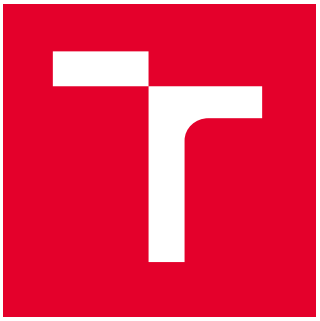


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

GENOMICKÁ PREDIKCE ZALOŽENÁ NA HLUBOKÉM UČENÍ POMOCÍ SÍTÍ LSTM

GENOMIC PREDICTION BASED ON DEEP LEARNING USING LSTM NETWORKS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Daniel Komjaty**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. et Ing. Jana Schwarzerová, MSc**
SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: Daniel Komjaty ***ID:*** 226396

Ročník: 3 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Genomická predikce založená na hlubokém učení pomocí sítí LSTM

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s technikami, které se používají pro genomické predikce. 2) Prostudujte predikční metody založené na lineárních a nelineárních metodách. 3) Předzpracujte si data tak, abyste je mohli vhodně aplikovat do predikční analýzy a otestujte aspoň dva predikční přístupy. 4) Práci rozšiřte o predikční přístup založený na hlubokém učení pomocí LSTM sítí. 5) Dále se zaměřte na konkrétní parametry modelu a síť dostatečně natrénujte. 6) Proveďte diskusi k výsledkům.

DOPORUČENÁ LITERATURA:

[1] SIDAK, D., SCHWARZEROVA, J., WECKWERTH, W., and WALDHERR, S. (2022). Interpretable machine learning methods for predictions in systems biology from omics data. Frontiers in Molecular Biosciences, 9, 926623.

[2] WEISZMANN, Jakob, et al. Metabolome plasticity in 241 Arabidopsis thaliana accessions reveals evolutionary cold adaptation processes. Plant Physiology, 2023, kiad298.

[3] SCHWARZEROVA, Jana, et al. Linear Predictive Modeling for Immune Metabolites Related to Other Metabolites. In: International Work-Conference on Bioinformatics and Biomedical Engineering. Cham: Springer International Publishing, 2022. p. 16-27.

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 29.5.2024

Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc

Konzultant: Univ.-Prof. Dr. Wolfram Weckwerth

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Tato bakalářská práce se věnuje problematice genomické predikce s využitím predikčních metod založených na strojovém učení. V první části se práce zabývá teoretickou rešerší s užším zaměřením na genomickou predikci a její aplikaci v rámci rostlinných dat. Dále se zabývá predikčními algoritmy a modely založenými na strojovém učení, které se využívají pro genomické predikce. Další část obsahuje podrobnější popis použitých genomických a metabolomických dat poskytnutých od vedoucí práce. Ve čtvrté části je popsána samotná implementace vybraných modelů strojového učení. Poslední pátá část se zabývá zhodnocením modelů strojového učení a diskuzí k výsledkům.

KLÍČOVÁ SLOVA

predikční metody, strojové učení, hřebenová regrese, operátor nejmenšího absolutního zmenšení a výběru, náhodný les, sítě s dlouhou-krátkodobou pamětí, genomická predikce

ABSTRACT

This bachelor's thesis deals with the problem of genomic prediction using machine learning based prediction methods. The first part of the thesis deals with theoretical review with a narrower focus on genomic prediction and its application to plant data. Thesis then discusses prediction algorithms and machine learning based models that are used for genomic prediction. The following section contains a more detailed description of the used genomic and metabolomic data, provided by the thesis supervisor. The fourth section describes the actual implementation of the selected machine learning models. The last fifth section deals with the evaluation of the machine learning models and discussion of the results.

KEYWORDS

prediction methods, machine learning, ridge regression, least absolute shrinkage and selection operator, random forest, long short-term memory networks, genomic prediction

KOMJATY, Daniel. *Genomická predikce založená na hlubokém učení pomocí sítí LSTM*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024, 58 s. Bakalářská práce. Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Daniel Komjaty
VUT ID autora:	226396
Typ práce:	Bakalářská práce
Akademický rok:	2023/24
Téma závěrečné práce:	Genomická predikce založená na hlubokém učení pomocí sítí LSTM

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucí mé bakalářské práce paní Ing. et Ing. Janě Schwarzerové, MSc, za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci. Také děkuji mé rodině, která mě podporovala po celou dobu mého studia.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Genomická predikce je postup používaný pro modelování vztahu genotyp-fenotyp a pomáhá pochopit otázky spojené s interpretací genomu. Stále detailnější schopnost pochopení skrytých pochodů na pozadí tohoto vztahu genotyp-fenotyp vede k snadnějšímu pokroku v mnoha oblastech a to jak v biomedicíně, tak v zemědělství a ekologii – při šlechtění rostlin. Globální změna klimatu, zvýšená poptávka po bioenergii a rostoucí světová populace jsou současnými výzvami v rámci udržitelnosti. Proto je zvyšování světové produkce plodin důležitou současnou výzvou a otevírá velké možnosti pro využití predikčních algoritmů v aplikaci genomické predikce u rostlin.

Tato bakalářská práce se zaměřuje na implementaci predikčních modelů pro genomickou predikci rostlin, využívajících genomická data popsána pomocí jednonukleotidových polymorfismů. K implementaci predikčních modelů jsou použity metody strojového učení, konkrétněji hřebenové regrese, operátoru nejmenšího absolutního zmenšení a výběru, náhodného lesa a především sítí s dlouhou-krátkodobou pamětí (tzv. LSTM sítí).

Teoretická část práce se soustředí na teorii v rámci genomické predikce. První část je zaměřena na genomickou predikci, což je problém vyplývající z biologie, jež se zabývá zkoumáním vztahu mezi genotypem a fenotypem. Tato část zahrnuje také přehled celogenomových asociačních studií, které slouží jako základ pro genomické predikce, a důvody pro použití genomické predikce u rostlin. Dále se práce zaměřuje na rešerši publikovaných studií, které se zabývají genomickou predikcí rostlin - a popisem analyzovaných dat, konkrétně modelového organismu *Arabidopsis thaliany*.

Praktická část práce se dále věnuje predikčním metodám. V základu konvenčním modelům založeným na strojovém učení, které jsou využívány pro genomické predikce. Dále je důraz kladen také na hloubkové učení a specifické typy architektur neuronových sítí. Přičemž hlavní pozornost je dána na LSTM sítě, které jsou hlavním tématem této práce, a jejich aplikace v oblasti predikcí. Na závěr práce popisuje hodnotící metriky predikčních modelů.

Praktická stránka práce se zabývá konkrétní implementací vybraných konvenčních modelů strojového učení a sítí LSTM. Obsahuje předzpracování dat a použití konkrétních funkcí pro realizaci jednotlivých modelů. V poslední části této bakalářské práce jsou prezentovány a diskutovány výsledky GP a jim odpovídající výsledky hodnotících metrik pro jednotlivé predikční přístupy.

1 Genomická predikce

Genomická Predikce (GP) se využívá pro modelování vztahu genotyp-fenotyp [1]. Správná interpretace modelů popisujících vztah genotyp-fenotyp povede k nalezení nových vlastností, které se skrývají na pozadí. Informace ukryté právě v těchto vlastnostech tak povedou k efektivnějšímu rozvoji genetiky, molekulární biologie a otevřou nové možnosti v medicínské prognóze. Celo-genomová asociační studie (*angl. Genome-Wide Association Study* – GWAS) je jednou z metod, které se dají využít při interpretaci genomu. GWAS se používá pro identifikaci individuálních alel a interpretaci variant genů ve studiích celo-genomového sekvenování (*angl. Whole Genome Sequencing* – WGS).

Studie WGS mají potenciál identifikovat každou formu genetické variace vzhledem k referenčnímu genomu. Proto je mnohem pravděpodobnější, že WGS data obsahují soubor variant, které ovlivňují studované fenotypy. V minulosti byla problémem dostupnost dat WGS, protože cena sekvenačních technologií byla vysoká. V dnešní době se tento problém přenesl z dostupnosti dat právě na jejich interpretaci. V molekulární biologii je velkou výzvou v této interpretaci hledání tzv. kauzálních variant pro daný fenotyp. Kauzální variantou lze rozumět variantu v genomu, která je odpovědná za fenotypové změny. [2, 3]

1.1 Vztah genotyp-fenotyp

Genotyp označuje soubor všech genů daného jedince, přesněji soubor všech alel tzn. konkrétních variant genů, které jedinec zdědil [4]. Variabilita genetické informace je způsobena její změnou, mezi které patří mutace, delece, indel (inzerce a delece). Tato změna je vztažena k referenčnímu genomu. Soubor genů jedince kóduje polypeptidy jejichž funkce jsou základem pro fenotypové vlastnosti daného jedince. Fenotyp je potom souborem všech pozorovaných vlastností jedince, např. barva květů. [4]

Vztah genotyp-fenotyp je reprezentován mírou variability daného genu k variabilitě pozorovaného fenotypu. Ve skutečnosti individuální gen sám o sobě nezpůsobí pozorovaný fenotypový znak, ani nemusí být potřebný, nebo postačující pro vznik pozorovaných vlastností [4]. Pro sledovaný účinek na organismus je potřeba kombinované působení více alel, tzn. kombinace dominantních a recesivních alel, a jejich kumulativní účinky. Geny mimo jiné interagují s vlivy prostředí, např. abiotickými faktory nebo symbionty, které v důsledku vytvářejí fenotyp. Proto vlivy prostředí také výrazně ovlivňují vztah genotyp-fenotyp. Díky tomu má vztah genotyp-fenotyp komplexní charakter, jež se projevuje nelineárními vztahy. Avšak, komplexní charakter komplikuje genomické predikce, které jsou často modelovány pomocí lineárních predikčních metod. [2, 4]

Jednonukleotidové polymorfismy

Jednonukleotidové polymorfismy (*angl. Single Nucleotide Polymorphism* – SNP) jsou pozice určitých páru bází v DNA, na kterých se u normálních jedinců v dané populaci vyskytují různé varianty. Podmínkou dané varianty je, aby byla její četnost v pozorované populaci 1 % nebo více. Alternativy s nižší četností patří mezi vzácné varianty. SNP v rámci DNA zahrnují různé typy, tj. tranzice C⇔T/G⇔A a transverze G⇔T/G⇔C/A⇔T/A⇔C. Daná tranzice/transverze má synonymní, nebo nesynonymní účinek na translaci aminokyseliny. Souvislost SNP s genotypem je znázorněn na obrázku 1.1. Obrázek 1.1 ilustruje, jak malé změny v genomu formují genotyp a existenci vztahu mezi genotypem a daným fenotypem. Data charakterizovaná pomocí SNP jsou užitečná v řadě bioinformatických a výpočetních aplikacích. [5]

Obr. 1.1: Schéma vztahu genotyp-fenotyp. Převzato a upraveno z [6].

1.2 Celogenomová asociační studie

GWAS zahrnuje testování genetických variant napříč genomy mnoha jedinců s cílem identifikovat vztahy/spojení mezi genotypem a fenotypem. GWAS zkoumá genetické odchylky napříč celým genomem, což z ní činí komplexní nástroj pro identifikaci vztahů mezi konkrétními genetickými markery a znaky organismu. Současnou analýzou tisíců až milionů SNP ve velkých a různorodých populacích může GWAS určit běžné genetické varianty, které přispívají k určitému projevu. [7, 8]

Prvními kroky GWAS analýzy jsou identifikace pozorovaného znaku a k němu vhodný výběr studované populace, neboli neurčitý vzorek populace pro určitý fenotypový znak. Genotypizaci lze provést pomocí WGS, nebo dat charakterizovaných

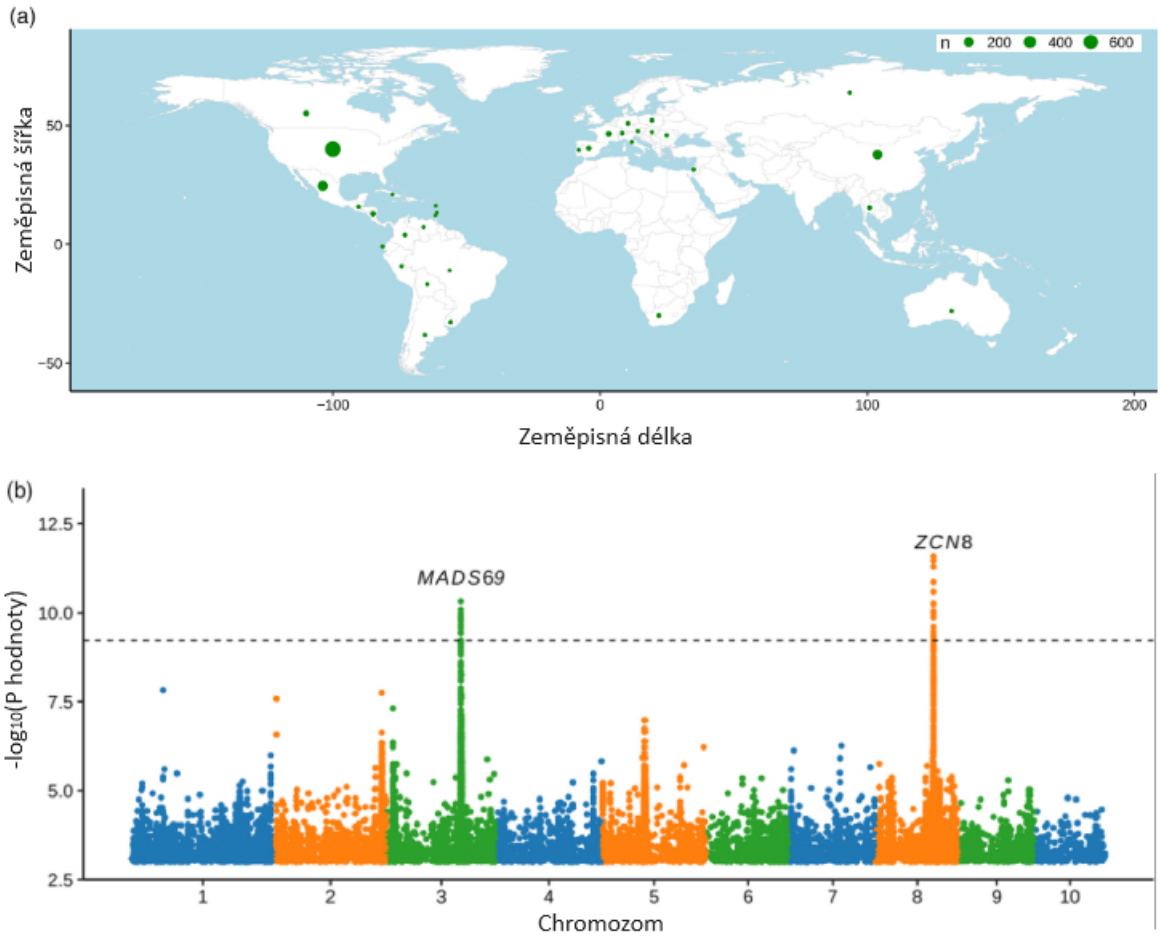
SNP maticemi, které jsou v rámci GWAS stále nejvíce využívané. Především díky nižším nákladům, protože provádění WGS u velmi velkých vzorků je v současné době nákladově náročné. K identifikaci oblastí genomu spojených se zájmovým fenotypem s celo-genomovou významností se používají asociační testy. [7, 8].

Zjištěná spojení pomocí GWAS vedou k identifikaci významných alel v kontextu sledovaných fenotypů. Například ve studii Grzybowski a kol. [9] se zaměřili na identifikaci alel souvisejících s kvetením samic kukuřice. Obrázek 1.2a znázorňuje zeměpisné rozložení jedinců kukuřice použitých v této studii [9]. Na obrázku 1.2b je Manhattanský graf, který se využívá pro vizualizaci výsledků GWAS analýzy. Na grafu je vidět, že asociační testy identifikovaly dvě významné alely MADS69 a ZCN8 související s kvetením samic kukuřice [9]. GWAS analýzy stále vedou k pokroku v klinické péči, např. identifikace nových cílů pro léčiva a biomarkerů onemocnění, a v personalizované medicíně, např. predikce rizika a optimalizace terapie na základě genotypu. Avšak, i přes všechny výhody má GWAS svá omezení, jako je například nedostatečná schopnost určit příčinné varianty a geny, omezená klinická výpovědní hodnota a neschopnost identifikovat všechny genetické determinanty komplexních znaků. [8]

1.3 Genomická predikce u rostlin

GP je přístup založený na predikci pomocí celého genomu, jejíž potenciál se zdá vhodný pro aplikaci při zlepšování plodin, díky velké dostupnosti genomických dat. Například ve studii Raimondi a kol. [2] se povedlo identifikovat 36 nových genů, které jsou pravděpodobně spojeny s rysy kvetení. Přičemž pro 6 z těchto 36 nově identifikovaných genů existují důkazy pomocí laboratorních experimentů publikované v literatuře [2]. Další výsledky studií naznačují, že GP je výkonný doplňkový přístup při šlechtění hybridů pro vysoce polygenní znaky [3]. Cílem většiny šlechtitelských programů je křížení inbredních linií s vhodnými partnery, u kterých se snaží předpovědět genetické hodnoty a umožnit tak cílené kombinace žádoucích alel, aby se vytvořili hybridi s vyšší vitalitou a zvýšeným výnosem [3, 10].

Výzvy v rámci udržitelnosti v současné době čelí globální změně klimatu, zvýšené poptávce po bioenergie a rostoucí světové populaci. Mimo jiné je potřeba udržovat přirozenou biodiverzitu naší planety. Vzhledem k těmto výzvám je důležité zaměřit se na adaptaci rostlin, pro zvýšení jejich celkové odolnosti a výnosu. Což otevírá velké možnosti predikčním algoritmům v aplikaci na genomickou predikci rostlin. [2, 3, 10]



Obr. 1.2: Obrázek znázorňuje zeměpisné rozložení použitých vzorků a Manhattanský graf s identifikovanými alelami. Převzato a upraveno z [9].

2 Predikční techniky

Predikční modely jsou založené na metodách strojového učení (*angl. Machine Learning* — ML), které spadá do problematiky umělé inteligence (*angl. Artificial Intelligence* — AI). ML metody se zaměřují na vytvoření modelu, který na základě vstupních informací, které mohou reprezentovat genetickou informaci, predikuje výstupní hodnotu - v našem případě fenotyp. Jinými slovy, model se snaží naučit vztahy, vzory nebo struktury identifikované v datech. [11]

Metody ML se v základu dělí podle přístupu k učení modelu, které se dá např. rozdělit na zpětnovazebné učení (*angl. reinforcement learning*), učení bez učitele (*angl. unsupervised learning*) a učení s učitelem (*angl. supervised learning*). V této práci se budeme zaměřovat na učení s učitelem, které je obecně rozšířeno v kontextu prediktivní systémové biologie [11]. Učení s učitelem spočívá v tom, že se pro učení modelu využívají trénovací data, která obsahují soubor vstupů a k nim odpovídající cílovou proměnou, kterou chceme predikovat. Existují dvě hlavní úlohy při učení s učitelem, které se liší podle toho, zda je cílová proměnná kategoriální nebo spojitá. U kategoriální proměnné se jedná o úlohu klasifikace. U spojité proměnné se jedná o úlohu regrese. Vstupem může být jakákoli hodnota, u které očekáváme, že se bude podílet na predikci cílové proměnné. [11, 12]

2.1 Konvenční metody

Klasická Vícenásobná Lineární Regrese (VLR) je základem pro velkou skupinu konvenčních predikčních metod, které se často využívají pro GP. VLR modeluje lineární vztahy mezi vstupy a cílovou proměnnou. Genomická data obvykle obsahují mnoho vstupů a obsažené informace mohou být vysoce korelované, což způsobuje při aplikaci VLR problémy. Metody regularizovaných lineárních regresí nebo sofistikovanější metody založené na stromových strukturách lépe řeší problémy spojené s aplikací VLR na genomická data. [13]

2.1.1 Regularizované regresní metody

Z metod regularizovaných lineárních regresí jsou zde detailněji zmíněny hřebenová regrese (*angl. Ridge Regression* — RR) a operátor nejmenšího absolutního zmenšení a výběru (*angl. Least Absolute Shrinkage and Selection Operator* — LASSO), které se řadí mezi často používané pro GP. Tyto metody se zabývají dvěma problémy, a to odhadem regresních parametrů a výběrem vstupních hodnot, tzn. vybírají relevantní vstupy vzhledem k požadované predikci. Odhad regresních parametrů se provádí pomocí metody nejmenších čtverců (*angl. Least Squares* — LS), která se

běžně používá u lineární regrese. Výběr vstupů užitečných pro predikci je v těchto metodách zajištěn pomocí zavedené regularizace, která je řešena pomocí tzv. penalizace [14]. Díky regularizaci, metody lépe řeší problém multikolinearity, který vyplývá z komplexnosti použitých dat a problém nadměrně velkého rozptylu, který má negativní vliv na přesnost predikce u testovacích dat, tzn. model se nadměrně přizpůsobí trénovacím datům. [13, 14, 15]

Základem obou metod je lineární regresní model, který je popsán následující rovnicí:

$$Y = X\beta + \epsilon, \tag{2.1}$$

kde Y je n-rozměrný vektor proměnných cílové predikce, X je matice příznaků o velikosti n x p (n je počet pozorování a p je počet vstupů), β je vektor regresních koeficientů vstupů a ϵ je vektor náhodných chyb. [13]

RR využívá k vytvoření modelu reziduálního součtu čtverců a tzv. L2 penalizace [14]. Míru působení L2 penalizace určuje ladící parametr $\lambda \geq 0$, tzn. λ řídí míru zmenšení koeficientů. Vyšší hodnota parametru λ způsobí větší zmenšení koeficientů. Koeficienty se u RR modelu zmenšují směrem k nule, ale nikdy nemohou být vynulovány a tím způsobit vyřazení některých vstupních hodnot. λ je hyperparametrem RR modelu, tzn. musí být vhodně zvolen. K určení λ se běžně používá křížová validace aplikovaná na data. [13, 14, 15]

LASSO využívá k vytvoření modelu reziduálního součtu čtverců a tzv. L1 penalizace [14]. Míru působení L1 penalizace určuje stejně jako u RR ladící parametr $\lambda \geq 0$. Rozdíl oproti RR je v penalizaci L1, která umožňuje metodě LASSO zmenšení některých regresních koeficientů na nulu. LASSO tedy dokáže vyřadit některé vstupní hodnoty. Proto je metoda LASSO vhodnou volbou při aplikaci pro výběr vstupních hodnot. [13, 14, 15]

2.1.2 Náhodné lesy

Metody náhodného lesa (*angl. Random Forest* — RF) jsou schopny zpracovat velké množství vstupů a určit jejich význam v kontextu cílové proměnné. Mají potenciál identifikovat a modelovat komplexní nelineární vztahy mezi vstupy a cílovou proměnnou. Proto bývají RF využívány pro GP. Základem metody RF jsou rozhodovací stromy. [16]

Rozhodovací strom (*angl. Decision Tree* — DT) je stromově strukturovaný graf, který je rozdělený na uzly, větve a listy, nebo-li koncové uzly stromu. V rámci DT lze rozlišit dvě různé metodiky: klasifikační strom a regresní strom (*angl. Regression Tree* — RT). Dále se zaměříme podrobněji na RT, protože je v této práci řešen problém regrese. RT používá kritérium součtu čtverců k rozdělení dat na postupně homogennější podmnožiny obsažené v uzlech. Ke každému z koncových uzlů

je připojena jednoduchá regrese, která platí pouze v tomto uzlu. V rámci jednoho regresního stromu lze tedy na různé podmnožiny dat aplikovat různé regrese, které mohou představovat různé odezvy řízené různými vstupy. Tyto vlastnosti dodávají RT robustnost, což umožňuje metodě nalézt skryté nelineární vztahy v datech. Z tohoto důvodu se tato práce zaměřuje také na aplikaci metody RF, založené na RT. RF je v kontrastu k regresním metodám, které dokáží nalézt jen lineární vztahy v datech. [16, 17, 18]

RF je technika, která kombinuje výkonnost mnoha algoritmů DT pro klasifikaci nebo regresi. V případě RT dokáže tedy provádět vícerozměrnou nelineární regresi a kombinovat výkonnost mnoha algoritmů regresních stromů. RF při regresní analýze ze vstupního vektoru, který se skládá z různých vstupních hodnot analyzovaných pro cílovou proměnnou, sestaví řadu regresních stromů, které následně zprůměruje a vytvoří tak výsledný strom. RF nechává růst jednotlivé stromy z různých podmnožin trénovacích dat, vytvořených postupem zvaným „bagging“, aby zabránil korelaci mezi stromy. Jakmile je proces vytváření jednotlivých stromů ukončen, lze použít metody prořezávání, jejichž cílem je zlepšit zobecňovací schopnost stromů snížením jejich strukturální složitosti. Za kritérium prořezávání lze považovat např. počet případů v uzlech. Pro metodu RF se dále využívají tyto tréninkové parametry: počet stromů v lese; počet vstupů vyzkoušených při každém dělení; a minimální velikost uzlu, pod kterou nejsou listy dále děleny. [16, 17, 18, 19]

2.2 Metody hloubkového učení

Hlavním cílem této práce, zaměřující se na genomické predikce, je otestování metody založené na hloubkovém učení (*angl. Deep Learning* – DL), které se řadí mezi přístupy strojového učení. Rozdíl oproti klasickým statistickým metodám je, že metody DL vytvářejí neparametrické modely. Neparametrické modely jsou vhodné pro data s komplikovanými nelineárními vztahy mezi vstupy a cílovou proměnnou, tzn. jsou schopné se přizpůsobit složitým a skrytým vzorům neznámé struktury v datech. Což se jeví jako vhodný přístup pro aplikaci GP, vzhledem ke složitým vztahům na pozadí vztahu genotyp-fenotyp. Metody DL potřebují ke správnému naučení modelů dostatečně velká trénovací data, které jsou mnohonásobně větší než u konvenčních predikčních modelů. Genomická data jsou obsáhlá, proto se hodí pro aplikaci DL modelů. Na základě současné literatury metody DL nemají jasnou převahu nad konvenčními predikčními modely z hlediska predikční síly [20]. Nicméně existují jasné důkazy, že modely DL zachycují nelineární vzory efektivněji než konvenční modely, proto je vhodné zkoumat jejich výkonnost v rámci GP [20]. [20]

Metody DL využívají umělé neuronové sítě, které byly inspirovány skutečnými biologickými neurony a jejich propojeními [11]. Neuron je v rámci DL matematická

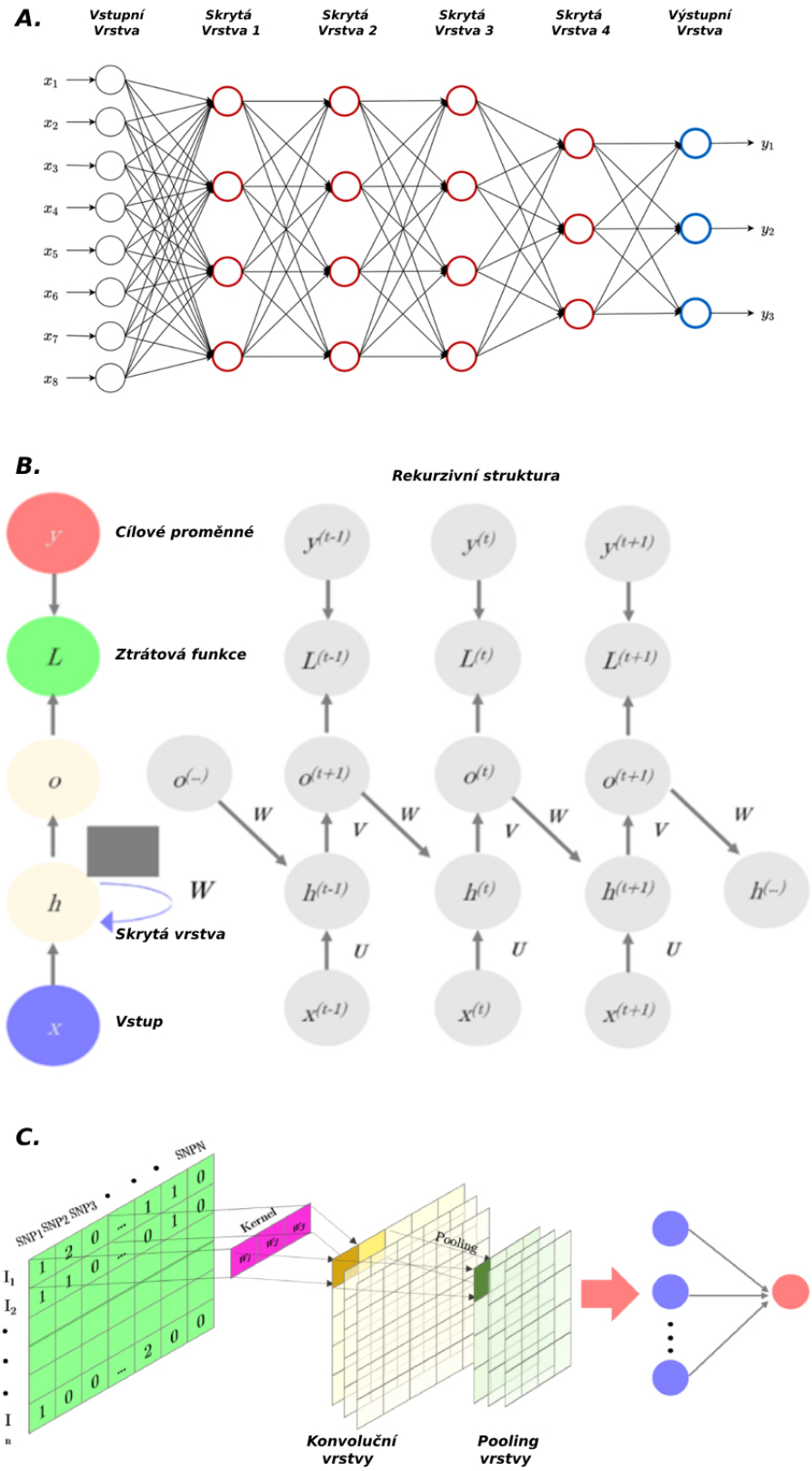
funkce, která přijímá vstupní vektor hodnot a vhodně jej transformuje. Transformace je prováděna pomocí aktivační funkce, která určuje výstupní hodnotu na základě adaptivních vah a prahu. Výstupní hodnota představuje aktuální aktivaci neuronu. Práh je učenlivý parametr, který se přidává do aktivační funkce a určuje, jak snadno se neuron aktivuje. Hodnoty vah kódují naučené informace získané z trénovacích dat a určují sílu spojení mezi jednotlivými neurony. [10, 11, 20]

Mezi nejpoužívanější topologie neuronových sítí patří dopředná neuronová síť (*angl. Feedforward Neural Network* – FNN), rekurentní neuronová síť (*angl. Recurrent Neural Network* – RNN) a konvoluční neuronová síť (*angl. Convolutional Neural Network* — CNN), jenž jsou znázorněné na obrázku 2.1. FNN 2.1A a RNN 2.1B obsahují neurony ve skrytých vrstvách navzájem hustě propojené. RNN navíc obsahuje rekurzivní spojení, tzn. informace se v síti šíří oběma směry. V levé části 2.1B je zobrazena obecně celá struktura RNN. X představuje vstupy, h jsou skryté vrstvy, o jsou výstupy, y jsou pozorované cílové proměnné a L je ztrátová funkce modelu RNN. Ztrátová funkce měří, jaké jsou kvantitativní rozdíly mezi pozorovanými a předpovídanými proměnnými. V pravé části 2.1B je znázorněna rekurzivní struktura RNN. CNN 2.1C obsahuje navíc oproti FNN a RNN konvoluční vrstvy v kombinaci s tzv. pooling vrstvami. Pooling vrstvy provádí podvzorkování, tzn. slučují výstupy z různých po sobě jdoucích pozic. Pooling vrstvy nezavádí do sítě CNN žádné nové parametry, provádí jen redukci rozměrů a odstraňují šum. [20, 21]

Neuronové sítě se učí v epochách, tzn. v jednotlivých průchodech neuronovou sítí dopředu a dozadu. Při průchodu dozadu jde o zpětné šíření, během nějž je chyba predikce modelu sledována zpět k jednotlivým odhadovaným parametrům modelu a je možné je vhodně upravit, tzn. umožňuje efektivní učení sítě. S rostoucím počtem epoch se v neuronové síti mění váhy a model přechází z podtrénování (*angl. underfitting*) do optimálního natrénování (*angl. fitting*), nebo do fáze přetrénování (*angl. overfitting*). [10, 11]

2.2.1 Konvoluční neuronové sítě

CNN jsou výkonné nástroje pro práci s daty, která obsahují vstupní hodnoty se známými prostorovými vztahy. CNN efektivně zachycují prostorové a časové závislosti vstupních dat [20]. Aplikují se na jednorozměrná i vícerozměrná data. Do aplikace na jednorozměrná data se řadí např. data reprezentovaná pomocí SNP, jak je znázorněno na obrázku 2.1C. Modely CNN zmenšují velikost vstupu a sdílejí parametry. Což vede ke snížení počtu parametrů, které je třeba odhadnout, a zvýšení výkonnosti. [11, 20, 21]

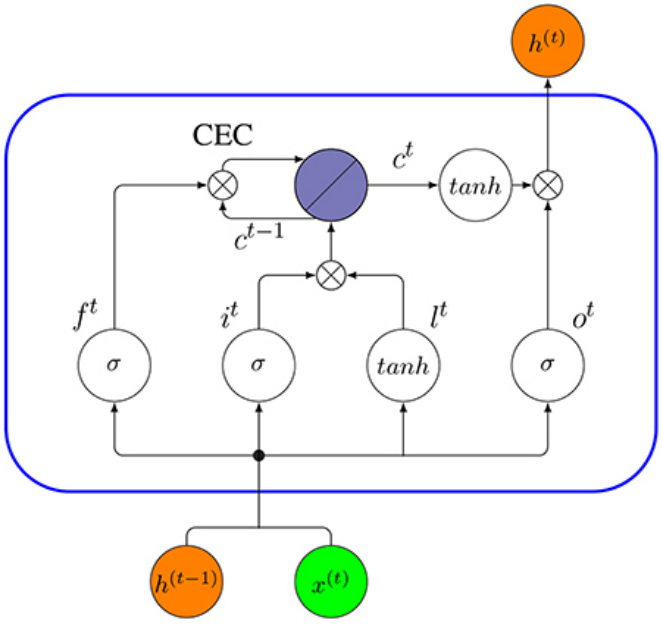


Obr. 2.1: A: Dopředná pětivrstvá neuronová síť, převzato a upraveno z [19]. B: Schéma fungování rekurentní neuronové sítě, převzato a upraveno z [21]. C: Jednotlivé kroky konvoluční neuronové sítě, převzato a upraveno z [21].

2.2.2 Sítě s dlouhou-krátkodobou pamětí

Podrobněji než v předešlé podkapitole se v této práci zaměříme na sítě s dlouhou-krátkodobou pamětí (*angl. Long Short-Term Memory* — LSTM), jejichž implementace v rámci GP je hlavním cílem této práce. LSTM sítě jsou speciálním druhem RNN a jsou určeny k naučení dlouhodobých závislostí v datech [21]. RNN jsou obecně navrženy pro modelování časoprostorových struktur v datech, čehož se dá využít u sekvenčních dat při GP [21].

LSTM sítě lépe řeší problémy při zpracovávání velmi dlouhodobých závislostí v datech, s kterými mají klasické RNN problémy [22]. LSTM sítě jsou známé svými příznivými konvergenčními vlastnostmi, tzn. odchylky dopočítávané ztrátovou funkcí mají tendenci konvergovat k nule [20]. V síti LSTM se základní jednotka nazývá paměťový blok. Každý blok se skládá z buňky, paměťové části a tří bran: vstupní brány, výstupní brány a zapomínající brány. Vstupní brána je jednotka, která řídí tok informací do buňky. Zapomínající brána určuje, které informace z předchozích kroků buňky mají být zapamatovány a které zapomenuty. Stav buňky nám udává, které informace jsou v buňce uchované z minulých běhů. Výstupní brána je jednotka, která řídí tok informací ven z buňky. Paměťový blok si dokáže zapamatovat hodnoty v libovolných časových intervalech v rámci dat/sekvence. Což by se dalo využít při GP, protože jednotlivé geny spolu v rámci genomu/sekvence navzájem interagují. Obecná struktura sítě LSTM je podobná jako v RNN. Rozdíl je v neuronech ve skrytých vrstvách, jenž jsou nahrazeny paměťovými bloky. [22]



Obr. 2.2: Vnitřní struktura paměťového bloku sítě LSTM. Převzato z [22].

Na obrázku 2.2 je znázorněn jeden paměťový blok sítě LSTM. Uvnitř bloku je naznačen způsob propojení jednotlivých částí, které jsou popsány výše [22]. $h^{(t)}$ představuje výstup aktuálního časového kroku. $h^{(t-1)}$ představuje výstup předchozího časového kroku. $x^{(t)}$ je vstup aktuálního časového kroku. Výstup $h^{(t-1)}$ a vstup $x^{(t)}$ jsou vstupem do bloku v aktuálním časovém kroku t. Výstup $h^{(t)}$ v časovém kroku t pak bude vstupem do stejného bloku v dalším časovém kroku $(t+1)$. [22]

2.3 Hodnotící metriky predikčních modelů

Mezi nejpoužívanější metriky pro hodnocení predikčních modelů patří střední kvadratická chyba (*angl. Mean Squared Error* – MSE) a Korelační Koeficient (KK). MSE poskytuje průměrný čtvercový rozdíl mezi cílovou a predikovanou hodnotou. MSE je definována následovně:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 [23], \tag{2.2}$$

kde N je počet vzorků, y_i jsou cílové hodnoty predikce a $\hat{y}_i$ jsou predikované hodnoty modelu. [23]

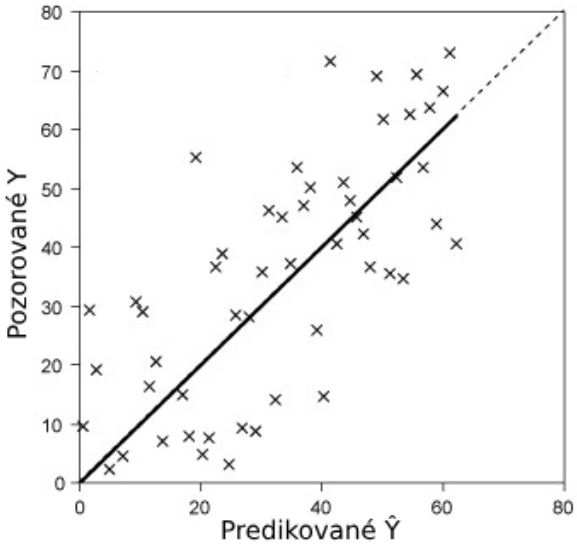
KK je statistická veličina, která nám udává míru korelace mezi dvěma proměnnými. Korelace je metoda, která nám udává míru možné lineární závislost mezi dvěma spojitými proměnnými. Mezi hlavní typy KK se řadí i Pearsonův KK, na který se v rámci této práce více zaměříme. Extrémní hodnoty mohou sílu Pearsonova KK nadhodnotit nebo podhodnotit, a proto je nevhodný, pokud jedna nebo obě proměnné nejsou normálně rozděleny. V této práci byl použit Pearsonův KK dostupný z [24]. Pearsonův KK je definován následovně:

$$r = \frac{\sum\left(x - m_x\right)\left(y - m_y\right)}{\sqrt{\sum\left(x - m_x\right)^2 \sum\left(y - m_y\right)^2}} [24], \tag{2.3}$$

kde r je Pearsonův KK pro vektory proměnných x a y, m_x je průměrná hodnota vektoru x a m_y průměrná hodnota vektoru y. [25]

Pearsonův KK může nabývat hodnot v rozmezí $\langle -1, 1\rangle$. Nulová hodnota KK znamená, že mezi proměnnými neexistuje žádný lineární vztah. Hodnota $KK = 1$ znamená dokonalý pozitivní lineární vztah, proměnné spolu přímo souvisí, tzn. s rostoucí hodnotou jedné proměnné má tendenci růst i hodnota druhé proměnné. Hodnota $KK = -1$ znamená dokonalý negativní lineární vztah, proměnné spolu souvisejí nepřímo, tzn. s růstem hodnoty jedné proměnné má hodnota druhé tendenci klesat. Pearsonův KK dokáže poukázat jen na lineární vztah mezi proměnnými. Jakýkoliv nelineární vztah bude z hlediska Pearsonova KK nulový. V této práci je zkoumán lineární vztah mezi predikovanými a reálnými hodnotami, proto je pro

hodnocení modelů použit Pearsonův KK. Znázornění hledaného lineárního vztahu mezi reálnou a predikovanou cílovou hodnotou je vidět na obrázku 2.3. [25]



Obr. 2.3: Graf znázorňující hledanou lineární závislost mezi reálnými a predikovanými cílovými hodnotami. Převzato a upraveno z [26].

3 *Arabidopsis thaliana*

Huseníček rolní (*lat. Arabidopsis thaliana* – AT) je drobná kvetoucí rostlina, řadící se do čeledi Brukvovitých [2]. AT se stala klíčovým objektem studia a modelovým organismem v molekulární biologii a genetice rostlin. Tento status modelového organismu získala díky několika faktorům, které jsou popsané v následujícím odstavci. Výzkum AT přinesl hlubší pochopení základních fyziologických, buněčných a molekulárních procesů [27]. Dále významně rozvinul poznatky o vnitrodruhové genetické variabilitě [27].

AT disponuje genomem o velikosti ∼125 Mb a krátkým životním cyklem [2]. Je samosprašným druhem s rozšířením do různých zeměpisných šířek, což ji vystavuje široké škále klimatických podmínek [28]. Široké rozšíření a převážně samosprašný typ rozmnožování vedly ke vzniku velkého množství geneticky odlišných homozygotních, inbredních linií [27], dále nazývaných linie vzorků. V roce 2000 proběhlo sekvenování genomu AT, které významně přispělo k nárůstu dostupných sekvenačních dat, zahrnující data WGS a detailní fenotypové anotace [2].

3.1 Genomická informace

Použitý dataset v této bakalářské práci reprezentující genomickou informaci obsahuje 241 genomů vybraných linií vzorků zmíněných níže, v části metabolomické informace. Genomická informace v datasetu je popsána pomocí SNP, které jsou dostupné ze studie Weiszmann a kol. [28]. Genomická informace studie Weiszmann a kol. [28] vychází z dostupných genomických informací studie Alonso-Blanco a kol. [27]. Studie [27] se zabývá analýzou genomického sekvenování více než 1 000 přirozených inbredních linií AT. Analýza odhaluje globální populační strukturu, migrační vzorce, evoluční historii a poskytuje bohatý genetický zdroj pro studium fenotypové variability a adaptace AT [27].

Studie [27] prezentuje 1 135 genomů, sekvenovaných na řadě platforem Illumina v průběhu několika let. Jednotlivé varianty byly vyvolány pomocí MPI-SHORE a GMIGATK, blíže popsané ve studii [27]. Výsledkem byly soubory VCF s vysokou kvalitou. Volání variant bylo porovnáno pomocí celo-genomových zarovnání jednoho dlouhého čtení (Pacific Biosciences) a tří krátkých čtení (Illumina) *de novo* sestavených proti referenci [27]. Průměrná hodnota pravdivě pozitivních výsledků byla 98 %, průměrná hodnota falešně negativních výsledků 1,5 % [27]. Po filtrování obsahovaly jaderné genomy 10 707 430 bialelických SNP a 1 424 879 indelů [27]. To představuje v průměru jednu variantu na každých 10 bp jediné kopie genomu, což je nejhustší mapa variant pro jakýkoli organismus [27].

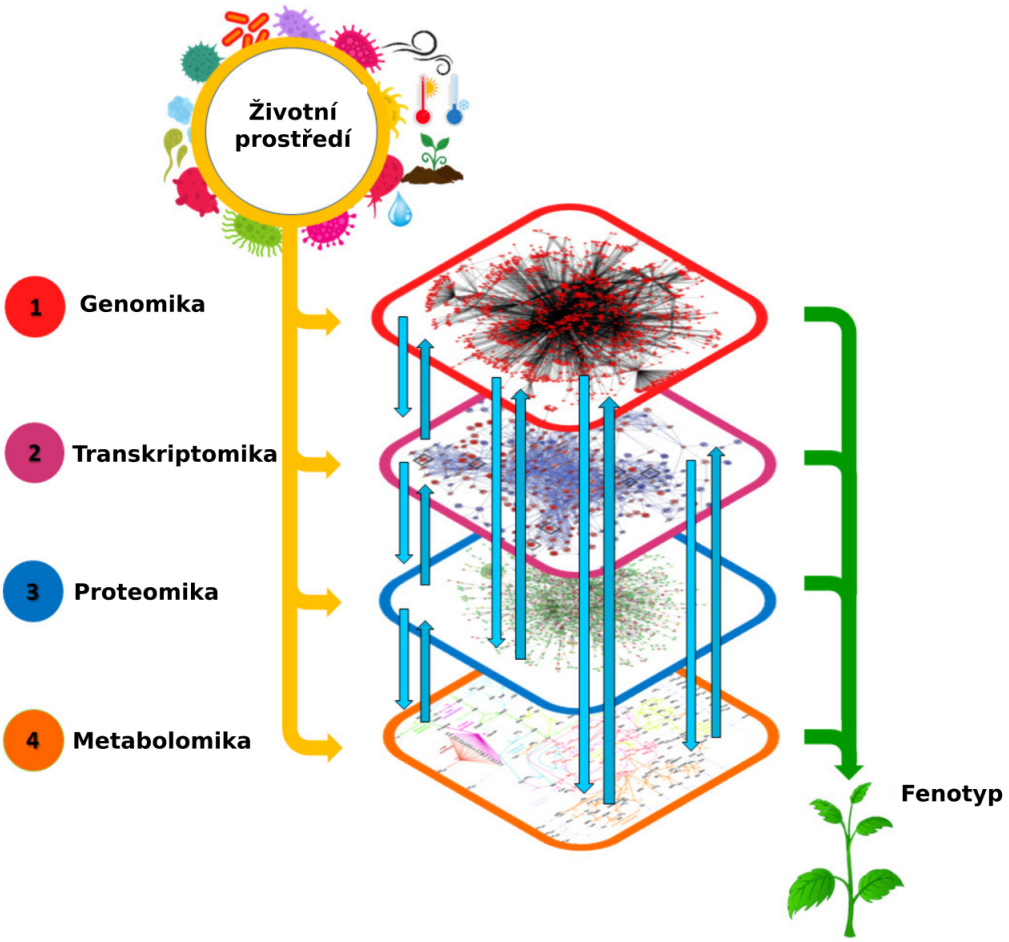
Ve studii [28] byl pro SNP dataset kladen požadavek, aby se minoritní alela vyskytovala alespoň ve 13 liniích vzorků, tzn. frekvence minoritní alely > 5 % ze všech 241 linií vzorků. Výsledkem bylo 1 756 214 bialelických SNP. Tedy soubor s vysokým rozlišením přibližně jeden SNP na 77 bp, ve studii [28] byla data použita pro výpočty GWAS. Pro účely GP byl vytvořen unikátní soubor dat SNP. V souboru nejsou žádné opakující se profily SNP napříč 241 liniemi vzorků. Soubor obsahuje vektory o délce 241 s prvky "1"a "-1". Alely byly kódovány jako "1"a "-1", aby odrážely dva různé diploidní homozygotní genotypy [28]. Z původního souboru dat byly vybrány pouze SNP, které vyžadovaly imputování nejvýše jedné chybějící alely, tzn. soubor s vysokou spolehlivostí [28]. Výsledkem byl soubor 16 544 bialelických SNP, tedy jeden SNP na $\sim$8 kb. Pokrytí SNP bylo považováno za dostatečné pro účely GP, bližší informace ve studii [28].

3.2 Metabolomická informace

Metabolomika nabízí široký pohled na biochemický a fyziologický stav organismu [29]. Změny na úrovni metabolomu jsou odrazem změn na úrovni genomu, transkriptomu a proteomu [29], což je možné vidět na obrázku 3.1. Metabolom je tedy považován za základní biochemickou vrstvu odrážející všechny informace vyjádřené a modulované v ostatních vrstvách omiky, což z něj činí nejbližší spojení s fenotypem [29]. Proto mohly být v této práci použiti hladiny jednotlivých metabolitů jako cílové hodnoty GP.

Metabolomická data použitá v této práci pochází ze studie [28]. V této studii bylo 241 přírodních linií vzorků AT pěstováno při 16 °C a 6 °C [28]. Pro tyto teploty byly zaznamenávány růstové parametry spolu s profily metabolitů, aby bylo možné zkoumat vliv genomu a prostředí na variabilitu metabolomu [28]. Metabolity byly extrahovány a měřeny na plynové chromatografii spojené s hmotnostní spektrometrií, blíže popsané ve studii [28]. Výsledkem byly datasety obsahující růstové parametry a hladiny 37 metabolitů pro 241 linií vzorků AT, vždy pro 16 °C a 6 °C. [28]

V rámci předzpracování získaných dat byly ve studii [28] nejprve normalizovány absolutní hladiny metabolitů, pro dosažení Gaussovského rozložení. Dále byla data metabolomického profilu každého vzorku centrována na medián, bližší informace ve studii [28]. Poté byl každý metabolit standardizován, tzn. průměr $= 0$ a rozptyl $= 1$. Nakonec byla data zprůměrována přes všechna opakovaní dané linie vzorků. Výsledkem byla jedna hodnota pro každý z 37 metabolitů pro všech 241 linií vzorků v obou teplotních přístupech. Na takto předzpracovaná metabolomická data dále navazuje tato bakalářské práce. [28]



Obr. 3.1: Obrázek znázorňující tok biologické informace od genomu k fenotypu. Převzato a upraveno z [29].

4 Implementace predikčních modelů

Tato kapitola se zaměřuje na praktickou aplikaci teoretických znalostí získaných v kapitolách 1 a 2. První část popisuje předzpracování dat. Druhá část se zaměřuje na implementaci konvenčních predikčních modelů RR, LASSO a RF. Třetí část se věnuje implementaci a optimalizaci LSTM sítě. Pro úpravu a práci s daty jsou použité knihovny *NumPy* [30] a *Pandas* [31]. Pro implementaci konvenčních predikčních modelů je použita ML knihovna *Scikit-learn* [32]. V rámci implementace sítě LSTM je použita ML knihovna *TensorFlow* [33]. Všechny kódy jsou zpracovány v programovacím jazyce *Python* verze 3.12.2 [34].

4.1 Předzpracování dat

Prvním krokem předzpracování dat je přetransformování SNP datasetu na binární reprezentaci SNP. Původní SNP dataset je reprezentován pomocí desetinných čísel, které jsou v rozmezí hodnot $\langle 0, 1\rangle$. Neceločíselné hodnoty v původním SNP datasetu udávají, že na dané pozici je jen určitá míra polymorfismus. Pro neceločíselné hodnoty je v rámci binární reprezentace přiřazena hodnota 1, tzn. na dané pozici se vyskytuje polymorfismus. Výsledkem je binární reprezentace SNP datasetu.

Dále se práce v rámci předzpracování dat věnuje několika částem kontroly datasetů. První část se věnuje kontrole zachování sekvenční posloupnosti SNP datasetu. Jednotlivé pozice SNP datasetu jsou charakterizovány číslem chromozomu, na kterém se nacházejí a číslem pozice SNP na daném chromozomu. Proto byla naprogramovaná funkce pro kontrolu sekvenční posloupnosti těchto jednotlivých chromozomů a pozic SNP, které hrají významnou roli v aplikací LSTM sítí. Druhá část se věnuje kontrole binárních vlastností přetransformovaného binárního SNP datasetu. Třetí část se věnuje odstranění prázdných hodnot reprezentujících jako NaN (*angl. Not a Number*) hodnot binárního SNP datasetu a metabolomických datasetů. Čtvrtá část předzpracování dat se věnuje kontrole odlehlých hodnot (*angl. outliers*) u metabolomických dat. Tato kontrola byla provedena vizuální kontrolou box-plotů všech 37 metabolitů.

Díky fázi předzpracování jsou data vhodně transformována na vstupy pro algoritmy predikčního modelování. Data jsou nejprve načtena za pomoci knihovny *Pandas* [31]. Poté jsou data vhodně transformována na vstupní formát vybraných predikčních metod. Formát predikčních metod je v podobě matic/vektorů hodnot. Pro SNP dataset je to matice hodnot jednotlivých SNP pro všech 241 linií vzorků AT, kde každý řádek matice reprezentuje 27 081 hodnot SNP pro jednu linii vzorků AT. Pro metabolomický dataset je to vektor 241 hodnot hladin daného metabolitu. Výsledkem jsou hodnoty SNP pro jednotlivé linie vzorků a jim odpovídající hladiny

metabolitů. Následně jsou tyto matice hodnot pro jednotlivých 241 linií vzorků AT náhodně rozděleny na trénovací a testovací set za pomoci funkce *train_test_split()*, dostupné z knihovny *Scikit-learn* [32]. Trénovací a testovací set je rozdělen v poměru 180/61 linii vzorků AT, kdy je velikost trénovacího setu rovna přibližně trojnásobku setu testovacího. Trénovací set slouží k natrénování jednotlivých modelů. Testovací set slouží k otestování úspěšnosti jednotlivých modelů.

4.2 Regularizované regresní modely a náhodný les

Pro každý predikční model byl vytvořen skript, v kterém se nachází vytvořená funkce s obecným názvem *Model_computation()*. Vstupy a výstupy funkce *Model_computation()* jsou vidět na obrázku v příloze A. Pro natrénování modelů je použit předzpracovaný binární SNP dataset a metabolomické data. Predikce je provedena pro každý metabolit a pro obě teplotní kultivační podmínky, tedy 6 a 16°C.

Každá funkce *Model_computation()* nejprve rozdělí data na trénovací a testovací dataset. Následně za pomoci vhodné funkce z knihovny *Scikit-learn* [32] natrénuje model. Nakonec jsou za pomoci modelu predikované výsledné hodnoty pro trénovací a testovací dataset a jsou pro model vypočteny hodnotící metriky. Obecná struktura funkce *Model_computation()* je znázorněna na diagramu v příloze A.

Pro natrénování RR modelu je použita funkce *RidgeCV()* dostupná z knihovny *Scikit-learn* [32]. Jedná se o RR se zabudovanou křížovou validací pro volbu nejvhodnějšího hyperparametru λ, který ovlivňuje nastavení jednotlivých parametrů modelu, pro co nejlepší natrénování modelu. Funkce je z hlediska hyperparametru λ použita v základním nastavení, kdy funkce vybírá z předem definovaného vektoru hodnot λ [32]. Navíc model používá i křížovou validaci, pro zobecnění a reprodukovatelnost výsledků modelových predikcí. Pro natrénování LASSO modelu je použita funkce *LassoCV()* dostupná z knihovny *Scikit-learn* [32]. Funkce *LassoCV()* navíc oproti *RidgeCV()* umožňuje iterativní automatické nastavení hodnot λ. Funkce *LassoCV()* také využívá křížovou validaci. Dále je pro funkci nastaven maximální počet iterací na hodnotu 10 000, aby docházelo ke konvergenci při iterativním hledání nejvhodnějšího hyperparametru λ.

Pro natrénování RF modelu je použita funkce *RandomForestRegressor()* dostupná z knihovny *Scikit-learn* [32]. Pro zvolení nejlepší kombinace hyperparametrů RF modelu je použita funkce *GridSearchCV* dostupná z knihovny *Scikit-learn* [32]. Do funkce *GridSearchCV* vstupuje zvolený model, hodnota k-násobné křížové validace a zvolená mřížka hyperparametrů, která byla vhodně nastavena při předzpracování hyperparametrů modelu RF, podrobněji popsané v následujících dvou odstavcích. Funkce *GridSearchCV* se v mřížce snaží najít tu nejlepší kombinaci hyperparametrů. K-násobná křížová validace je nastavena na hodnotu 3, tzn. každá

podmnožina obsahuje 60 vzorků trénovacích dat. I přes omezení hyperparametrů je natrénování RF modelu výrazně časově náročnější než u regularizovaných regresních modelů.

Nastavení hyperparametrů modelu RF

Pro náhodný les je proveden krok předzpracování navíc, který se zaměřuje na optimalizaci natrénování modelu. Optimalizace se věnuje velkému množství hyperparametrů, které je potřeba vhodně nastavit. Pro tyto účely je vytvořena mřížka vhodných hyperparametrů, ze které model vybírá během svého trénování. Mřížka vhodných hyperparametrů je vytvořena za pomoci funkce *RandomizedSearchCV()*, dostupné z knihovny *Scikit-learn* [32]. Tato funkce využívá křížové validace k nastavení a otestování náhodných kombinací hyperparametrů modelu, které se nastavují z námi vložené mřížky hyperparametrů. Výsledkem je nejlepší kombinace hyperparametrů modelu, která byla během námi nastavených iterací nalezena. Pro potřeby této práce je nastaven počet iterací na hodnotu 100 a křížová validace na hodnotu 3, tedy jeden násobek křížové validace je roven 60 vzorkům trénovacích dat.

Mezi vložené hyperparametry v mřížce patří:

* počet stromů,
* maximální počet vstupů branných v úvahu při každém dělení stromu,
* minimální počet vzorků potřebných pro rozdělení uzlu,
* minimální počet vzorků potřebných v koncovém uzlu/listu,
* volba metody pro výběr vzorků při trénování jednotlivých stromů.

Nakonec je podle nalezené nejlepší kombinace hyperparametrů vhodně vytvořena výsledná mřížka hyperparametrů, která je použita při trénování modelu náhodného lesa.

4.3 Sítě s dlouhou-krátkodobou pamětí

Sestavení a natrénování sítí LSTM je implementováno v jazyce *python* a ve funkci *LSTM_single_metabolit_predictor*. Tato funkce umožňuje natrénování sítě LSTM vždy pro jeden metabolit. Mezi vstupy funkce *LSTM_single_metabolit_predictor* patří: počet epoch, velikost dávky (*angl. batch size*), rychlost učení (*angl. learning rate*), identifikační číslo LSTM modelu, cesta k uloženým datům, název metabolitu pro který má být síť LSTM natrénována a název metabolomického datasetu. Počet epoch, velikost dávky a rychlost učení jsou hyperparametry sítě LSTM, na které se tato práce zaměřuje z hlediska optimalizace sítě. Výstupem funkce

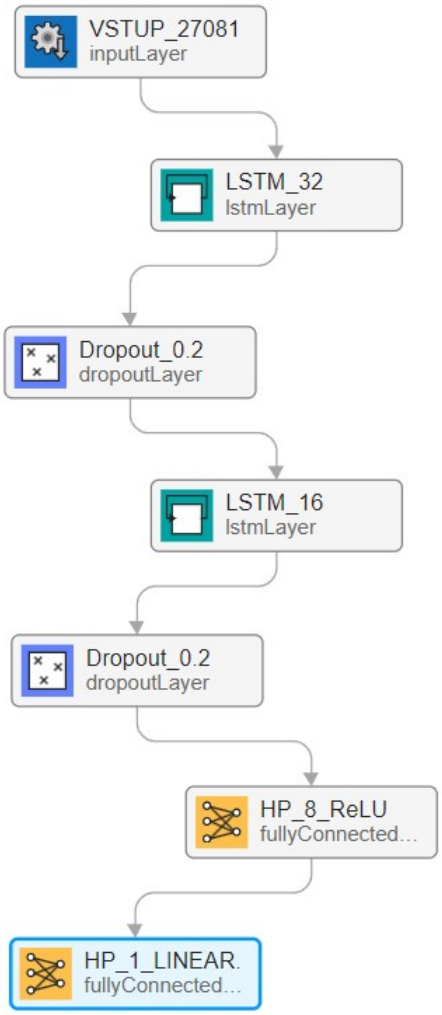
LSTM_single_metabolit_predictor jsou predikované hladiny metabolitu, hodnoty Pearsonova KK a MSE pro trénovací a testovací množinu dat. Pro natrénování sítě LSTM je použit předzpracovaný binární SNP dataset a metabolomická data. Predikce je provedena pro každý metabolit naměřený za teplotních podmínek: 6 °C a 16 °C.

Funkce *LSTM_single_metabolit_predictor* nejprve načte data a vhodně je upraví na trénovací a testovací dataset. Následně za pomoci funkcí *Sequential()*, *InputLayer()*, *LSTM()* a *Dense()*, dostupných z knihovny *TensorFlow* [33], sestaví architekturu sítě LSTM.

Pro ztrátovou funkci, která hodnotí odchylku predikovaných hodnot od reálných, je nastavena metrika MSE, která je použita i jako hodnotící metrika modelů. Pro optimalizaci je použit tzv. Adam Optimizer [35]. Adam Optimizer patří mezi nejpoužívanější iterační optimalizační algoritmy, které se používají k minimalizaci ztrátové funkce při trénování neuronových sítí.

Dále je funkcí *LSTM_single_metabolit_predictor* nastavena kontrola, která se snaží zabránit přeučení sítě LSTM. Tato kontrola spočívá v monitorování validační odchylky ztrátové funkce. Pokud v pěti epochách po sobě nedojde ke snížení této validační odchylky, funkce *LSTM_single_metabolit_predictor* ukončí trénování sítě a uloží nejlepší nalezené váhy sítě LSTM. Následně je za pomoci sestavené architektury a funkce *model.fit()*, dostupné z knihovny *TensorFlow* [33], síť natrénována. Mezi vstupní parametry funkce *model.fit()* patří: trénovací množina dat, počet epoch, velikost dávky, validační poměr dat a výše zmíněná kontrola přeučení sítě LSTM. Validační poměr je nastaven na hodnotu 30 % z celkových 180 vzorků trénovacích dat, což odpovídá 126 vzorkům trénovacích dat, 54 vzorkům validačních dat a 61 vzorkům testovacích dat. Validační data slouží k otestování generalizační schopnosti modelu sítě a k omezení nadměrné adaptace sítě na trénovací data během trénování sítě. Výsledkem natrénování sítě je uložení nejlepších nalezených vah modelu. Nakonec funkce *LSTM_single_metabolit_predictor* použije natrénovanou síť LSTM k predikci na trénovacích a testovacích datech a dopočítá hodnotící metriky modelu LSTM.

Funkce *LSTM_single_metabolit_predictor* vytváří jednoduchou jednosměrnou architekturu sítě LSTM, která z mnoha vstupů predikuje jeden výstup, v našem případě hladinu konkrétního metabolitu. Schéma navrhnuté architektury LSTM sítě je možné vidět na obrázku 4.1. Architektura LSTM sítě začíná vstupní vrstvou, jejíž velikost odpovídá počtu vstupních hodnot modelu sítě. Následují dvě vrstvy LSTM, které obsahují hustě propojené paměťové jednotky LSTM. Ke každé vrstvě LSTM je připojena tzv. *Dropout vrstva*, která se snaží zabránit přeučení modelu. *Dropout vrstva* u přiřazené vrstvy nastavuje některé jednotky na nulovou hodnotu. Nulováním jednotek *dropout vrstva* v každém běhu vytváří náhodně modifikovanou



Obr. 4.1: Schéma znázorňující vytvořenou architekturu sítě s dlouhou-krátkodobou pamětí za pomoci funkce *LSTM_single_metabolit_predictor*. Vytvořeno v MATLABU [36].

architekturu sítě, tzn. v každém běhu je model vystaven jiné množině vstupních hodnot. Hodnota *dropout vrstev* je nastavena na 20 %, tzn. že v každém běhu je 20 % jednotek v přiřazených vrstvách vynulováno. LSTM a *dropout vrstvy* následují dvě hustě propojené vrstvy neuronů. Tyto hustě propojené vrstvy zpracovávají informace získané z předchozích vrstev LSTM. Druhá hustě propojená vrstva, která je zároveň poslední vrstvou architektury sítě, obsahuje jen jeden neuron. Tento jeden neuron obsahuje lineární aktivační funkci, která odpovídá finální regresi, jejíž výstupem je hladina daného metabolitu. Počet paměťových jednotek v LSTM vrstvách je nastaven na hodnoty 32 a 16. Počet neuronů v první hustě propojené vrstvě je nastaven na hodnotu 8 a jejich aktivační funkce je nastavena na tzv. ReLU ak-

tivační funkci. Výsledkem této architektury sítě LSTM je 7 633 adaptovatelných parametrů modelu. Počet jednotek v jednotlivých vrstvách ovlivňuje výsledný počet adaptovatelných parametrů, které ovlivňují výslednou výpočetní a časovou náročnost modelu sítě LSTM. Tato výpočetní a časová náročnost je velkou výzvou, proto je na ni kladen důraz při vytváření architektury LSTM sítě.

Optimalizace hyperparametrů pro sítě LSTM

Tato práce se zaměřuje z hlediska optimalizace sítě LSTM na tyto hyperparametry: počet epoch, velikost dávky a rychlost učení. Je zvolen experimentální přístup optimalizace, kdy jsou pozorovány průměrné hodnoty hodnotících metrik modelu sítě pro různá nastavení již zmíněných hyperparametrů. Pro počet epoch jsou pozorovány výsledky hodnotících metrik pro hodnoty 4, 8 a 12 epoch. Zbylé dva hyperparametry jsou při testování počtu epoch nastaveny následovně: velikost dávky = 60 a rychlost učení = 0.001. Pro velikost dávky jsou pozorovány výsledky hodnotících metrik pro hodnoty 20, 40 a 60. Nastavení zbylých hyperparametrů je při testování velikosti dávky následovné: počet epoch = 4 a rychlost učení = 0.001. Pro rychlost učení jsou pozorovány výsledky hodnotících metrik pro hodnoty 0.01, 0.001 a 0.0001. Ostatní hyperparametry jsou při testování rychlosti učení nastaveny následovně: počet epoch = 4 a velikost dávky = 60.

Z těchto výše zmíněných nastavení je vybrána nejlepší kombinace hyperparametrů, u kterých vycházely nejlepší průměrné hodnoty hodnotících metrik modelu sítě LSTM. Výsledná nejlepší kombinace hyperparametrů vyšla následovně: počet epoch = 12, velikost dávky = 20 a rychlost učení = 0.01. Tato nalezená nejlepší kombinace hyperparametrů je použita pro finální natrénování a predikci za pomoci modelu sítě LSTM.

5 Genomická predikce v *Arabidopsis thaliana*

Tato kapitola shrnuje a hodnotí výsledky GP získané za pomoci implementovaných predikčních modelů popsaných v kapitolách 2 a 4. Kapitola je rozdělena do dvou částí. V první části jsou popsány a zhodnoceny výsledky konvenčních predikčních modelů RR, LASSO a RF. Druhá část popisuje a hodnotí výsledky GP získané za pomoci implementovaných sítí LSTM.

5.1 Genomické predikce za pomoci RR, LASSO a RF

Výsledky GP konvenčních modelů pro všech 37 metabolitů při 6 °C a 16 °C je možné vidět v přiložených tabulkách v příloze B. Tyto tabulky obsahují vypočtené hodnotící metriky pro modely RR, LASSO a RF. Každá hodnotící metrika je uvedena pro trénovací a testovací část dat. Pro vizualizaci výsledků GP konvenčních modelů jsou uvedeny následující dva přiložené grafy v příloze C. Každý graf znázorňuje výsledné hodnoty Pearsonova KK v % pro testovací část dat. První graf obsahuje výsledky pro metabolity při 6 °C a druhý graf pro metabolity při 16 °C.

Na prvním grafu C.1 lze vidět, že nejlepší výsledky predikce pro metabolity při 6 °C dosahují hodnot kolem 60 % Pearsonova KK. Z pohledu všech konvenčních modelů se mezi nejlépe predikované metabolity při 6 °C řadí fruktóza, kyselina fumarová a threitol. Tyto tři metabolity překročily hranici 40 % Pearsonova KK u všech konvenčních přístupů. Model RR překonal hranici 50 % Pearsonova KK pro kyselinu citronovou, galaktózu, maltózu a serin. Model LASSO nejlépe predikoval fruktózu, galaktinol, glukózu, kyselinu glutamovou a raffinózu, jejichž hodnoty překročily hranici 50 % Pearsonova KK. Model RF si nejlépe vedl při predikci fruktózy, kyseliny fumarové, kyseliny mléčné, myo-inositolu a spermidinu, jejichž hodnoty také překročily hranici 50 % Pearsonova KK. Pro predikci spermidinu model RF dokonce překročil hranici 60 % Pearsonova KK.

Na druhém grafu C.2 pro metabolity při 16 °C dosahují nejlepší predikce hodnot 50 až 60 % Pearsonova KK. Hranici 40 % Pearsonova KK pro všechny tři konvenční přístupy překročila predikce galaktinolu, glycinu, kyseliny mléčné, raffinózy a kyseliny threoniové. U kyseliny threoniové model RF dokonce dosahoval predikce vyšší než 60 % Pearsonova KK. Modely RR a RF si navíc vedly dobře při predikci kyseliny fumarové, kyseliny jablečné, myo-inositolu a spermidinu, jejichž predikce přesáhly hodnotu 40 % Pearsonova KK. Pro myo-inositol a modely RR a RF predikce sahala přes hodnotu 55 % Pearsonova KK.

Z grafů se dá vyčíst, že je rozdíl mezi predikcemi metabolitů při 6 °C a predikcemi metabolitů při 16 °C. Například jde vidět, že model LASSO si lépe vedl při predikci metabolitů při 6 °C. Dále je vidět, že určité metabolity při 6 °C vykazují dobré výsledky predikce a naopak stejné metabolitů při 16 °C nevykazují zas tak dobré výsledky predikce. Stejně tak určité metabolity při 16 °C vykazují lepší hodnoty predikce než u 6 °C. Například predikce fruktózy při 6 °C dosahovala hodnot 50 % Pearsonova KK. Naopak predikce fruktózy při 16 °C nedosahovala ani hodnot 40 % Pearsonova KK. Predikce kyseliny threoniové zase vykazovala velice dobré výsledky při 16 °C, ale při 6 °C jsou výsledky predikce výrazně horší.

Pro lepší zhodnocení toho, který model všeobecně dosahoval nejlepší výkonnosti při predikci jednotlivých metabolitů, jsou zde přiloženy následující dvě tabulky 5.1 a 5.2. Tabulky obsahují průměrné hodnoty hodnotících metrik pro predikci na testovacích datech za pomoci konvenčních modelů a každá tabulka obsahuje jeden z teplotních přístupů 6 °C a 16 °C.

Model:	**|KK| [%]**	**MSE**	**Teplota**
RR	28.566	0.210	6 °C
LASSO	26.532	0.206	6 °C
RF	30.955	0.200	6 °C

Tab. 5.1: Tabulka průměrných hodnot hodnotících metrik konvenčních modelů pro predikci na testovacích datech a pro metabolity při 6 °C.

Model:	**|KK| [%]**	**MSE**	**Teplota**
RR	29.786	0.219	16 °C
LASSO	22.379	0.223	16 °C
RF	30.336	0.212	16 °C

Tab. 5.2: Tabulka průměrných hodnot hodnotících metrik konvenčních modelů pro predikci na testovacích datech a pro metabolity při 16 °C.

Z tabulek je patrné, že si v obou teplotních přístupech nejlépe vedl model RF, následující modelem RR a nejhůře si vedl model LASSO. Jediná výjimka nastala z hlediska průměrné hodnoty MSE při 6 °C, kdy si vedl model LASSO o trošičku lépe než model RR, ale rozdíl je téměř nepatrný a z hlediska KK si lépe vedl model RR. Dále je možné vidět, že si modely vedly lépe při predikci metabolitů při 6 °C, až na výjimku modelu RR, který si z hlediska průměrné hodnoty KK vedl lépe u metabolitů při 16 °C. Všechny průměrné hodnoty hodnotících metrik jsou si však v obou teplotních přístupech blízké. Největší rozdíl je z hlediska průměrné hodnoty KK pro model LASSO, jenž byl patrný i z grafů.

V přiřazených tabulkách B je vidět, že úspěšnost predikce na trénovacích datech je u modelů RR a RF velmi vysoká, neklesá pod hodnotu 93 % Pearsonova KK. To ukazuje na přílišné přizpůsobení modelů RR a RF trénovacím datům a nižší generalizační schopnost modelů. Pro model LASSO se průměrné hodnoty KK pohybují kolem 60 %, čemuž však odpovídají i nižší hodnoty KK při predikci na testovacích datech oproti modelům RR a RF.

Výsledky predikce získané pomocí modelů RR, LASSO a RF dokazují, že jsou modely schopny natrénování se na dostupných genomických datech. Získané výsledky však poukazují i na nedostatky metod RR a LASSO založených na VLR. VLR není vhodnou volbou pro řešení problému vícerozměrné predikce, jelikož data v takovém případě obsahují často spoustu příznaků, nebo-li vysokou dimenzionalitu a vykazují složitou korelační strukturu mezi příznaky. Dalším problémem pro VLR je, když počet příznaků převyšuje počet pozorování. Genomická data, která jsou základem této práce, obsahují mnohem více příznaků, než je počet pozorování a obsažené informace mohou být vysoce korelované. Z těchto důvodů nejsou metody používající VLR nejvhodnější volbou pro aplikaci na genomická data a je potřeba hledat další přístupy v rámci ML, které by mohly dosahovat lepších výsledků a lépe hledat složité vztahy obsažené v genomických datech. Sofistikovanější metody založené na stromových strukturách, jako je RF, by mohly být lepší volbou, jak nám dokazují i výsledky v tabulkách 5.1 a 5.2. Dále je potřeba více prozkoumat i složitější metody založené na DL, čemuž se věnuje tato práce a další část této kapitoly.

5.2 Genomická predikce za pomoci sítí LSTM

Výsledky GP za pomoci sítí LSTM pro všech 37 metabolitů při 6 °C a 16 °C je možné vidět v přiložené tabulce v příloze D. Tato tabulka obsahuje vypočtené hodnotící metriky pro sítě LSTM. Každá hodnotící metrika je uvedena pro trénovací a testovací část dat. Pro vizualizaci výsledků GP za pomoci sítí LSTM jsou uvedeny heat mapy dostupné v příloze E. Heat mapy znázorňují reálné hodnoty hladin metabolitů a predikované hodnoty hladin metabolitů za pomoci sítí LSTM, pro obě teplotní podmínky 6 a 16 °C.

Na heat mapách je možné porovnat rozsahy reálných a predikovaných hodnot. Z hlediska rozsahu hodnot můžeme vidět, že rozsah predikovaných hodnot je ve všech případech menší než rozsah hodnot reálných. Dále je u predikovaných hodnot vidět, že variabilita je oproti reálným hodnotám pro některé metabolity téměř nulová, tzn. že se síť pro tyto metabolity nebyla schopna efektivně natrénovat. Naopak například u fruktózy a kyseliny mléčné při 6 °C a leucinu a serinu při 16 °C je vidět, že se zde nachází určitá míra variability predikovaných hodnot, což naznačuje určitou míru

natrénování sítě LSTM. Pro tyto výše zmíněné metabolity vycházejí i vyšší hodnoty Pearsonova KK, což je možné vidět na přiloženém grafu F.1 v příloze F.

Graf predikcí za pomoci sítí LSTM F.1 zobrazuje hodnoty Pearsonova KK v % pro testovací část dat a teplotní podmínky 6 a 16 °C. Z grafu je patrné, že si LSTM všeobecně pro všechny metabolity nevedla tak dobře jako konvenční modely. Rozdíl oproti konvenčním modelům je také v tom, že se u predikce alaninu a glutaminu při 6 °C a u kyseliny oxoglutarové při 16 °C vyskytla hodnota zajímavá z pohledu záporné korelace Pearsonova KK, což u konvenčních modelů nenastalo. Tato záporná hodnota by také mohla naznačovat určitou míru natrénování sítě LSTM. Nejvíce zajímavé jsou však hodnoty predikce u kyseliny mléčné při 6 °C a u kyseliny asparagové, kyseliny glutamové a leucinu při 16 °C, u nichž hodnoty predikce za pomoci sítí LSTM předčily modely RR, LASSO a RF, z pohledu Pearsonova KK.

V přiřazené tabulce D je vidět, že úspěšnost predikce na trénovacích datech také všeobecně nedosahuje vysokých hodnot. Dokonce pro některé metabolity jsou hodnoty pro trénovací dataset nízké, ale pro testovací dataset jsou vyšší. To ukazuje na nižší generalizační schopnost a nedostatečné natrénování sítí LSTM.

Výsledky GP získané za pomoci sítí LSTM výkonnostně nepřekonaly konvenční modely RR, LASSO a RF. Přesto se nalezly případy, kdy se LSTM dařilo a někdy dokonce lépe než konvenčním modelům, což podporuje myšlenku toho, že jsou LSTM sítě schopny nalézt jiné hlubší vztahy v datech. Přístupy DL jsou však velice náročné na velikost trénovacích dat, což může být jeden z důvodů, který komplikoval predikci za pomoci sítí LSTM v této práci. Dále by bylo potřeba více prozkoumat optimalizaci sítí LSTM, protože v tomto ohledu se nachází spousta možností, které si však žádají velké množství času a hlubší zkoumání. Například by se dalo zaměřit více na optimalizaci konkrétního metabolitu, než na optimalizaci z pohledu všech zkoumaných metabolitů v této práci. Také by se dala zkoumat vícerozměrná GP, např. GP více metabolitů najednou, což by mohlo vést zase k jiným zajímavým výsledkům.

Závěr

Tato bakalářská práce se zabývá problematikou genomické predikce u rostlin s využitím predikčních metod založených na strojovém učení. Práce využívá genomická data popsána pomocí jednonukleotidových polymorfismů. Cílem práce bylo zpracovat potřebnou teorii, která je zaměřena na genomickou predikci a její aplikace u rostlin, na predikční metody a modely strojového učení využívané pro genomické predikce rostlin a hodnotící metriky predikčních modelů.

První část práce je zaměřená na teoretickou část a popisuje i detailněji analyzovaná data z *Arabidopsis thaliany*. Cílem praktické části práce bylo implementovat vybrané konvenční modely strojového učení a sítě LSTM pro genomickou predikci rostlin v programovacím jazyce *Python*. Implementace zahrnovala také předzpracování dat a transformaci dat na vstupy, které jsou vhodné pro algoritmy predikčního modelování. Cílem implementace bylo zautomatizovat predikci pro dostupných 37 hladin metabolitů, které jsou uvedeny pro dvě teplotní podmínky 6 a 16 °C. Pro realizaci implementace byly použity funkce ze známých ML knihoven se správnou implementací do vytvořených funkcí v rámci této bakalářské práce. U sítí LSTM se práce navíc zaměřila na optimalizaci, která byla provedena experimentálním přístupem. Výsledkem práce bylo uložení výsledných hodnot hodnotících metrik modelů pro následné zhodnocení genomické predikce.

Hodnocení úspěšnosti predikce modelů bylo provedeno pomocí Pearsonova korelačního koeficientu a MSE. Na dostupných datech se povedlo otestovat a natrénovat konvenční modely RR, LASSO a RF. V nejlepších případech u nich predikce některých metabolitů dosahovaly kolem 60 % Pearsonova KK. Z pohledu všech predikovaných metabolitů si nejlépe vedl model RF, jehož predikce průměrně dosahovaly hodnot mezi 30 až 31 % Pearsonova KK. Nejhůře si vedl model LASSO, který měl v některých případech problém zpracovat předzpracovaná data této práce a nebyl schopen správného natrénování. Výsledky konvenčních přístupů poukázaly na nedostatky lineárních metod RR a LASSO. Naopak RF prokázal svou schopnost nalézt v datech složitější vztahy a vedl si lépe. Nakonec byly úspěšně implementovány sítě LSTM. Genomická predikce u nich v nejlepších případech dosahovala hodnot v rozmezí 20 až 55 % Pearsonova KK. Pro sítě LSTM se navíc objevily hodnoty zajímavé i z pohledu záporné korelace, dosahovaly hodnot -20 až -30 % Pearsonova KK. Průměrné hodnoty MSE pro genomickou predikci za pomoci sítí LSTM dosahovaly o něco horších výsledků než pro modely RR, LASSO a RF.

Literatura

[1] HO, Daniel Sik Wai; SCHIERDING, William; WAKE, Melissa; SAFFERY, Richard a O-SULLIVAN, Justin. Machine Learning SNP Based Prediction for Precision Medicine. Online. *Frontiers in Genetics*. 2019, roč. 10. ISSN 1664-8021. Dostupné z: https://doi.org/10.3389/fgene.2019.00267. [cit. 2024-04-24].

[2] RAIMONDI, Daniele; CORSO, Massimiliano; FARISELLI, Piero a MOREAU, Yves. From genotype to phenotype in Arabidopsis thaliana: in-silico genome interpretation predicts 288 phenotypes from sequencing data. Online. *Nucleic Acids Research*. 2022, roč. 50, č. 3, s. e16-e16. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkab1099. [cit. 2023-12-28].

[3] RIEDELSHEIMER, Christian; CZEDIK-EYSENBERG, Angelika; GRIEDER, Christoph; LISEC, Jan; TECHNOW, Frank et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Online. *Nature Genetics*. 2012, roč. 44, č. 2, s. 217-220. ISSN 1061-4036. Dostupné z: https://doi.org/10.1038/ng.1033. [cit. 2024-01-03].

[4] ORGOGOZO, Virginie; MORIZOT, Baptiste a MARTIN, Arnaud. The differential view of genotype-phenotype relationships. Online. *Frontiers in Genetics*. 2015, roč. 6, s. 179. ISSN 1664-8021. Dostupné z: https://doi.org/10.3389/fgene.2015.00179. [cit. 2024-01-03].

[5] BROOKES, Anthony J. The essence of SNPs. Online. *Gene*. 1999, roč. 234, č. 2, s. 177-186. ISSN 03781119. Dostupné z: https://doi.org/10.1016/S0378-1119(99)00219-X. [cit. 2024-01-03].

[6] HURTA, Martin; SCHWARZEROVA, Jana; NAEGELE, Thomas; WECKWERTH, Wolfram; PROVAZNÍK, Valentine et al. Utilizing Genetic Programming to Enhance Polygenic Risk Score Calculation. Online. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2023, s. 3782-3787. ISBN 979-8-3503-3748-8. Dostupné z: https://doi.org/10.1109/BIBM58861.2023.10385615. [cit. 2024-04-24].

[7] VISSCHER, Peter M.; BROWN, Matthew A.; MCCARTHY, Mark I. a YANG, Jian. Five Years of GWAS Discovery. Online. *The American Journal of Human Genetics*. 2012, roč. 90, č. 1, s. 7-24. ISSN 00029297. Dostupné z: https://doi.org/10.1016/j.ajhg.2011.11.029. [cit. 2024-01-03].

[8] TAM, Vivian; PATEL, Nikunj; TURCOTTE, Michelle; BOSSÉ, Yohan; PARÉ, Guillaume et al. Benefits and limitations of genome-wide association studies.

Online. *Nature Reviews Genetics*. 2019, roč. 20, č. 8, s. 467-484. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/s41576-019-0127-1. [cit. 2024-01-03].

[9] GRZYBOWSKI, Marcin W.; MURAL, Ravi V.; XU, Gen; TURKUS, Jonathan; YANG, Jinliang et al. A common resequencing-based genetic marker data set for global maize diversity: a genetic marker dataset with increased marker density facilitates association studies in maize. Online. *The Plant Journal*. 2023, roč. 113, č. 6, s. 1109-1121. ISSN 0960-7412. Dostupné z: https://doi.org/10.1111/tpj.16123. [cit. 2024-01-03].

[10] MONTESINOS-L-PEZ, Abelardo; MONTESINOS-L-PEZ, Osval A; GIANOLA, Daniel; CROSSA, José a HERNÁNDEZ-SUÁREZ, Carlos M. Multi-environment Genomic Prediction of Plant Traits Using Deep Learners With Dense Architecture. Online. *G3 Genes|Genomes|Genetics*. 2018, roč. 8, č. 12, s. 3813-3828. ISSN 2160-1836. Dostupné z: https://doi.org/10.1534/g3.118.200740. [cit. 2024-01-03].

[11] SIDAK, David; SCHWARZEROVÁ, Jana; WECKWERTH, Wolfram a WALDHERR, Steffen. Interpretable machine learning methods for predictions in systems biology from omics data. Online. *Frontiers in Molecular Biosciences*. 2022, roč. 9, s. 926623. ISSN 2296-889X. Dostupné z: https://doi.org/10.3389/fmolb.2022.926623. [cit. 2024-01-03].

[12] GONZÁLEZ GARCÍA, Cristian; NÚÑEZ-VALDEZ, Edward; GARCÍA-DÍAZ, Vicente; PELAYO G-BUSTELO, Cristina a CUEVA-LOVELLE, Juan Manuel. A Review of Artificial Intelligence in the Internet of Things. Online. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2019, roč. 5, č. 4, s. 9-20. ISSN 1989-1660. Dostupné z: https://doi.org/10.9781/ijimai.2018.03.004. [cit. 2024-01-03].

[13] CULE, Erika a DE IORIO, Maria. Ridge Regression in Prediction Problems: Automatic Choice of the Ridge Parameter: Automatic Choice of the Ridge Parameter. Online. *Genetic Epidemiology*. 2013, roč. 37, č. 7, s. 704-714. ISSN 0741-0395. Dostupné z: https://doi.org/10.1002/gepi.21750. [cit. 2024-01-03].

[14] OGUTU, Joseph O; SCHULZ-STREECK, Torben a PIEPHO, Hans-Peter. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions: ridge regression, lasso, elastic net and their extensions. Online. *BMC Proceedings*. 2012, roč. 6, č.

S10, s. p. 1-6. ISSN 1753-6561. Dostupné z: https://doi.org/10.1186/1753-6561-6-S2-S10. [cit. 2024-01-03].

[15] MUTHUKRISHNAN, R a ROHINI, R. LASSO: A feature selection technique in predictive modeling for machine learning: A feature selection technique in predictive modeling for machine learning. Online. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. 2016, s. 18-20. ISBN 978-1-5090-3769-8. Dostupné z: https://doi.org/10.1109/ICACA.2016.7887916. [cit. 2024-01-03].

[16] RODRIGUEZ-GALIANO, Victor F.; SANCHEZ-CASTILLO, Manuel; DASH, Jadunandan; ATKINSON, Peter M. a OJEDA-ZUJAR, Jose. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Online. *Biogeosciences*. 2016, roč. 13, č. 11, s. 3305-3317. ISSN 1726-4189. Dostupné z: https://doi.org/10.5194/bg-13-3305-2016. [cit. 2024-01-03].

[17] RODRIGUEZ-GALIANO, V.; SANCHEZ-CASTILLO, M.; CHICA-OLMO, M. a CHICA-RIVAS, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines: An evaluation of neural networks, random forest, regression trees and support vector machines. Online. *Ore Geology Reviews*. 2015, roč. 71, s. 804-818. ISSN 01691368. Dostupné z: https://doi.org/10.1016/j.oregeorev.2015.01.001. [cit. 2024-01-03].

[18] JOHANSSON, Ulf; BOSTRÖM, Henrik; LÖFSTRÖM, Tuve a LINUSSON, Henrik. Regression conformal prediction with random forests. Online. *Machine Learning*. 2014, roč. 97, č. 1-2, s. 155-176. ISSN 0885-6125. Dostupné z: https://doi.org/10.1007/s10994-014-5453-0. [cit. 2024-01-03].

[19] PALMER, David S.; O'BOYLE, Noel M.; GLEN, Robert C. a MITCHELL, John B. O. Random Forest Models To Predict Aqueous Solubility. Online. *Journal of Chemical Information and Modeling*. 2007, roč. 47, č. 1, s. 150-158. ISSN 1549-9596. Dostupné z: https://doi.org/10.1021/ci060164k. [cit. 2024-01-03].

[20] MONTESINOS-L-PEZ, Osval Antonio; MONTESINOS-L-PEZ, Abelardo; PÉREZ-RODRÍGUEZ, Paulino; BARR-N-L-PEZ, José Alberto; MARTINI, Johannes W. R. et al. A review of deep learning applications for genomic selection. Online. *BMC Genomics*. 2021, roč. 22, č. 1, s. 1-23. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/s12864-020-07319-x. [cit. 2024-01-03].

[21] PÉREZ-ENCISO, Miguel a ZINGARETTI, Laura M. A Guide for Using Deep Learning for Complex Trait Genomic Prediction. Online. *Genes*. 2019, roč. 10, č. 7, s. 553. ISSN 2073-4425. Dostupné z: https://doi.org/10.3390/genes10070553. [cit. 2024-01-03].

[22] EMMERT-STREIB, Frank; YANG, Zhen; FENG, Han; TRIPATHI, Shailesh a DEHMER, Matthias. An Introductory Review of Deep Learning for Prediction Models With Big Data. Online. *Frontiers in Artificial Intelligence*. 2020, roč. 3, s. 4. ISSN 2624-8212. Dostupné z: https://doi.org/10.3389/frai.2020.00004. [cit. 2024-01-03].

[23] FROST, Jim. *Mean Squared Error (MSE)*. Online. FROST, Jim. Statistics By Jim. C2023. Dostupné z: https://statisticsbyjim.com/regression/mean-squared-error-mse/. [cit. 2023-12-28].

[24] THE SCIPY COMMUNITY. *Scipy.stats.pearsonr*. Online. THE SCIPY COMMUNITY. SciPy documentation. C2008-2024. Dostupné z: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html. [cit. 2024-04-26].

[25] MUKAKA, M.M. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. Online. *Malawi Medical Journal*. 2012, roč. 24, č. 3, s. 69-71. ISSN 1995-7262. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/. [cit. 2024-04-26].

[26] PIÑEIRO, Gervasio; PERELMAN, Susana; GUERSCHMAN, Juan P. a PARUELO, José M. How to evaluate models: Observed vs. predicted or predicted vs. observed? Online. *Ecological Modelling*. 2008, roč. 216, č. 3-4, s. 316-322. ISSN 03043800. Dostupné z: https://doi.org/10.1016/j.ecolmodel.2008.05.006. [cit. 2024-05-15].

[27] ALONSO-BLANCO, Carlos; ANDRADE, Jorge; BECKER, Claude; BEMM, Felix; BERGELSON, Joy et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Online. *Cell*. 2016, roč. 166, č. 2, s. 481-491. ISSN 00928674. Dostupné z: https://doi.org/10.1016/j.cell.2016.05.063. [cit. 2024-01-03].

[28] WEISZMANN, Jakob; WALTHER, Dirk; CLAUW, Pieter; BACK, Georg; GUNIS, Joanna et al. Metabolome plasticity in 241 Arabidopsis thaliana accessions reveals evolutionary cold adaptation processes. Online. *Plant Physiology*. 2023, roč. 193, č. 2, s. 980-1000. ISSN 0032-0889. Dostupné z: https://doi.org/10.1093/plphys/kiad298. [cit. 2024-01-03].

[29] HAMANY DJANDE, Claude Y.; PRETORIUS, Chanel; TUGIZIMANA, Fidele; PIATER, Lizelle A. a DUBERY, Ian A. Metabolomics: A Tool for Cultivar Phenotyping and Investigation of Grain Crops. Online. *Agronomy*. 2020, roč. 10, č. 6. ISSN 2073-4395. Dostupné z: `https://doi.org/10.3390/agronomy10060831`. [cit. 2024-05-18].

[30] NUMPY TEAM. *NumPy*. Online. 2005. Dostupné z: `https://numpy.org/`. [cit. 2023-12-29].

[31] THE PANDAS DEVELOPMENT TEAM. *Pandas*. Online. 2009. Dostupné z: `https://pandas.pydata.org/`. [cit. 2023-12-29].

[32] PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre a MICHEL, Vincent. *Scikit-learn*. Online. C2007-2023. Dostupné z: `https://scikit-learn.org/stable/index.html`. [cit. 2023-12-29].

[33] ABADI, Martín; AGARWAL, Ashish; BARHAM, Paul; BREVDO, Eugene; CHEN, Zhifeng et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Online. 2015. Dostupné z: `https://www.tensorflow.org`. [cit. 2024-05-23].

[34] PYTHON SOFTWARE FOUNDATION. *Python*. Online. C2001-2023. Dostupné z: `https://www.python.org/`. [cit. 2023-12-29].

[35] BOCK, Sebastian a WEIS, Martin. A Proof of Local Convergence for the Adam Optimizer. Online. *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019, s. 1-8. ISBN 978-1-7281-1985-4. Dostupné z: `https://doi.org/10.1109/IJCNN.2019.8852239`. [cit. 2024-05-27].

[36] THE MATHWORKS, INC. *MathWorks*. Online. C1994-2024. Dostupné z: `https://www.mathworks.com/`. [cit. 2024-05-25].

Seznam symbolů a zkratek

GP Genomická Predikce

GWAS Celogenomová Asociační Studie – *angl. Genome-Wide Association Study*

WGS Celogenomové Sekvenování – *angl. Whole Genome Sequencing*

SNP Jednonukleotidové Polymorfismy – *angl. Single Nucleotide Polymorphism*

ML Strojové Učení – *angl. Machine Learning*

VLR Vícenásobná Lineární Regrese

RR Hřebenová Regrese – *angl. Ridge Regression*

LASSO Operátor Nejmenšího Absolutního Zmenšení a Výběru – *angl. Least Absolute Shrinkage and Selection Operator*

LS Nejmenší Čtverce – *angl. Least Squares*

RF Náhodný Les – *angl. Random Forest*

DT Rozhodovací Strom – *angl. Decision Tree*

RT Regresní Strom – *angl. Regression Tree*

DL Hloubkové Učení – *angl. Deep Learning*

FNN Dopředná Neuronová Síť – *angl. Feedforward Neural Network*

RNN Rekurentní Neuronová Síť – *angl. Recurrent Neural Network*

CNN Konvoluční Neuronová Síť – *angl. Convolutional Neural Network*

LSTM Síť s Dlouhou-Krátkodobou Pamětí – *angl. Long Short-Term Memory*

CEC Konstantní Chybový Kolotoč – *angl. Constant Error Carousel*

MSE Střední Kvadratická Chyba – *angl. Mean Squared Error*

KK Korelační Koeficient

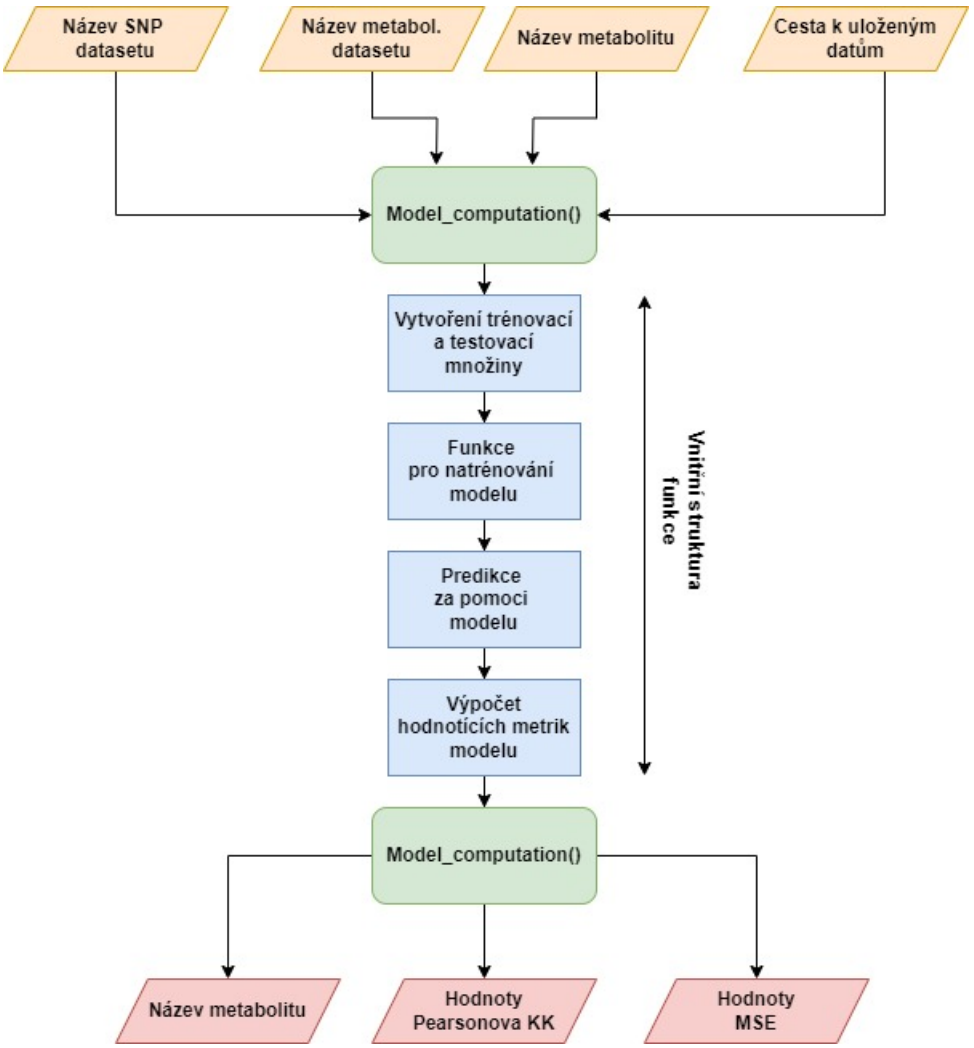
AT Huseníček rolní – *Arabidopsis thaliana*

NaN Prázdná Hodnota – *angl. Not a Number*

Seznam příloh

A Diagram pro funkci *Model_computation()*

V této části přílohy je vyobrazen diagram znázorňující obecnou strukturu funkce *Model_computation()*, která je použita pro implementaci konvenčních predikčních modelů RR, LASSO a RF.



Obr. A.1: Diagram znázorňující obecnou strukturu funkce *Model_computation()*.

B Tabulky výsledků genomické predikce pro konvenční modely

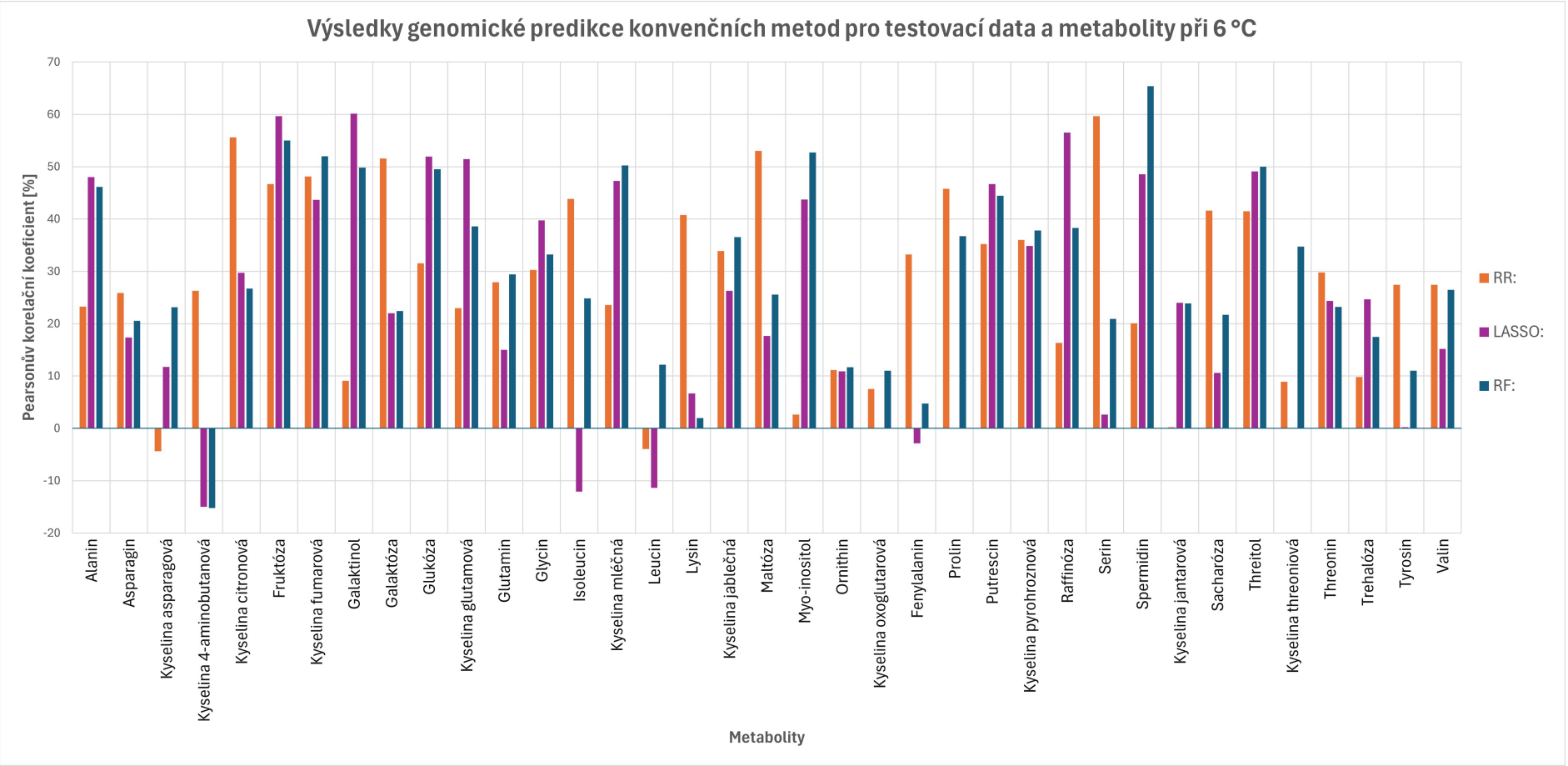
RR 6C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test	LASSO 6C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test	RF 6C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test
Alanin	99.883	23.278	0.000	0.181	Alanin	66.796	48.000	0.155	0.170	Alanin	94.221	46.138	0.044	0.164
Asparagin	99.887	25.844	0.000	0.180	Asparagin	59.194	17.341	0.159	0.177	Asparagin	97.607	20.534	0.025	0.174
Kyselina asparagová	99.976	-4.355	0.000	0.158	Kyselina asparagová	69.475	11.747	0.096	0.166	Kyselina asparagová	97.467	23.144	0.018	0.159
Kyselina 4-aminobutanová	99.996	26.309	0.000	0.196	Kyselina 4-aminobutanová	33.765	-14.975	0.146	0.164	Kyselina 4-aminobutanová	94.603	-15.252	0.041	0.177
Kyselina citronová	99.985	55.646	0.000	0.234	Kyselina citronová	62.740	29.748	0.165	0.230	Kyselina citronová	97.233	26.735	0.031	0.230
Fruktóza	99.996	46.687	0.000	0.387	Fruktóza	86.067	59.678	0.200	0.354	Fruktóza	96.317	54.990	0.067	0.380
Kyselina fumarová	99.981	48.111	0.000	0.304	Kyselina fumarová	72.662	43.683	0.197	0.334	Kyselina fumarová	97.860	51.978	0.030	0.296
Galaktinol	99.922	9.059	0.000	0.159	Galaktinol	87.561	60.151	0.096	0.132	Galaktinol	97.397	49.795	0.029	0.155
Galaktóza	100.000	51.602	0.000	0.284	Galaktóza	66.499	22.027	0.152	0.250	Galaktóza	98.046	22.440	0.020	0.250
Glukóza	99.492	31.534	0.000	0.437	Glukóza	88.943	51.919	0.194	0.417	Glukóza	97.271	49.529	0.063	0.431
Kyselina glutamová	99.716	22.978	0.001	0.097	Kyselina glutamová	67.639	51.482	0.071	0.085	Kyselina glutamová	96.284	38.622	0.014	0.090
Glutamin	99.982	27.889	0.001	0.177	Glutamin	57.703	14.999	0.138	0.177	Glutamin	96.158	29.421	0.031	0.167
Glycin	99.980	30.273	0.000	0.143	Glycin	63.905	39.755	0.159	0.138	Glycin	95.019	33.230	0.044	0.136
Isoleucin	99.921	43.877	0.000	0.173	Isoleucin	43.867	-12.118	0.137	0.183	Isoleucin	97.222	24.837	0.021	0.167
Kyselina mléčná	99.999	23.594	0.001	0.293	Kyselina mléčná	99.613	47.303	0.004	0.321	Kyselina mléčná	96.793	50.253	0.039	0.260
Leucin	99.708	-3.964	0.000	0.271	Leucin	69.945	-11.389	0.167	0.296	Leucin	96.319	12.143	0.035	0.270
Lysin	100.000	40.749	0.001	0.233	Lysin	53.963	6.649	0.155	0.206	Lysin	96.970	1.985	0.022	0.217
Kyselina jablečná	99.997	33.896	0.000	0.238	Kyselina jablečná	74.106	26.287	0.148	0.265	Kyselina jablečná	97.721	36.530	0.026	0.248
Maltóza	99.940	53.042	0.000	0.215	Maltóza	81.330	17.632	0.146	0.223	Maltóza	96.492	25.545	0.046	0.219
Myo-inositol	99.787	2.620	0.000	0.132	Myo-inositol	89.586	43.729	0.051	0.144	Myo-inositol	97.489	52.713	0.015	0.136
Ornithin	99.865	11.164	0.001	0.159	Ornithin	57.373	10.880	0.141	0.131	Ornithin	96.467	11.660	0.026	0.135
Kyselina oxoglutarová	99.976	7.543	0.001	0.259	Kyselina oxoglutarová	0.000	0.000	0.237	0.244	Kyselina oxoglutarová	94.860	10.987	0.053	0.243
Fenylalanin	100.000	33.240	0.000	0.465	Fenylalanin	59.997	-2.842	0.262	0.428	Fenylalanin	96.861	4.738	0.055	0.430
Prolin	99.998	45.755	0.000	0.119	Prolin	0.000	0.000	0.145	0.131	Prolin	96.650	36.754	0.024	0.115
Putrescin	99.404	35.189	0.000	0.176	Putrescin	65.722	46.695	0.139	0.189	Putrescin	95.390	44.468	0.034	0.187
Kyselina pyrohroznová	99.983	36.009	0.002	0.201	Kyselina pyrohroznová	62.459	34.844	0.116	0.204	Kyselina pyrohroznová	95.441	37.822	0.030	0.198
Raffinóza	99.944	16.323	0.000	0.151	Raffinóza	81.027	56.494	0.122	0.119	Raffinóza	97.257	38.274	0.027	0.141
Serin	99.992	59.653	0.000	0.220	Serin	99.867	2.650	0.001	0.287	Serin	96.911	20.885	0.027	0.199
Spermidin	99.998	20.048	0.000	0.114	Spermidin	89.630	48.566	0.058	0.136	Spermidin	94.112	65.417	0.037	0.103
Kyselina jantarová	100.000	0.219	0.000	0.093	Kyselina jantarová	84.232	24.003	0.044	0.087	Kyselina jantarová	94.870	23.884	0.021	0.086
Sacharóza	100.000	41.596	0.000	0.210	Sacharóza	67.579	10.619	0.117	0.178	Sacharóza	97.102	21.697	0.021	0.174
Threitol	99.974	41.512	0.000	0.106	Threitol	71.546	49.125	0.080	0.097	Threitol	95.568	49.999	0.022	0.095
Kyselina threoniová	99.938	8.925	0.000	0.116	Kyselina threoniová	0.000	0.000	0.153	0.138	Kyselina threoniová	94.445	34.757	0.037	0.121
Threonin	99.999	29.812	0.000	0.092	Threonin	72.457	24.370	0.074	0.075	Threonin	93.865	23.183	0.026	0.076
Trehalóza	99.997	9.816	0.000	0.188	Trehalóza	82.606	24.633	0.151	0.195	Trehalóza	97.393	17.479	0.027	0.200
Tyrosin	99.917	27.410	0.000	0.198	Tyrosin	51.913	0.192	0.169	0.163	Tyrosin	94.880	11.018	0.043	0.168
Valin	99.917	27.410	0.000	0.393	Valin	59.589	15.164	0.232	0.401	Valin	97.887	26.488	0.030	0.387

Tab. B.1: Tabulka výsledků hodnotících metrik pro konvenční modely a metabolity při 6 °C.

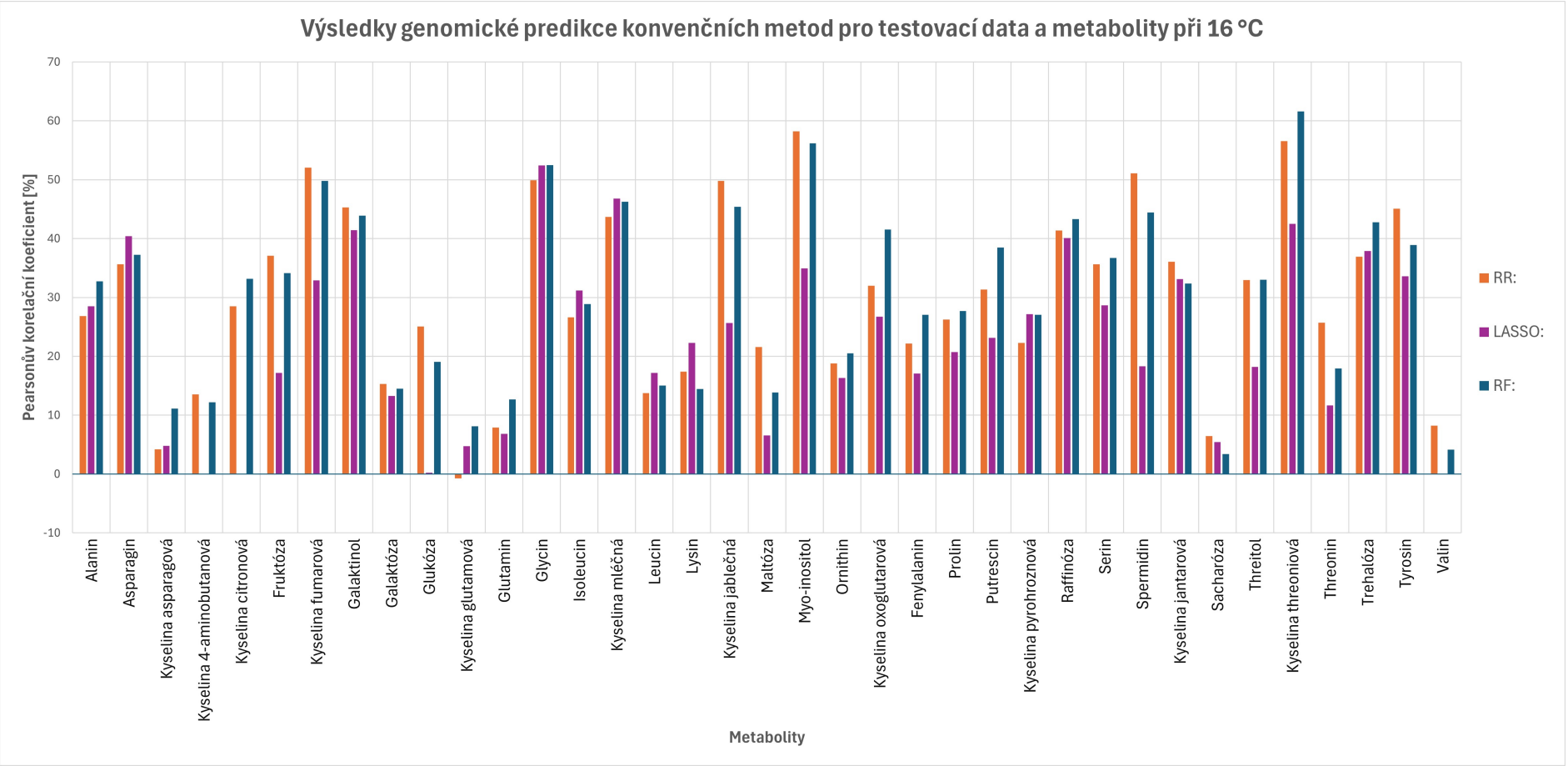
RR 16C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test	LASSO 16C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test	RF 16C:	PK:train [%]	PK:test [%]	MSE:train	MSE:test
Alanin	99.973	26.833	0.000	0.311	Alanin	76.823	28.486	0.165	0.302	Alanin	96.279	32.725	0.041	0.293
Asparagin	99.920	35.620	0.000	0.243	Asparagin	62.001	40.407	0.189	0.216	Asparagin	96.367	37.240	0.034	0.215
Kyselina asparagová	98.980	4.200	0.003	0.170	Kyselina asparagová	91.837	4.783	0.043	0.152	Kyselina asparagová	95.503	11.128	0.027	0.141
Kyselina 4-aminobutanová	100.000	13.549	0.000	0.244	Kyselina 4-aminobutanová	0.000	0.000	0.188	0.233	Kyselina 4-aminobutanová	97.997	12.220	0.019	0.232
Kyselina citronová	99.915	28.487	0.000	0.175	Kyselina citronová	0.000	0.000	0.155	0.187	Kyselina citronová	97.437	33.191	0.020	0.168
Fruktóza	99.945	37.099	0.000	0.271	Fruktóza	97.241	17.205	0.031	0.339	Fruktóza	95.383	34.113	0.065	0.271
Kyselina fumarová	99.268	52.043	0.005	0.241	Kyselina fumarová	63.890	32.882	0.259	0.298	Kyselina fumarová	95.951	49.797	0.058	0.257
Galaktinol	99.951	45.290	0.000	0.248	Galaktinol	89.288	41.438	0.078	0.257	Galaktinol	97.830	43.911	0.030	0.253
Galaktóza	99.999	15.331	0.000	0.176	Galaktóza	55.083	13.266	0.125	0.156	Galaktóza	97.083	14.495	0.021	0.162
Glukóza	99.927	25.044	0.000	0.196	Glukóza	38.281	0.256	0.216	0.205	Glukóza	96.667	19.058	0.040	0.198
Kyselina glutamová	99.600	-0.728	0.001	0.228	Kyselina glutamová	58.036	4.727	0.153	0.188	Kyselina glutamová	97.282	8.133	0.018	0.200
Glutamin	99.739	7.877	0.001	0.155	Glutamin	60.414	6.812	0.135	0.126	Glutamin	96.569	12.672	0.024	0.132
Glycin	99.985	49.886	0.000	0.135	Glycin	74.630	52.425	0.142	0.136	Glycin	94.222	52.509	0.045	0.134
Isoleucin	99.999	26.630	0.000	0.179	Isoleucin	54.102	31.176	0.135	0.168	Isoleucin	96.255	28.890	0.029	0.163
Kyselina mléčná	99.833	43.708	0.001	0.167	Kyselina mléčná	63.945	46.810	0.138	0.169	Kyselina mléčná	97.049	46.274	0.024	0.167
Leucin	99.959	13.764	0.000	0.326	Leucin	29.263	17.195	0.256	0.299	Leucin	95.989	15.028	0.053	0.302
Lysin	99.894	17.379	0.000	0.185	Lysin	76.059	22.281	0.111	0.168	Lysin	96.534	14.469	0.030	0.170
Kyselina jablečná	99.999	49.775	0.000	0.153	Kyselina jablečná	65.920	25.634	0.155	0.189	Kyselina jablečná	97.416	45.401	0.024	0.161
Maltóza	99.894	21.566	0.000	0.166	Maltóza	32.945	6.560	0.203	0.161	Maltóza	95.895	13.884	0.046	0.162
Myo-inositol	99.967	58.211	0.000	0.249	Myo-inositol	74.044	34.923	0.207	0.329	Myo-inositol	97.853	56.209	0.028	0.278
Ornithin	98.942	18.782	0.005	0.159	Ornithin	65.448	16.343	0.173	0.145	Ornithin	96.810	20.495	0.028	0.145
Kyselina oxoglutarová	99.956	31.993	0.000	0.257	Kyselina oxoglutarová	74.357	26.757	0.144	0.255	Kyselina oxoglutarová	94.625	41.524	0.044	0.232
Fenylalanin	99.652	22.157	0.002	0.429	Fenylalanin	66.582	17.092	0.231	0.431	Fenylalanin	96.856	27.066	0.048	0.413
Prolin	99.999	26.276	0.000	0.236	Prolin	43.073	20.723	0.180	0.245	Prolin	96.611	27.722	0.035	0.231
Putrescin	99.997	31.371	0.000	0.135	Putrescin	43.256	23.164	0.159	0.143	Putrescin	97.557	38.503	0.019	0.127
Kyselina pyrohroznová	99.995	22.303	0.000	0.235	Kyselina pyrohroznová	49.081	27.179	0.151	0.226	Kyselina pyrohroznová	97.287	27.069	0.021	0.219
Raffinóza	99.774	41.371	0.001	0.250	Raffinóza	99.542	40.117	0.003	0.258	Raffinóza	98.437	43.323	0.015	0.243
Serin	99.578	35.661	0.002	0.314	Serin	83.129	28.646	0.122	0.329	Serin	97.066	36.718	0.030	0.306
Spermidin	99.560	51.104	0.001	0.107	Spermidin	74.116	18.295	0.088	0.138	Spermidin	93.081	44.461	0.033	0.109
Kyselina jantarová	99.999	36.082	0.000	0.142	Kyselina jantarová	66.109	33.112	0.087	0.146	Kyselina jantarová	95.843	32.389	0.017	0.144
Sacharóza	98.484	6.480	0.005	0.190	Sacharóza	33.903	5.435	0.153	0.166	Sacharóza	93.981	3.422	0.040	0.174
Threitol	99.999	32.941	0.000	0.097	Threitol	53.780	18.190	0.092	0.098	Threitol	95.752	33.017	0.022	0.092
Kyselina threoniová	99.555	56.555	0.002	0.179	Kyselina threoniová	67.042	42.506	0.158	0.229	Kyselina threoniová	95.686	61.609	0.039	0.188
Threonin	99.923	25.737	0.000	0.108	Threonin	51.267	11.681	0.126	0.110	Threonin	97.200	17.917	0.022	0.109
Trehalóza	99.995	36.939	0.000	0.447	Trehalóza	89.986	37.908	0.135	0.450	Trehalóza	97.444	42.790	0.044	0.435
Tyrosin	99.997	45.088	0.000	0.222	Tyrosin	59.844	33.597	0.208	0.246	Tyrosin	95.161	38.925	0.062	0.236
Valin	99.926	8.218	0.000	0.389	Valin	0.000	0.000	0.282	0.352	Valin	97.475	4.124	0.034	0.371

Tab. B.2: Tabulka výsledků hodnotících metrik pro konvenční modely a metabolity při 16 °C.

C Grafy výsledků genomické predikce pro konvenční modely a metabolity při 6 a 16 °C



Obr. C.1: Graf znázorňující výsledky Pearsonova korelačního koeficientu pro predikci konvenčních modelů a metabolity při 6 °C.



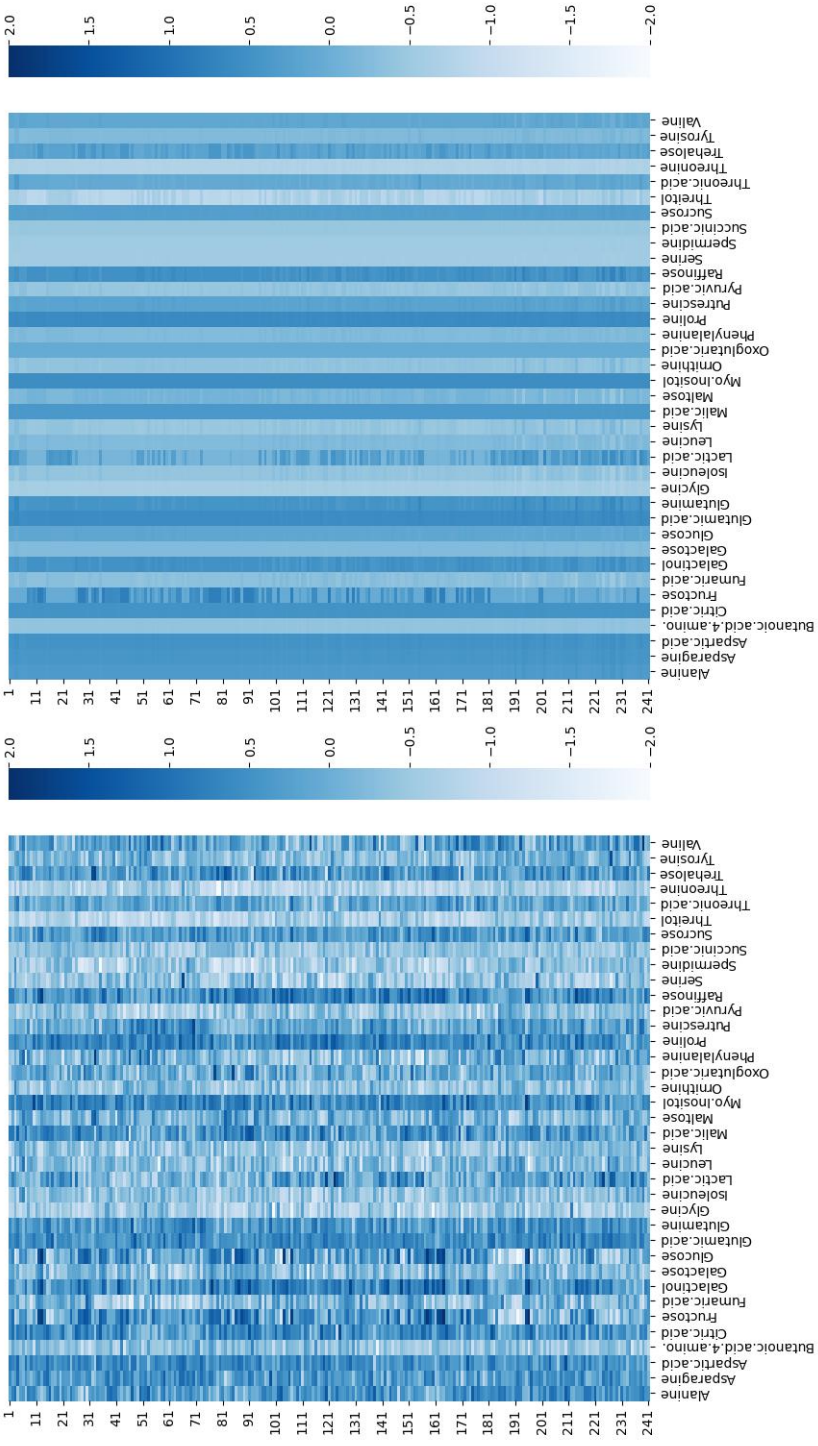
Obr. C.2: Graf znázorňující výsledky Pearsonova korelačního koeficientu pro predikci konvenčních modelů a metabolity při 16°C.

D Tabulka výsledků genomické predikce za pomoci sítí s dlouhou-krátkodobou pamětí

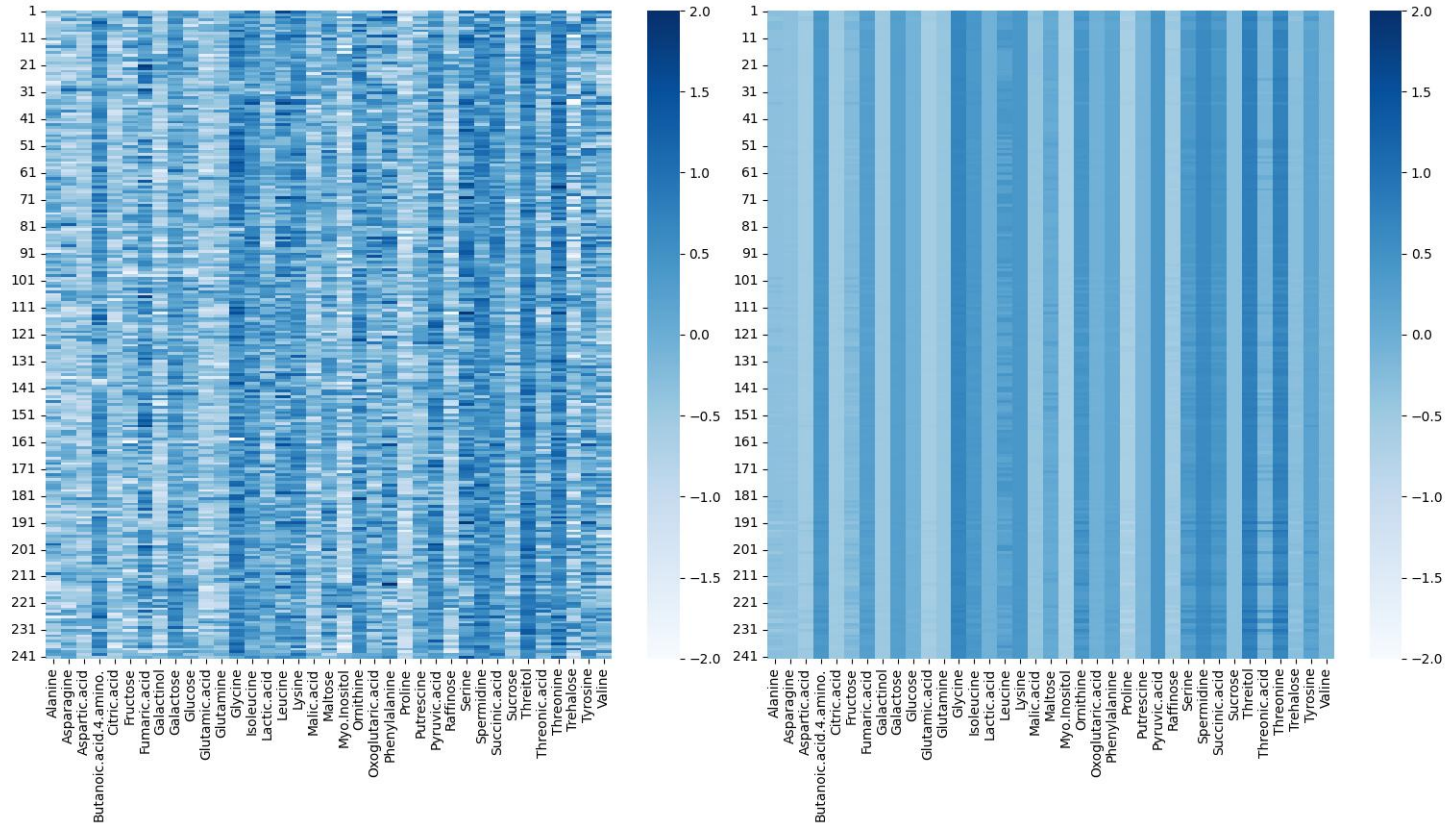
LSTM 6C:	**PK:train [%]**	**PK:test [%]**	**MSE:train**	**MSE:test**	**LSTM 16C:**	**PK:train [%]**	**PK:test [%]**	**MSE:train**	**MSE:test**
Alanin	-2.629	-26.624	0.237	0.218	Alanin	14.687	-6.380	0.269	0.338
Asparagin	8.662	-8.085	0.203	0.190	Asparagin	0.183	-1.071	0.256	0.240
Kyselina asparagová	2.295	11.776	0.138	0.166	Kyselina asparagová	19.292	28.709	0.155	0.127
Kyselina 4-aminobutanová	10.510	-5.310	0.151	0.156	Kyselina 4-aminobutanová	7.284	-1.247	0.187	0.233
Kyselina citronová	8.035	7.196	0.215	0.247	Kyselina citronová	3.052	10.883	0.156	0.187
Fruktóza	47.669	33.056	0.488	0.483	Fruktóza	0.693	19.015	0.296	0.300
Kyselina fumarová	21.019	5.779	0.316	0.391	Kyselina fumarová	-1.284	6.702	0.351	0.330
Galaktinol	17.556	1.613	0.301	0.214	Galaktinol	0.957	0.725	0.254	0.311
Galaktóza	8.110	-6.633	0.221	0.265	Galaktóza	-1.330	-13.676	0.149	0.158
Glukóza	26.356	-1.790	0.693	0.569	Glukóza	10.063	17.831	0.226	0.203
Kyselina glutamová	13.126	-9.287	0.103	0.106	Kyselina glutamová	12.842	28.466	0.174	0.182
Glutamin	18.043	-27.125	0.162	0.199	Glutamin	4.344	-0.938	0.167	0.123
Glycin	3.984	5.590	0.205	0.157	Glycin	-8.338	-11.817	0.236	0.177
Isoleucin	17.173	-0.292	0.145	0.181	Isoleucin	14.297	6.611	0.150	0.175
Kyselina mléčná	45.670	54.751	0.267	0.245	Kyselina mléčná	16.618	28.028	0.199	0.200
Leucin	9.844	-11.122	0.257	0.269	Leucin	25.795	40.852	0.242	0.264
Lysin	20.954	-12.700	0.178	0.217	Lysin	4.294	19.088	0.181	0.159
Kyselina jablečná	21.680	4.303	0.240	0.284	Kyselina jablečná	9.561	-10.649	0.212	0.204
Maltóza	22.565	-2.952	0.256	0.231	Maltóza	11.888	-7.653	0.208	0.169
Myo-inositol	8.017	29.265	0.170	0.181	Myo-inositol	5.689	23.095	0.322	0.364
Ornithin	21.420	11.198	0.168	0.132	Ornithin	18.423	-3.771	0.234	0.151
Kyselina oxoglutarová	-3.078	0.370	0.237	0.246	Kyselina oxoglutarová	5.630	-23.389	0.243	0.277
Fenylalanin	6.676	-8.193	0.324	0.419	Fenylalanin	11.828	5.551	0.293	0.441
Prolin	6.797	-5.843	0.148	0.133	Prolin	14.771	-4.214	0.187	0.254
Putrescin	10.646	-5.659	0.199	0.231	Putrescin	4.494	6.539	0.170	0.147
Kyselina pyrohroznová	16.342	12.431	0.157	0.233	Kyselina pyrohroznová	-6.051	-15.379	0.171	0.240
Raffinóza	12.865	-3.657	0.248	0.179	Raffinóza	11.416	4.955	0.206	0.297
Serin	6.842	8.869	0.252	0.204	Serin	15.814	30.450	0.311	0.345
Spermidin	21.523	11.847	0.231	0.175	Spermidin	0.228	-11.172	0.163	0.129
Kyselina jantarová	18.644	7.445	0.098	0.090	Kyselina jantarová	15.979	5.822	0.123	0.161
Sacharóza	6.509	23.770	0.164	0.172	Sacharóza	-1.847	-0.717	0.159	0.162
Threitol	42.215	35.582	0.108	0.109	Threitol	8.223	-4.421	0.110	0.108
Kyselina threoniová	24.063	-1.401	0.148	0.141	Kyselina threoniová	24.793	18.093	0.216	0.253
Threonin	10.427	3.000	0.126	0.079	Threonin	0.841	-2.780	0.154	0.118
Trehalóza	9.598	-12.123	0.292	0.258	Trehalóza	11.496	-5.958	0.440	0.536
Tyrosin	26.106	-1.043	0.190	0.161	Tyrosin	12.640	7.204	0.262	0.264
Valin	22.717	19.648	0.266	0.404	Valin	12.664	6.565	0.280	0.350

Tab. D.1: Tabulka výsledků hodnotících metrik pro sítě LSTM a metabolity při 6 a 16 °C.

E Heat mapy reálných hodnot a predikovaných hodnot pomocí sítí LSTM

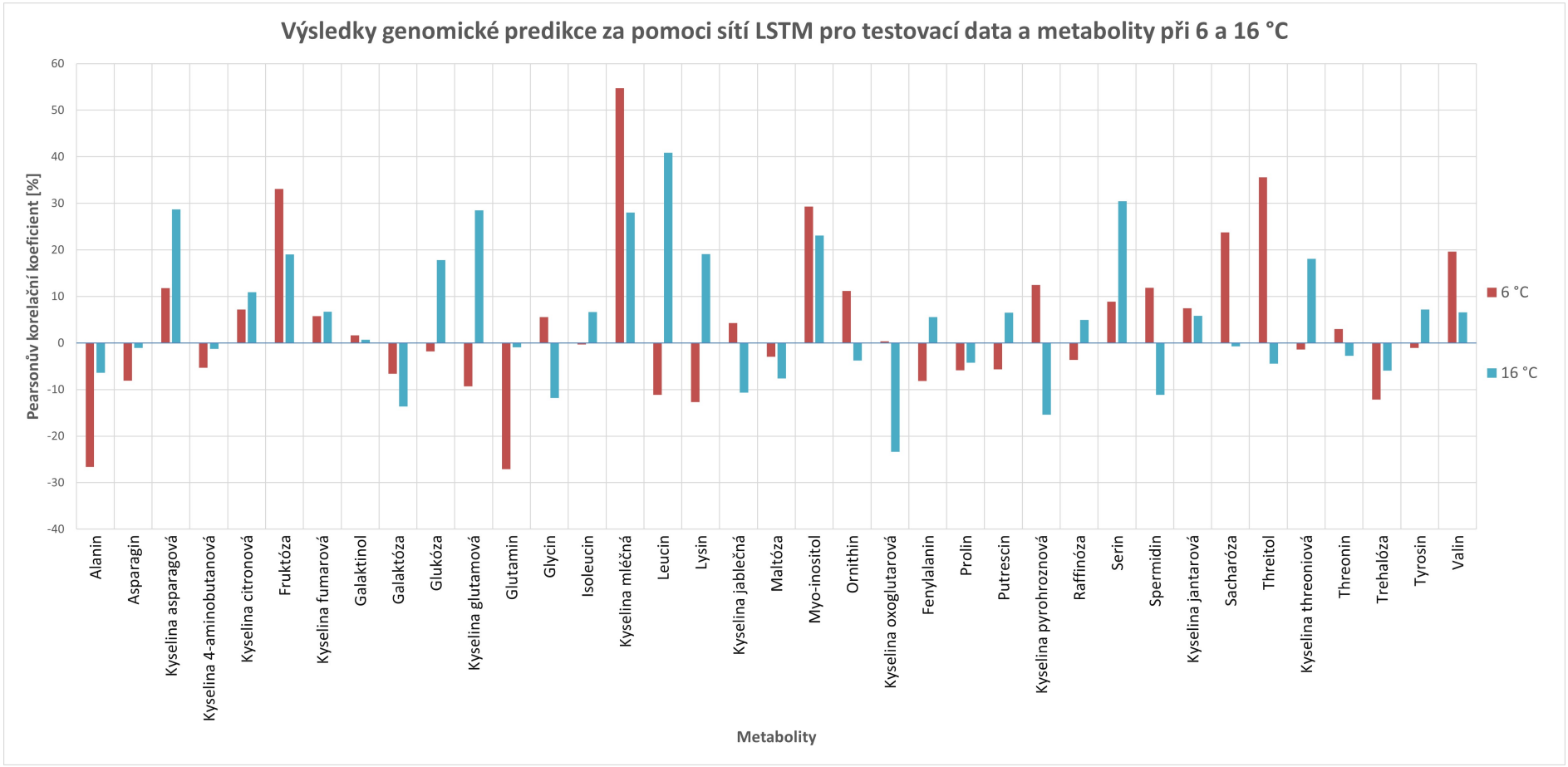


Obr. E.1: Vlevo heat mapa reálných hodnot hladin metabolitů při 6 °C. Vpravo heat mapa predikovaných hodnot hladin metabolitů při 6 °C pro predikci za pomoci sítí LSTM.



Obr. E.2: Vlevo heat mapa reálných hodnot hladin metabolitů při 16 °C. Vpravo heat mapa predikovaných hodnot hladin metabolitů při 16 °C pro predikci za pomoci sítí LSTM.

F Graf výsledků Pearsonova korelačního koeficientu pro predikci za pomoci sítí LSTM a metabolity při 6 a 16 °C



Obr. F.1: Graf znázorňující výsledky Pearsonova korelačního koeficientu pro predikci za pomoci sítí LSTM a metabolity při 6 a 16 °C.

G Obsah elektronické přílohy

Elektronická příloha této práce obsahuje vytvořené kódy, v kterých je aplikována genomická predikce za pomoci ML modelů, a excel soubory obsahující výsledky genomické predikce pro dané modely. Všechny kódy jsou vytvořeny v programovacím jazyce *Python* verze 3.12.2.

```
/ ..... kořenový adresář
├── Kódy ..... kódy pro jednotlivé metody a předzpracování dat
│   ├── Predzpracovani.py
│   ├── Regularizovane_regresni_modely.py
│   ├── Nahodny_les.py
│   └── LSTM.py
└── Výsledky ..... excel soubory s výsledky
    ├── Vysledky_konvencni_metody.xlsx
    ├── LSTM_optimalizace_vysledky.xlsx
    └── Vysledky_LSTM.xlsx
```