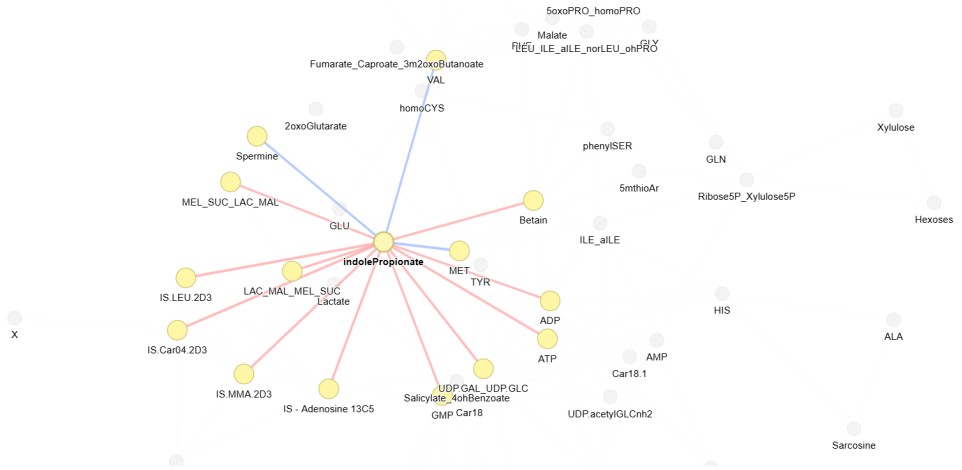
V obou zmíněných metodách je navíc nezbytné aplikovat korekci výsledných p-hodnot kvůli problému vícenásobného testování.

Vzhledem k vysokému počtu analyzovaných proměnných nebylo možné prezentovat klasickou heatmapu. Jako alternativní a přehlednější řešení pro vizualizaci výsledků byly proto zvoleny interaktivní korelační sítě a barevně formátované matice exportované do excelové tabulky.

Výsledná korelační tabulka je rozdělena na dvě části. Dolní trojúhelníková matice obsahuje samotné hodnoty korelačních koeficientů. Tyto hodnoty jsou vizualizovány pomocí barevné škály, kde modrá barva indikuje pozitivní korelaci, zatímco červená barva značí korelaci negativní (přičemž intenzita barvy odpovídá síle závislosti). Silné korelace, pro které platí (např.) $|r| > 0{,}8$, jsou navíc hodnoty zvýrazněny tučným písmem. Horní trojúhelníková matice pak obsahuje příslušné p-hodnoty, jež slouží k posouzení statistické významnosti jednotlivých vztahů. Pokud je zjištěná p-hodnota menší než předem stanovená hladina významnosti $\alpha = 0{,}05$, je v tabulce rovněž vyznačena tučně.

Jako další metoda grafické vizualizace byly využity také korelační sítě. Tento přístup je obzvláště vhodný a přehledný v případech s vysokou dimenzionalitou dat. Základem těchto sítí jsou uzly, které reprezentují jednotlivé proměnné (metabolity), a hrany, které vizualizují zjištěné statisticky významné vztahy mezi nimi. Barevné kódování hran odpovídá směru zjištěné korelace: modrá barva indikuje pozitivní vztah, zatímco červená barva značí vztah negativní. Sítě jsou vždy generovány odděleně pro jednotlivé zkoumané skupiny pacientů. Příklad takové vizualizace pro proměnnou *indolePropionate* ve skupině ALG je znázorněn na obrázku 9.

Pro detailnější analýzu byl s podporou nástrojů umělé inteligence [8] vytvořen interaktivní HTML report. Tento nástroj umožňuje komplexní vizu-



Obrázek 9: Výběrový Pearsonův korelační koeficient pro skupin *ALG*

alizaci sítí založených jak na Pearsonově, tak na Spearmanově korelačním koeficientu. Uživatelské rozhraní nabízí přepínání mezi jednotlivými skupinami. Významným prvkem interaktivity je možnost výběru konkrétního uzlu; po vybrání určité proměnné se síť automaticky přeskupí a zvýrazní pouze ty hrany, které s daným uzlem přímo souvisí. Současně je vygenerována přehledná tabulka obsahující přesné hodnoty korelačních koeficientů a odpovídajících p-hodnot pro daný výběr. Kompletní sada těchto interaktivních reportů je čtenáři k dispozici v přílohách *Pearson_networks.html* a *Spearman_networks.html* této práce.

Pro výpočet samotných korelačních matic byla využita knihovna *pandas* a její funkce `corr` s příslušným parametrem pro volbu metody (např. `method="pearson"` či `method="spearman"`). Pro vizualizaci a generování interaktivních HTML reportů byly využity knihovny `json` a `networkx`.

3.9 ROC analýza

ROC křivka (z angl. *Receiver Operating Characteristic*) je populární grafický nástroj pro současné zobrazení obou typů chyb klasifikace napříč všemi mož-

nými prahovými hodnotami. V této kapitole čerpáme ze zdrojů [7], [17], [22] a [18].

Souhrnnou charakteristikou ROC křivky je plocha pod křivkou, tzv. *area under the curve* (AUC), která ukazuje celkový výkon klasifikátoru shrnutý přes všechny možné prahové hodnoty. Ideální ROC křivka se bude blížit k hornímu levému rohu. Platí čím větší hodnota AUC, tím lepší klasifikátor. Očekáváme, že klasifikátor, který nevykazuje lepší výkon než náhoda, bude mít hodnotu AUC = 0,5 (pokud je hodnocen na nezávislé testovací sadě, která nebyla použita při trénování modelu).

ROC křivky jsou užitečné pro porovnávání různých klasifikátorů, protože zohledňují všechny možné prahové hodnoty. Změna prahu klasifikátoru mění jeho míru pravdivě pozitivních a falešně pozitivních výsledků. Tyto metriky jsou také označovány jako senzitivita a (1 – specificita) našeho klasifikátoru. ROC křivka tedy vizualizuje vztah mezi těmito dvěma chybami: na ose y je vynesena míra skutečně pozitivních nálezů (senzitivita: podíl správně identifikovaných případů při dané prahové hodnotě) a na ose x míra falešně pozitivních nálezů ((1 – specificita): podíl nesprávně klasifikovaných případů při stejné prahové hodnotě).

V případě metabolomických dat neuvažujeme složitý klasifikační model, ale hodnotíme schopnost samotné koncentrace jednotlivého biomarkeru klasifikovat stav onemocnění pomocí analýzy ROC. Pro nalezení optimální prahové hodnoty využíváme Youdenův index (J), který představuje maximální potenciální účinnost biomarkeru a je běžnou souhrnnou mírou ROC křivky [22]. Tento index je formálně definován jako

$$J = \max_{c}\{Se(c) + Sp(c) - 1\}.$$

Prahová hodnota, která tohoto maxima dosahuje, se označuje jako opti-

mální (c^*), neboť optimalizuje diskriminační schopnost biomarkeru za předpokladu, že senzitivitě (Se) i specificitě (Sp) je přikládána stejná váha.

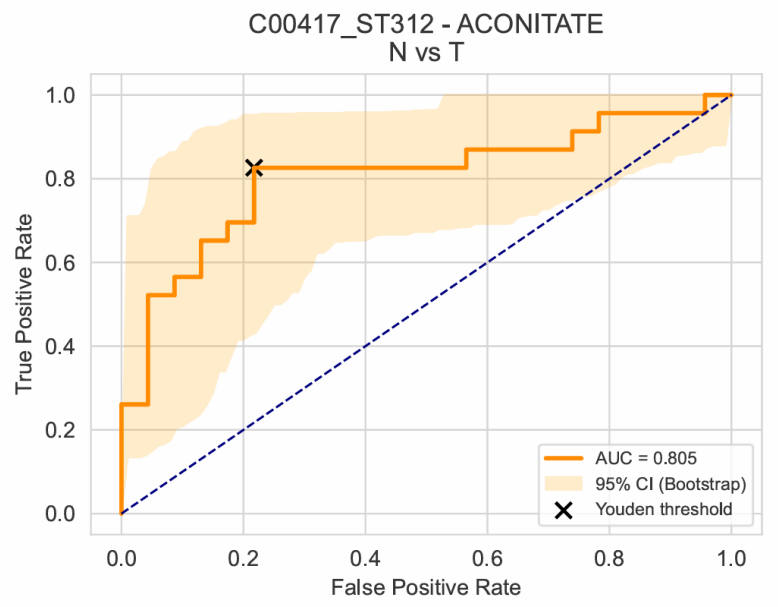
Sestrojení samotné ROC křivky a výpočet celkové plochy pod křivkou (AUC) jsou provedeny pomocí funkce `roc_curve` a `auc` z knihovny *scikit-learn*. Jelikož je teoretické rozdělení tohoto empirického odhadu neznámé, odhadované standardní chyby a 95% intervaly spolehlivosti získáme pomocí základní percentilové bootstrap metody. Tento přístup spočívá v neparametrickém opakovaném výběru dat, kdy jsou všechna pozorování u pacientů i zdravých kontrol nezávisle vybírána s opakováním. Pro každý metabolit se provede 1000 bootstrapových výběrů (náhodných výběrů s opakováním). Výsledné 95% intervaly spolehlivosti pro AUC i pro konfidenční pás ROC křivky zkonstruujeme na základě percentilu 2,5 a 97,5 z hodnot získaných z těchto výběrů [22].

K otestování statistické významnosti diskriminační schopnosti jednotlivých metabolitů je formulována nulová hypotéza, že AUC je rovna $0,5$ (tj. biomarker nemá lepší rozlišovací schopnost než náhodný odhad) [18]. Testování probíhá pomocí asymptotického Z-testu definovaného

$$Z = \frac{AUC - 0,5}{SE_{\mathrm{AUC}}}$$

s využitím bootstrapové směrodatné odchylky (SE_{AUC}) a kde AUC představuje pozorovanou plochu pod křivkou. Tato veličina má asymptoticky standardní (normované) normální rozdělení $N(0,1)$, tedy p-hodnota odpovídající oboustranné alternativě bude ve tvaru $p = 2 \cdot [1 - \Phi(|Z|)]$, [12].

Pro usnadnění biologické interpretace je u každého metabolitu rovněž hodnocen směr rozdílu mezi skupinami. Tento směr (určující, u které ze skupin vykazuje metabolit celkově vyšší hodnoty) je dedukován na základě toho, zda se hodnota AUC nachází nad nebo pod teoretickou hranicí 0,5. Kvůli vel-



Obrázek 10: ROC analýza

kému množství testovaných metabolitů je nutné korigovat p-hodnoty proti kumulaci chyby I. druhu (falešné pozitivity).

Vzhledem k tomu, že analyzovaná data hemolymfy neobsahují rozdělení na pacienty a zdravé kontroly, je pro ilustraci ROC analýzy využit ukázkový graf z jiného datového souboru (obrázek 10). Na ose x je vynesena (1 – specificita) a na ose y senzitivita. Oranžová křivka ukazuje výkonnost biomarkeru napříč všemi možnými prahovými hodnotami. Pásmo ve světlejším odstínu vizualizuje 95% konfidenční interval zkonstruovaný pomocí bootstrapu. Černým křížkem je na křivce vyznačen bod odpovídající Youdenovu indexu, který poskytuje nejlepší matematický kompromis mezi senzitivitou a specificitou. V tomto případě dosahuje Youdenův index hodnoty $J = 0,61$. Na základě jeho maximalizace byla stanovena optimální prahová hodnota biomarkeru na 0,66 (při senzitivitě 83 % a specificitě 78 %). Zjištěný směr rozdílu (N > T) indikuje, že u zdravých kontrol (N) jsou koncentrace tohoto metabolitu celkově vyšší než u pacientů (T). Z diagnostického hlediska

to znamená, že model predikuje přítomnost onemocnění v případě, že naměřená hodnota biomarkeru klesne pod tuto stanovenou optimální prahovou hodnotu.

3.10 Volcano plot

Volcano plot představuje kompaktní způsob vizualizace výsledků statistického testování rozdílů v koncentracích stovek či tisíců metabolitů. Jedná se o typ bodového grafu, kde každý bod reprezentuje jeden měřený prvek (např. metabolit) a ukazuje, jak se jeho koncentrace liší mezi dvěma testovanými skupinami.

Výpočet pro každou dvojici porovnávaných skupin probíhá ve třech hlavních krocích.

1. **Výpočet násobku změny (Fold Change):** Pro každý metabolit se určí střední hodnota (lze zvolit průměr nebo medián) v obou srovnávaných skupinách. Z těchto hodnot se spočítá $\log_2 FC$ jako rozdíl binárníchch logaritmů středních hodnot

$$\log_2 FC = \log_2(\bar{X}_A) - \log_2(\bar{X}_B),$$

   kde $\bar{X}_A$ a $\bar{X}_B$ jsou teoretické střední hodnoty (průměry či mediány) koncentrací daného metabolitu ve srovnávaných skupinách.

2. **Statistické testování:** Samotné testování probíhá pro každý metabolit zvlášť a využívá původní hodnoty. Podle možnosti předpokladu normality se provede buď parametrický Welchův t-test, nebo neparametrický Mann-Whitneyho U test.

3. **Korekce p-hodnot:** Vzhledem k problému vícenásobného testování

se v posledním kroku vypočítané p-hodnoty matematicky upravují (korigují), aby se minimalizoval počet falešně pozitivních výsledků, viz kapitola 3.4.

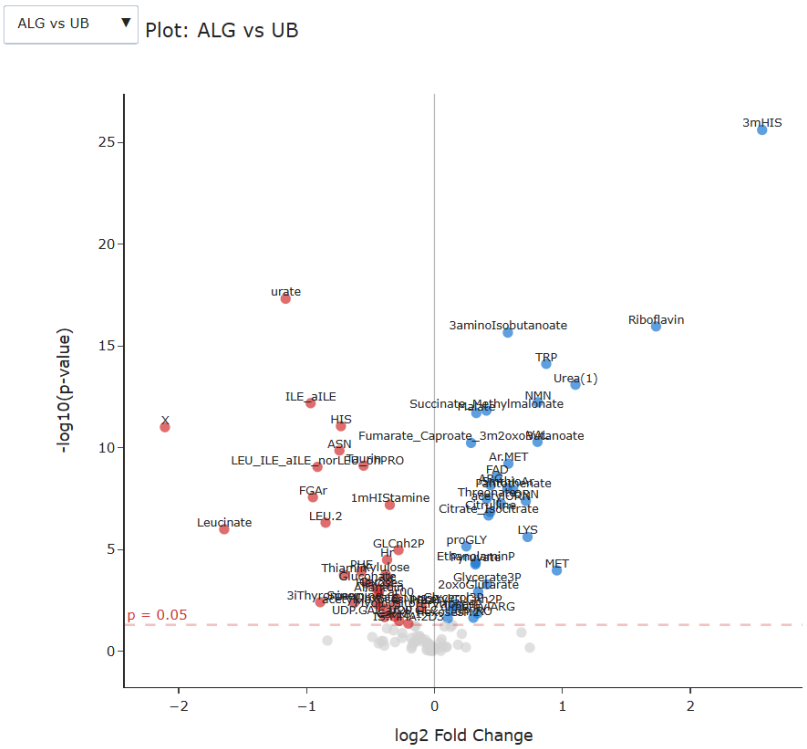
Po provedení těchto výpočtů jsou data zanesena do grafu. Podrobnější popis interpretace tohoto typu vizualizace v metabolomice lze nalézt v [14]. Osa y znázorňuje úroveň statistické významnosti jako $-\log_{10}(p)$. Statisticky nejvýznamnější metabolity jsou směrem nahoru (v levém či pravém horním rohu), protože mají nejnižší p-hodnoty. Osa x znázorňuje logaritmovaný Fold Change. Tato hodnota představuje poměr mezi koncentrací (hladinou) daného metabolitu ve dvou srovnávaných skupinách. Jak bylo zmíněno výše, pro lepší přehlednost v grafu využíváme binární logaritmus ($\log_2$), díky kterému je rozdělení symetrické.

- **Pravá strana (kladné hodnoty):** Pokud je hladina určitého metabolitu ve skupině A vyšší ve srovnání se skupinou B, bude $\log_2 FC$ pozitivní. Směrem doprava se tak nacházejí metabolity s relativně vyšší koncentrací ve skupině A.

- **Levá strana (záporné hodnoty):** Pokud je hladina určitého metabolitu ve skupině A nižší ve srovnání se skupinou B, bude $\log_2 FC$ negativní. Směrem doleva se nacházejí metabolity, které mají vyšší koncentraci ve skupině B ve srovnání se skupinou A.

- **Střed (nula):** Hodnota 0 na ose x znamená, že koncentrace metabolitu je v obou skupinách stejná.

Vzhledem k tomu, že data do analýzy typicky vstupují v přirozeném logaritmu, provádíme nejprve jejich zpětnou transformaci (odlogaritmování). Tím zajistíme, že vypočítaný $\log_2 FC$ bude správně reprezentovat poměr koncentrací metabolitů mezi srovnávanými skupinami.



Obrázek 11: Volcano plot

Obrázek 11 ukazuje Volcano plot vygenerovaný v Pythonu. V samotné práci je vložený jen PNG obrázek, ale v příloze čtenář najde i interaktivní soubor *Volcano.html*, vytvořený s pomocí AI [8]. V něm si může uživatel graf přiblížit nebo si prohlédnout tabulku významných metabolitů. Ty jsou v tabulce rozděleny podle toho, zda u nich došlo k nárůstu ($FC > 0$), nebo poklesu ($FC < 0$). Jak je patrné pro porovnání skupin ALG a UB, nejvýraznější izolovanou proměnnou je metabolit *3mHIS*. Tento bod se nachází v extrémní pravé horní části grafu, což ukazuje na vysoce statisticky významný rozdíl spojený se silným nárůstem koncentrace. Kladná hodnota logaritmického násobku změny indikuje, že hladina *3mHIS* je ve skupině ALG podstatně vyšší ve srovnání s kontrolní skupinou UB.

4 Vícerozměrná statistika

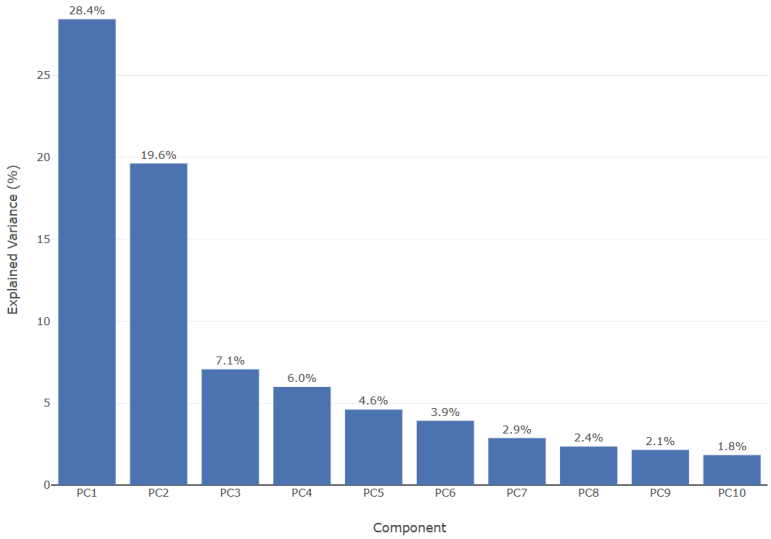
Poslední kapitola této práce se zabývá metodami z vícerozměrné statistiky, které jsou v tuto chvíli do workflow zaimplementovány (viz příloha *Multivariate_analysis.ipynb*). Jedná se o grafické nástroje pomocí PCA a shlukování. Podrobnější teorii lze nalézt v literatuře [17] a [28].

4.1 PCA

Analýza hlavních komponent (PCA) patří k základním nástrojům statistiky učení bez učitele (*unsupervised learning*). Protože jde o velmi známou a často používanou metodu, popíšeme si zde jen její hlavní principy.

Analýza hlavních komponent je metoda určená k redukci dimenze a vizualizaci vícedimenzionálních dat. Její princip spočívá v transformaci původních proměnných do nových, navzájem nekorelovaných latentních proměnných, označovaných jako hlavní komponenty. Tyto komponenty představují lineární kombinace původních p proměnných a jsou konstruovány tak, aby v datech zachytily co největší podíl celkové variability. PCA tak poskytuje nástroj pro nalezení nízkodimenzionální reprezentace dat (nejčastěji dvourozměrné), která nejlépe vystihuje rozdíly mezi jednotlivými pozorováními.

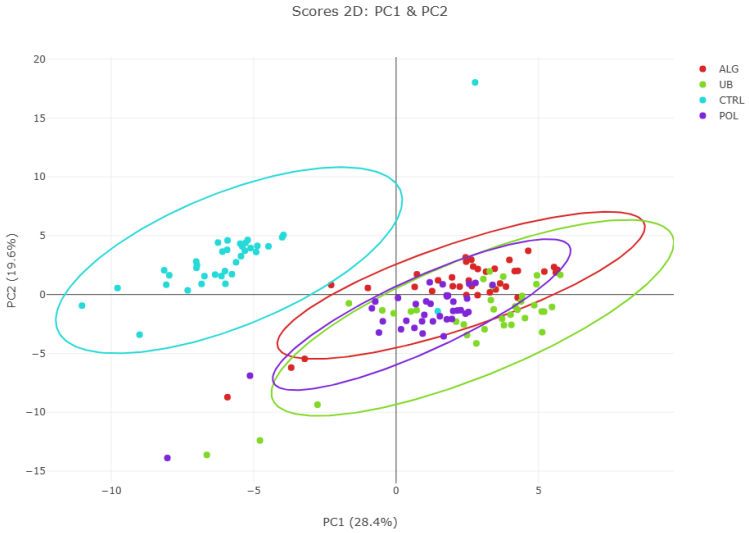
Výstupem PCA jsou dva typy informací: zátěže (tzv. *loadings*) a skóry (tzv. *scores*). Zátěže představují váhy, které určují vliv jednotlivých metabolitů na danou komponentu, zatímco skóry jsou souřadnice vzorků v tomto novém prostoru. Pro správnou funkci algoritmu je nezbytné data předem škálovat, viz kapitola 2.6, aby výsledky nebyly zkresleny. Důležitým ukazatelem je také podíl vysvětleného rozptylu (PVE z angl. *Proportion of variance explained*), který udává, kolik informací z původního souboru daná komponenta zachovává.



Obrázek 12: Scree plot

Pro určení kvality analýzy se využívá tzv. *scree plot*, který zobrazuje podíl vysvětleného rozptylu pro každou hlavní komponentu. Cílem je najít co nejmenší počet komponent, které v datech zachycují podstatnou část informace. V grafu hledáme tzv. „loket“, tedy bod, za kterým už další komponenty vysvětlují jen zanedbatelnou část celkové variability. Z obrázku 12 je patrné, že největší podíl variability vysvětlují první dvě komponenty, které dohromady pokrývají 48 % celkového rozptylu dat. Jelikož se po druhé komponentě nachází výrazný zlom a přínos dalších komponent je již nízký, byly pro následnou vizualizaci a interpretaci výsledků vybrány právě první dvě hlavní komponenty.

Na obrázku 13 vidíme graf skórů, který vizualizuje jednotlivá pozorování v novém prostoru hlavních komponent. Každý bod v grafu představuje jeden vzorek a jeho pozice odráží celkovou metabolomickou podobnost s ostatními – vzorky, které leží blízko sebe, mají v naměřených datech podobné vlastnosti. Při pohledu na graf lze identifikovat několik klíčových poznatků:



Obrázek 13: Graf skorů

- Skupina *CTRL* (světle modrá) se výrazně a jasně odděluje od ostatních tří skupin. To naznačuje, že kontrolní vzorky mají odlišný metabolomický profil než vzorky v ostatních skupinách.

- Skupiny *ALG* (červená), *UB* (zelená) a *POL* (fialová) vykazují značný vzájemný překryv a tvoří jeden velký shluk v pravé části grafu. To znamená, že jejich metabolomické profily jsou si navzájem velmi blízké.

Všechny prezentované výstupy, včetně grafu skórů a Scree plotu, jsou zpracovány v interaktivní podobě a spolu s 3D grafy či grafy zátěží jsou k nahlédnutí v příloze *PCA.html*. Skript pro generování těchto interaktviních vizualizací byl vytvořen v Pythonu za využití umělé inteligence [8].

4.2 Shlukování

Shlukování se vztahuje k velmi široké škále technik pro nalezení podskupin, neboli shluků, v datovém souboru. Když shlukujeme pozorování, snažíme

se je rozdělit do odlišných skupin tak, aby pozorování v rámci každé skupiny byla velmi podobná, zatímco pozorování v různých skupinách byla značně odlišná. Shlukování se snaží najít homogenní podskupiny mezi pozorováními.

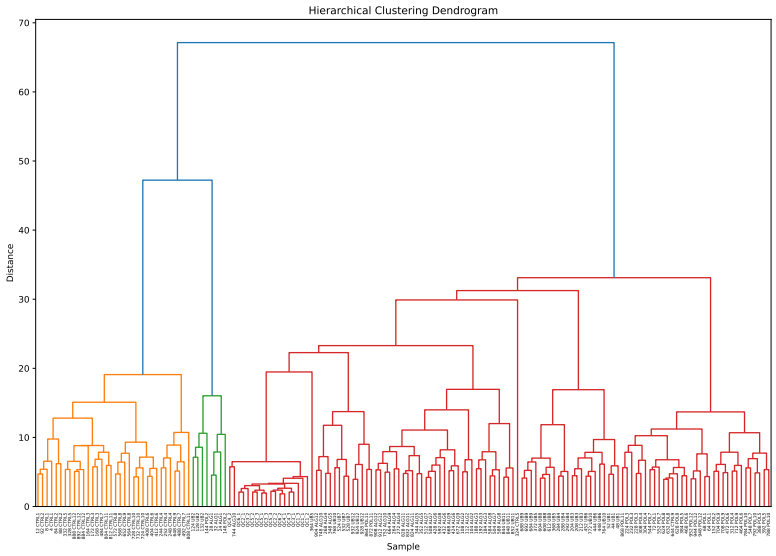
Zabýváme se aglomerativními metodami popsanými v [17] a [28]. Na první úrovni hierarchie každý z n objektů tvoří samostatný shluk, což vede k n shlukům. Na další úrovni jsou spojeny dva nejbližší shluky, a tak dále, až nakonec všechny objekty vytvoří jeden jediný shluk. Výsledné grafické reprezentaci tohoto postupu říkáme dendrogram.

Vzdálenost mezi dvěma shluky s indexy j a l lze potom určit různými metodami, přičemž nejčastěji se používají následující

- complete linkage: $\max_i \|\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\|$,
- single linkage: $\min_i \|\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\|$,
- average linkage: $\text{average}_i \|\mathbf{x}_i^{(j)} - \mathbf{x}_i^{(l)}\|$,
- Ward linkage: $\|\mathbf{c}_j - \mathbf{c}_l\| \cdot \frac{\sqrt{2n_j n_l}}{\sqrt{n_j + n_l}}$.

V těchto definicích jsme použili eukleidovskou vzdálenost, ale lze zvážit i jiné míry vzdálenosti. Complete linkage uvažuje vzdálenost mezi dvěma shluky jako vzdálenost dvou nejvzdálenějších pozorování j a l, naopak single linkage bere vzdálenost dvou nejbližších pozorování. Average linkage počítá průměrnou vzdálenost mezi všemi dvojicemi pozorování ve dvou shlucích, tj. počítá součet všech párových vzdáleností mezi shluky a dělí počtem párů $n_j n_l$. Wardova metoda používá vzdálenost mezi centroidy shluků $\boldsymbol{c}_j$ a $\boldsymbol{c}_l$, které jsou přeškálovány počtem pozorování n_j, n_l v jednotlivých shlucích. To zajišťuje rostoucí vzdáleností mezi centroidy v rámci postupného shlukování.

Na obrázku 14 máme výsledek Ward linkage metody použitím funkce `dendrogram` a `linkage` z modulu *scipy*. Opět jde hezky vidět rozdělení kontrolní skupiny *CTRL* (oranžová barva) oproti ostatním.



Obrázek 14: Dendrogram - Ward linkage

5 Výsledky analýzy

V rámci analýzy včelí hemolymfy jsme došli k několika závěrům. Nejvýznamnější rozdíl byl pozorován v rámci hladin aminokyselin. U kontrol krmených pouze roztokem 50 % sacharózy jsou pak pozorovatelné zvýšené hladiny sacharidů a snížené hladiny aminokyselin. Dále se oddělují karnitiny s dlouhým řetězcem (např.: *Car18, Car18.1, Car22, Car20*) v případě skupin *ALG*, *UB* a *POL*, které mají zvýšené hladiny těchto karnitinů, od kontrolní skupiny *CTRL*, viz obrázek 6b. V rámci vícerozměrné statistiky lze jasně pozorovat separaci jednotlivých skupin podle jejich metabolomického profilu, přičemž skupiny *UB* a *POL* vykazují mezi všemi skupinami největší vzájemnou podobnost. Detailní výsledky lze nalézt v přílohách práce.

Z hlediska včelařské praxe výsledky potvrzují, že jednostranná sacharidová výživa není pro včely nutričně dostačující, zejména z pohledu přísunu aminokyselin nezbytných pro růst, imunitu a celkovou vitalitu. Alternativní zdroje bílkovin, jako jsou pylové náhražky (např. Ultrabee nebo řasy), mohou efektivně doplnit chybějící živiny v obdobích nedostatku přirozeného pylu a přispět tak ke zlepšení kondice včelstev.

V době nedostatku nektaru, typicky na podzim, se včelám podává sacharidová výživa ve formě cukerného roztoku. Tuto stravu je tedy velmi vhodné doplňovat o proteinové příkrmy. Je však důležité zmínit, že žádná náhražka nedokáže plně nahradit přirozený pyl. Ten totiž včelám kromě proteinů poskytuje i celou řadu dalších nezbytných živin, jako jsou lipidy, sacharidy, škrob, steroly a minerály.

Závěr

Hlavním cílem této diplomové práce bylo vytvořit komplexní workflow v jazyce Python, které by pracovníkům Oddělení klinické biochemie FN Olomouc usnadnilo zpracování omických dat. Tento cíl se podařilo úspěšně splnit. Vytvořený nástroj pokrývá kompletní proces analýzy, jako je předzpracování dat, které zahrnuje LOWESS korekci, filtraci metabolitů, imputaci chybějících hodnot, transformaci, PQN normalizaci a škálování.

Následně workflow nabízí rozsáhlý aparát pro jednorozměrnou statistiku – od testování normality a detekce odlehlých hodnot, přes komplexní parametrické i neparametrické testování hypotéz (včetně metody ANCOVA), až po korelační a ROC analýzu či tvorbu volcano plotů. Pro celkové zhodnocení datových sad jsou v programu implementovány také populární metody vícerozměrné statistiky, konkrétně analýza hlavních komponent a hierarchická shluková analýza.

Aby byl nástroj pro oddělení skutečně přínosný, bylo klíčové jeho ovládání co nejvíce zjednodušit a zautomatizovat. Uživatelům tak stačí nahrát svá data a zbytek výpočtů i grafických výstupů jednoduše vygenerovat. Tuto funkčnost jsme v práci úspěšně demonstrovali na reálném datasetu včelí hemolymfy. Skript si s daty bez problémů poradil a vytvořil přehledné výstupy ve formátech HTML, PDF a tabulkách MS Excel.

Věřím, že implementace tohoto workflow kolegům výrazně usnadní a urychlí jejich každodenní práci s vysoce dimenzionálními daty. Celý skript je navíc navržen tak, aby jej bylo možné snadno upravovat a rozšiřovat. Do budoucna se nabízí hned několik možností doplnění: implementace metody *SERRF* jako alternativy k regresi LOWESS, zařazení metody *O-PLS-DA* do vícerozměrné analýzy a přidání vícerozměrných metod pro detekci odlehlých hodnot.

Reference

[1] ADIKARAM, K., HUSSEIN, M. A., EFFENBERGER, M., BECKER, T.: *Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation.* Journal of Applied Mathematics [online]. 2015 [cit. 2026-02-6]. Dostupné z: https://doi.org/10.1155/2015/708948

[2] BROADHURST, D, GOODACRE, R: *Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies.* Metabolomics [online]. 2018 [cit. 2026-01-27]. Dostupné z: https://www.researchgate.net/publication/232756497_Procedures_for_large-scale_metabolic_profiling_of_serum_and_plasma_using_gas_chromatography_and_liquid_chromatography_coupled_to_mass_spectrometry

[3] CLEVELAND, W.: *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association [online]. 1979, 74(368), 829-836 [cit. 2025-10-02]. Dostupné z: https://doi.org/10.1080/01621459.1979.10481038

[4] CONOVER, W. J.: *Practical nonparametric statistics.* 3rd ed. Wiley series in probability and statistics. Applied probability and statistics. New York: John Wiley, c1999. ISBN 978-0-471-16068-7.

[5] DIETERLE, F, ROSS, A. SCHLOTTERBECK, G., and SENN, H.: *Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics* Analytical Chemistry 2006. [cit. 2026-1-26].https://doi.org/10.1021/ac051632c

[6] DUNN, W, BROADHURST, D., et al: *Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.* Nature Protocols [online]. 2017 [cit. 2026-01-27]. Dostupné z: https://www.researchgate.net/publication/232756497_Procedures_for_large-scale_metabolic_profiling_of_serum_and_plasma_using_gas_chromatography_and_liquid_chromatography_coupled_to_mass_spectrometry

[7] FAWCETT, T.: *An introduction to ROC analysis.* Pattern Recognition Letters [online]. 2006, 27(8), 861-874 [cit. 2026-04-11]. Dostupné z: https://doi.org/10.1016/j.patrec.2005.10.010

[8] Google. (2026). Gemini 3 Flash [Velký jazykový model]. Dostupné z: https://gemini.google.com/

[9] HASIN, Y., SELDIN, M., LUSIS, A.: *Multi-omics approaches to disease.* Genome Biology [online]. 2017 [cit. 2026-01-27]. Dostupné z: https://www.researchgate.net/publication/316746666_Multi-omics_Approaches_to_Disease

[10] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. :*The Elements of Statistical Learning* [online]. Second Edition. Springer, 2009, [cit. 2024-01-26]. Dostupné z: https://hastie.su.domains/Papers/ESLII.pdf

[11] HEBÁK, P, J HUSTOPECKÝ a I MALÁ. *Vícerozměrné statistické metody (2).* 1. vydání. INFORMATORIUM Praha, 2005. ISBN 80-7333-036-9.

[12] HRON, K., KUNDEROVÁ, P. a VENCÁLEK, O.: *Základy počtu pravděpodobnosti a metod matematické statistiky.* 4. doplněné vydání. Olomouc: Univerzita Palackého v Olomouci, Přírodovědecká fakulta, 2021. ISBN 978-80-244-5990-5.

[13] HUITEMA, B.: *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies.* Second edition. New Jersey: A JOHN WILEY AND SONS, INC., PUBLICATION, 2011. ISBN 978-0-471-74896-0.

[14] HUR, M. a kol.: *A global approach to analysis and interpretation of metabolic data for plant natural product discovery.* Natural product reports vol. 30,4 (2013). [cit. 2026-2-20]. doi:10.1039/c3np20111b

[15] IDKOWIAK, J., DEHAIRS, J., SCHWARZEROVÁ, J. a kol.: *Best practices and tools in R and Python for statistical processing and visualization of lipidomics and metabolomics data.* Nature Communications [online]. 2025, 16, 8714 [cit. 2026-03-23]. Dostupné z: https://doi.org/10.1038/s41467-025-63751-1

[16] IGLEWICZ, B., HOAGLIN, D.: *How to Detect and Handle Outliers.* Vol. 16. The ASQC Basic References in Quality Control: Statistical Techniques, 1993.

[17] GARETH, J.: *An introduction to statistical learning: with applications in R.* [2nd ed.]. New York: Springer, 2021. Springer texts in statistics. ISBN 978-1-4614-7137-0.

[18] NAHM, F. S.: *Receiver operating characteristic curve: overview and practical use for clinicians.* Korean Journal of Anesthesiology [online]. 2022, 75(1), 25-36 [cit. 2026-04-11]. Dostupné z: https://doi.org/10.4097/kja.21209

[19] NIST/SEMATECH: *e-Handbook of Statistical Methods* [online]. 2012 [cit. 2026-02-6]. Dostupné z: https://doi.org/10.18434/M32189

[20] PATTI, G., YANES, O., SIUZDAK, G.: *Innovation: Metabolomics: The apogee of the omics trilogy.* Nature reviews. Molecular cell biology [online]. 2012 [cit. 2026-01-27]. Dostupné z: https://www.researchgate.net/publication/221903351_Innovation_Metabolomics_The_apogee_of_the_omics_trilogy

[21] PEDREGOSA, F. a kol.: *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research [online]. 2011 [cit. 2025-11-23]. Dostupné z: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[22] RUOPP, M. D., PERKINS, N. J., WHITCOMB, B. W., SCHISTERMAN, E. F.: *Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection.* Biometrical Journal [online]. 2008, 50, 419-430 [cit. 2026-03-23]. Dostupné z: https://doi.org/10.1002/bimj.200710415

[23] SEABOLD, S., PERKTOLD, J.: *statsmodels: Econometric and statistical modeling with python.* 9th Python in Science Conference [online]. 2010 [cit. 2025-11-23]. Dostupné z: https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf

[24] STEKHOVEN, D., BÜHLMANN, P.: *MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics [online]. 2012, 28(1), 112-118 [cit. 2026-01-26]. Dostupné z: https://doi.org/10.1093/bioinformatics/btr597

[25] THE PANDAS DEVELOPMENT TEAM: *pandas-dev/pandas: Pandas.* Zenodo [online]. 2020 [cit. 2025-11-23]. Dostupné z: https://doi.org/10.5281/zenodo.3509134

[26] TROYANSKAYA, O. a kol.: *Missing value estimation methods for DNA microarrays.* Bioinformatics [online]. 2001, 17(6), 520-525 [cit. 2026-02-06]. Dostupné z: https://doi.org/10.1093/bioinformatics/17.6.520

[27] VAN DEN BERG, R.A. :*Centering, scaling, and transformations: improving the biological information content of metabolomics data.* BMC Genomics [online]. 2006 [cit. 2025-11-14]. Dostupné z: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-142

[28] VARMUZA, K., FILZMOSER, P.: *Introduction to Multivariate Statistical Analysis in Chemometrics* [online]. Taylor & Francis Group, 2009 [cit. 2023-12-06]. ISBN 978-1-4200-5947-2.