DELTA – Střední škola informatiky a ekonomie, s.r.o.

Ke Kamenci 151, Pardubice

Identifikace a vizualizace strukturních motivů chemických látek klíčových pro rozhodování QSAR a QSPR modelů

Autoři: Michael Kolakovský, Petr Švestka

Třída: 4.B

Studijní obor: Informační technologie (18-20-m/01)

Školní rok: 2025/26

Zadání maturitního projektu z informatických předmětů

Jméno a příjmení	Michael Kolakovský
Pro školní rok	2025/2026
Třída	4.B
Obor	Informační technologie 18-20-M/01
Téma práce	Identifikace strukturních motivů chemických látek klíčových pro rozhodování u regresních QSPR modelů
Vedoucí práce	Ing. Ivan Čmelo, Ph.D.

Způsob zpracování, cíle práce, pokyny k obsahu a rozsahu práce

Cíl projektu

Cílem tohoto projektu je navrhnout a implementovat workflow pro identifikaci a vizualizaci strukturních motivů chemických látek, které hrají klíčovou roli v rozhodování u regresních Quantitative Structure-Property Relationship (QSPR) modelů. Projekt se zaměří na zvýšení vysvětlitelnosti tzv. „black-box“ modelů pomocí metodiky SHAP, a tím přispěje k lepšímu porozumění tomu, jak modely odvozují své predikce na základě molekulární struktury.

Specifikace projektu

1. **Teoretická příprava**
    - Bude zpracována krátká literární rešerše pro úvod do problematiky QSPR modelů
    - Bude vytvořen stručný úvod k základním metodám strojového učení a k SHAP metodice vč. Boruta-SHAP
2. **Příprava datových množin**

- Bude provedena analýza dostupných datových zdrojů a databází fyzikálních dat k látkám (např.: rozpustnost ve vodě)
- Budou sestaveny vhodné datové sady chemických látek pro regresní úlohy (např. predikce rozpustnosti)

3. **Návrh řešení**

   - Budou definovány strukturní vlastnosti vhodné pro vstup do QSPR modelů a převedeny na binární vektory (např. strukturní fingerprinty pomocí RDKit).
   - Bude navržena předběžná architektura workflow sestávající ze skriptů pro automatizované zpracování dat, natrénování modelů a následné SHAP, popř. Boruta-SHAP analýzy.

4. **Vývoj nástrojů a workflow**

   - Bude připraveno vývojové prostředí sjednocující potřebné nástroje a knihovny pro cheminformatickou analýzu, strojové učení a vizualizaci
   - Bude implementován skriptovaný proces pro přípravu dat, trénování a testování QSPR modelů (např. Random Forest, XGBoost, neuronové sítě).
   - SHAP analýza bude začleněna do trénovacího procesu pro určení významnosti jednotlivých vstupních rysů.

5. **Agregace dat**

   - Budou vyvinuty skripty pro automatické porovnání výkonu jednotlivých modelů
   - Budou vyvinuty skripty pro automatické sjednocení a porovnání SHAP, popř. Boruta-SHAP hodnot mezi různými modely na identických datech v rámci jedné Pandas dataframe

6. **Vizualizace a interpretace**

   - Bude vytvořena metoda pro automatickou vizualizaci významných strukturních motivů přímo na chemických strukturách (např. zvýraznění fragmentů molekul)
   - Výsledky budou zpracovány tak, aby byly srozumitelné a užitečné pro chemiky bez hlubšího IT zázemí..

7. **Dobře navržený workflow**
    - Workflow bude navržen modulárně s funkčně oddělenými vrstvami
    - Skripty ve workflow budou individuálně zdokumentovány

Požadované výstupy

- Celkově funkční workflow pro analýzu strukturních motivů v chemických datech pomocí SHAP.
- Sada skriptů pro přípravu dat, trénování modelů, interpretaci výstupů a vizualizaci.
- Uživatelská a vývojářská dokumentace.
- Otevřeně dostupný repozitář s kódem a ukázkovými daty.

Hodnocení

Projekt bude hodnocen na základě následujících kritérií:

- Funkčnost a celková použitelnost workflow jako celku i jeho individuálních skriptů
- Srozumitelnost a použitelnost vizualizace identifikovaných klíčových strukturních motivů
- Flexibilita workflow, jeho znovupoužitelnost při změnách v datech nebo modelech
- Úroveň dokumentace, čitelnosti kódu a strukturování skriptů
- Celková kreativita a inovativnost samotného workflow a výstupních vizualizací

Stručný časový harmonogram (s daty a konkretizovanými úkoly)

- **Září-říjen:** Analýza problému, rešeršní činnost, příprava dat
- **Listopad-prosinec:** Příprava prostředí, úvodní implementace workflow
- **Leden-únor:** Iterativní vylepšování workflow, ladění, rozšiřování funkcionality
- **Březen:** Dokumentace projektu, doladění vizualizací, jejich interpretace

Zadání maturitního projektu z informatických předmětů

Jméno a příjmení	Petr Švestka
Pro školní rok	2025/2026
Třída	4.B
Obor	Informační technologie 18-20-M/01
Téma práce	Identifikace strukturních motivů chemických látek klíčových pro rozhodování u klasifikačních QSAR modelů
Vedoucí práce	Ing. Ivan Čmelo, Ph.D.

Způsob zpracování, cíle práce, pokyny k obsahu a rozsahu práce

Cíl projektu

Cílem tohoto projektu je navrhnout a implementovat workflow pro identifikaci a vizualizaci strukturních motivů chemických látek, které hrají klíčovou roli v rozhodování u klasifikačních Quantitative Structure-Activity Relationship (QSAR) modelů. Projekt se zaměří na zvýšení vysvětlitelnosti tzv. „black-box“ modelů pomocí metodiky SHAP, a tím přispěje k lepšímu porozumění tomu, jak modely odvozují své predikce na základě molekulární struktury.

Specifikace projektu

1. **Teoretická příprava**
    - Bude zpracována krátká literární rešerše pro úvod do problematiky QSAR modelů
    - Bude vytvořen stručný úvod k základním metodám strojového učení a k SHAP metodice vč. Boruta-SHAP
2. **Příprava datových množin**

- Bude provedena analýza dostupných datových zdrojů a databází biologických aktivit látek (např.: ChEMBL)
- Budou sestaveny vhodné datové sady chemických látek pro klasifikační úlohy (např. inhibice vybraných receptorů jako glukokortikoidní, androgenní, opioidní, ...)

3. **Návrh řešení**

   - Budou definovány strukturní vlastnosti vhodné pro vstup do QSAR modelů a převedeny na binární vektory (např. strukturní fingerprinty pomocí RDKit).
   - Bude navržena předběžná architektura workflow sestávající ze skriptů pro automatizované zpracování dat, natrénování modelů a následné SHAP, popř. Boruta-SHAP analýzy.

4. **Vývoj nástrojů a workflow**

   - Bude připraveno vývojové prostředí sjednocující potřebné nástroje a knihovny pro cheminformatickou analýzu, strojové učení a vizualizaci
   - Bude implementován skriptovaný proces pro přípravu dat, trénování a testování QSAR modelů (např. Random Forest, XGBoost, neuronové sítě).
   - SHAP analýza bude začleněna do trénovacího procesu pro určení významnosti jednotlivých vstupních rysů.

5. **Agregace dat**

   - Budou vyvinuty skripty pro automatické porovnání výkonu jednotlivých modelů
   - Budou vyvinuty skripty pro automatické sjednocení a porovnání SHAP, popř. Boruta-SHAP hodnot mezi různými modely na identických datech v rámci jedné Pandas dataframe

6. **Vizualizace a interpretace**

   - Bude vytvořena metoda pro automatickou vizualizaci významných strukturních motivů přímo na chemických strukturách (např. zvýraznění fragmentů molekul)

- Výsledky budou zpracovány tak, aby byly srozumitelné a užitečné pro chemiky bez hlubšího IT zázemí..

7. **Dobře navržený workflow**
    - Workflow bude navržen modulárně s funkčně oddělenými vrstvami
    - Skripty ve workflow budou individuálně zdokumentovány

Požadované výstupy

- Celkově funkční workflow pro analýzu strukturních motivů v chemických datech pomocí SHAP.
- Sada skriptů pro přípravu dat, trénování modelů, interpretaci výstupů a vizualizaci.
- Uživatelská a vývojářská dokumentace.
- Otevřeně dostupný repozitář s kódem a ukázkovými daty.

Hodnocení

Projekt bude hodnocen na základě následujících kritérií:

- Funkčnost a celková použitelnost workflow jako celku i jeho individuálních skriptů
- Srozumitelnost a použitelnost vizualizace identifikovaných klíčových strukturních motivů
- Flexibilita workflow, jeho znovupoužitelnost při změnách v datech nebo modelech
- Úroveň dokumentace, čitelnosti kódu a strukturování skriptů
- Celková kreativita a inovativnost samotného workflow a výstupních vizualizací

Stručný časový harmonogram (s daty a konkretizovanými úkoly)

- **Září-říjen:** Analýza problému, rešeršní činnost, příprava dat
- **Listopad-prosinec:** Příprava prostředí, úvodní implementace workflow
- **Leden-únor:** Iterativní vylepšování workflow, ladění, rozšiřování funkcionality

- **Březen:** Dokumentace projektu, doladění vizualizací, jejich interpretace

Prohlašujeme, že jsme maturitní projekt vypracovali samostatně, výhradně s použitím uvedené literatury. Dále prohlašujeme, že při tvorbě této práce jsme použili nástroj generativního modelu AI [ChatGPT 5.4; https://chatgpt.com] za účelem pomoci se strukturováním textu. Po použití tohoto nástroje jsme provedli kontrolu obsahu a přebíráme za něj plnou zodpovědnost.

V Pardubicích 31. 3. 2026 .....

Michael Kolakovský

.....

Petr Švestka

Rozdělení práce

Práce na maturitním projektu byla mezi členy týmu rozdělena následovně. Michael Kolakovský se věnoval implementaci části zaměřené na QSPR. V textové části práce zpracoval úvod, praktickou část, diskuzi a závěr. Dále vytvořil skript, který propojuje části QSPR a QSAR do jednotné CLI utility využívající workflow navržené oběma autory. Současně vytvořil a průběžně spravoval celou dokumentaci v systému LaTeX.

Petr Švestka se věnoval implementaci části zaměřené na QSAR. V dokumentaci zpracoval teoretickou část a rozšířil praktickou část o grafy vycházející z QSAR části projektu. Současně se podílel na kontrole a korektuře zbytku práce.

Poděkování

Rádi bychom poděkovali Ing. Ivanu Čmelovi, Ph.D. z Vysoké školy chemicko-technologické v Praze za návrh tohoto projektu, jeho odborné vedení, cenné připomínky a čas, který nám během práce věnoval.

Dále bychom rádi poděkovali RNDr. Janu Koupilovi, Ph.D. z DELTy – Střední školy informatiky a ekonomie v Pardubicích za pomoc s přípravou na soutěže a s dokumentací.

Anotace

Tato práce se zabývá identifikací a vizualizací strukturních motivů chemických látek, které mají klíčový vliv na rozhodování modelů QSAR a QSPR. Cílem je vytvořit workflow, které spojuje metody cheminformatiky a strojového učení s nástroji vysvětlitelnosti tak, aby bylo možné nejen predikovat vybrané vlastnosti chemických látek, ale také porozumět tomu, na základě kterých strukturních motivů se modely rozhodují.

Klíčová slova: QSAR, QSPR, strojové učení, Random Forest, XGBoost, SMILES, ECFP, SHAP, Boruta-SHAP, vizualizace molekul, cheminformatika, vysvětlitelnost

Abstract

This work focuses on the application of QSAR and QSPR methods to the analysis of chemical compounds, with an emphasis on interpreting the decision-making of machine learning models. The aim of the project is to combine cheminformatics, predictive modeling, and explainable artificial intelligence tools in order not only to predict molecular properties, but also to determine which structural motifs have the greatest influence on the model's prediction.

Keywords: QSAR, QSPR, machine learning, Random Forest, XGBoost, SMILES, ECFP, SHAP, Boruta-SHAP, molecular visualization, cheminformatics, interpretability

Obsah

1 Úvod

Odhad vlastností chemických látek výpočetními „in silico“ metodami, tj. bez nutnosti časově a finančně náročného fyzického „in vitro“ experimentu je velmi důležitým nástrojem pro farmakologii, strukturní biologii, materiálové vědy a mnoho dalších s chemií souvisejících oborů. Prostřednictvím výpočetních modelů se rutinně předzpracovávají mnohamilionové datové množiny, pro vytipování často jen desítek či stovek chemických látek k fyzickému testování. Ačkoliv výpočetní metody mohou jen aproximovat výsledky reálných experimentů, jsou schopny operovat na mnohořádově jiných početních škálách, a jsou tak dnes již nedílnou součástí vývojových procesů chemických látek zejména u farmaceutických firem.

Výpočetní odhad vlastností chemických struktur tudíž výrazně zrychluje a zlevňuje vývoj nových chemických látek, protože umožňuje experimentátorům vytipovat ty nejnadějnější látky a zaměřit se výhradně na ně.

1.1 Teorie SAR

Takzvaná „Structure-activity relationship“ (SAR) teorie je základním kamenem výpočetního odhadu vlastností chemických látek. SAR postuluje, že látky s podobnou strukturou obecně mívají i podobné fyzikálně chemické vlastnosti. Toto lze i v praxi vidět u přirozených skupin látek jako například steroidy, které se fyzikálně chovají velmi podobně, mají velmi podobnou čtyřkruhovou strukturu a působí na velmi podobnou skupinu receptorů, u některých dokonce i s funkčním překryvem.

Výpočetní modely pro odhad vlastností chemických látek využívají SAR tak, že jsou trénovány na strukturách chemických látek u nichž je daná zájmová vlastnost známa. U takto známých látek se snaží různými způsoby reprezentovat jejich strukturní motivy a najít zejména ty kvantitativně asociované (ať už v pozitivním či negativním smyslu) se zájmovou vlastností - tento koncept je zvaný přímo „Quantitative structure-activity relationship“ (QSAR). Pokud model odhaduje místo přímo biologické aktivity nějakou jinou fyzikálně-chemickou vlastnost, často je označován obecněji jako „Quantitative structure-property relationship“ (QSPR). Tato práce se zabývá oběma variantami.

Po natrénování se modely opírají právě o takovéto naučené strukturní motivy ve snaze u neznámých látek odhadnout přítomnost či absenci zájmové vlastnosti, případně i její

míru. Zde je pro chemiky i výpočetní analytiky velmi užitečné vědět jaké přesně strukturní motivy asociuje natrénovaný model se zájmovou vlastností či její absencí.

SAR je totiž velmi užitečnou abstrakcí, nicméně jde stále toliko o abstrakci. Je velmi dobře zdokumentován tzv. SAR paradox, kdy velmi malá strukturní změna u chemické látky má za následek velký výkyv v reálné aktivitě. S tímto pojmem se spojují koncepty jako „Activity Cliff“, tedy pomyslný prudký propad aktivity i při malé změně, a jeho varianta „Magical Methyl“, kde pouhým přidáním či odebráním malé methylové skupiny ve struktuře molekuly mnohořádově skáče její biologická aktivita. Pokud je takovýto signál v trénovacích datech dostatečně reprezentovaný, modely se jej samozřejmě naučí a budou jej používat, a tento zlomový strukturní motiv a jeho umístění nevyhnutelně dostane interně vysokou míru důležitosti. Vhodnou interpretací takového modelu se k takovéto informaci může dostat i chemik.

Některé modely jsou snadno interpretovatelné, zejména ty jednodušší. Například různé formy regrese lze interpretovat přímo na základě příslušných natrénovaných koeficientů. U většiny široce používaných modelů to tak snadné není, zejména v chemické informatice stále velmi často používané modely typu Random Forest (RF) a XGBoost nelze přímo interpretovat.

1.2 Predikce v praxi

Trénování, testování a použití QSAR A QSPR modelů probíhá obdobně jako mnoho dalších použití strojového učení. Typicky se nejprve vytvoří datová sada látek, u nichž jsou známé jak strukturní informace, tak i sledované vlastnosti. Na základě těchto dat je následně natrénován model, který se snaží nalézt vztah mezi strukturou molekuly (propř. nějakou její reprezentací) a její zájmovou vlastností. Takto vytvořený model lze poté využít k predikci vlastností nových, dosud nezkoumaných látek.

V oblasti farmacie se tyto metody používají například při návrhu nových léčiv, kde mohou pomoci odhadnout jak již bylo zmíněno žádoucí biologickou aktivitu, ale i další důležité vlastnosti: v první řadě rozpustnost látky, aby ji bylo možno vůbec otestovat, a aby ji v případě úspěchu bylo možno využít jako léčivo. S tímto se pojí i odhad perorální dostupnosti dané látky, tj. zda je možné látku případně podávat jako pilulku, kapsli, sirup či jakoukoliv jinou formulaci podávanou ústy, což je výrazně více žádoucí než třeba injekční podání. Možná forma podání je pak jen jedním z aspektů často výpočetně odhadovaných tzv. ADMET vlastností (Absorption, Distribution, Metabolism, Excretion, Toxicity), tedy

vlastností souvisejících s absorbcí látky do organismu, její šíření v rámci organismu, její metabolismus v rámci organizmu a mechanismus jejího vyloučení - a specificky i její toxicita, tj. nežádoucí forma biologické aktivity. Tyto procesy a s nimi spjaté modely umožňují prioritizovat látky, které mají největší potenciál pro další výzkum.

V tomto projektu však nehraje důležitou roli samotná predikce, ale jak již bylo zmíněno zejména její interpretace. Podstatou práce je vytvořit proces jak porozumět tomu, na základě jakých strukturních motivů se model rozhoduje. S tím souvisí i vizualizace výsledků, která umožňuje intuitivně zachytit vztahy mezi strukturními motivy molekuly a výslednou predikcí. Vizualizační metody tak usnadňují uživatelské porozumění chování modelu tím, že přímo zobrazují klíčové strukturní motivy v kontextu konkrétních chemických struktur.

1.3 Cíle práce

Cílem práce je natrénovat modely QSAR a QSPR schopné predikovat vybrané vlastnosti chemických látek na základě jejich struktury. Dalším cílem je analyzovat, jak jednotlivé strukturní motivy přispívají k výsledné predikci, a identifikovat ty, které mají největší vliv na sledované vlastnosti. Součástí práce proto bude také implementace a využití vhodných vizualizačních metod, které umožní tyto vztahy přehledně zobrazit.

2 Teoretická část

2.1 Reprezentace molekul

2.1.1 Řetězce SMILES

Abychom mohli pracovat s chemickou látkou v počítači, musíme ji nejprve převést na tvar, kterému bude rozumět. Například s obrázkem molekuly by si model poradit nedokázal.

SMILES (Simplified Molecular-Input Line-Entry System) je textový formát, kterým lze velmi efektivně chemické látky reprezentovat. Tento formát převádí dvojrozměrné molekuly do řetězce pomocí základních ASCII znaků. Jde v podstatě o tradiční reprezentaci chemické struktury jako obecného grafu, kde SMILES je toliko linearizovaným průchodem tohoto grafu.

Atomy jsou zde reprezentovány svými chemickými značkami a jednoduché vazby mezi nimi jsou implicitní. Jiné vazby, ale značíme explicitně. Například dvojnou vazbu značíme pomocí =, trojná je # . Pokud je v látce nějaký cyklus, tak ten značíme pomocí toho, že tuto část ohraničíme čísly. Větvící se struktury se anotují závorkami. Například ethanol lze zapsat jako **CCO** , benzen jako cyklickou strukturu **c1ccccc1** a rozvětvenější kyselinu benzoovou jako **O=C(O)c1ccccc1**.

Výhody SMILES jsou například v jeho kompaktnosti, kde i velmi složité molekuly lze reprezentovat pomocí pár desítek bajtů. Tím je pak i zpracování datasetu z csv formátu velmi rychlé a nenáročné na výpočetní výkon.

Výhodou také je, že formát SMILES je dnes běžným standardem a většina velkých chemických databází (včetně použitého datasetu AqSolDb) poskytuje vzorce sloučenin právě v tomto formátu.

Poslední z výhod je že knihovny jako RDKit jsou primárně optimalizované pro práci právě se SMILES, které umí převést do molekulových grafů a dále s nimi pracovat. [1]

2.1.2 Molekulový graf

V teorii grafů lze tento graf popsat jako neorientovaný graf, kde vrcholy jsou atomy a hrany jsou vazby mezi nimi. Vlastnosti vrcholů odpovídají vlastnostem atomů jako jsou atomové číslo, formální náboj nebo třeba počet navázaných vodíků. Hrany mají též vlastnosti a ty nesou informace hlavně o typu vazby mezi atomy (jednoduchá, dvojná, trojná, aromatická).

2.1.3 Extended Connectivity Fingerprint

Mezi grafovou reprezentací molekul a většinou standardních modelů je ale stále významná metodická mezera. Zatímco molekuly jsou reprezentovány grafy, modely často operují výhradně s vektory binárních, celých nebo racionálních čísel. U vektorů je taktéž často vyžadována fixní délka. Způsobů jak vektorově vyjádřit inherentně grafovou strukturu je vícero, ale zdaleka nejpoužívanějším způsobem jsou tzv. strukturní fingerprinty. Z této skupiny je stále nejpoužívanější skupina tzv. Extended Connectivity Fingerprints (ECFP), které převádí grafovou strukturu molekuly do binárního vektoru volitelné délky.

Samotný algoritmus ECFP funguje tak, že postupně iteruje přes všechny atomy a podle rádiusu, který si můžeme nastavit, se dívá i na jejich okolí a staví si z toho fragmenty. Tedy rádius nám říká kolik atomů od prvotního atomu bereme v potaz pro výpočet vlastností. Každý z těchto fragmentů potom projde skrz matematickou hashovací funkcí, která mu přidělí pozici ve výsledném fingerprintu. Velikost fingerprintu si můžeme libovolně nastavit a čím víc použijeme bitů tím je fingerprint přesnější, jelikož to sníží riziko kolizí (situaci kdy by dva různé fragmenty měly stejnou pozici). Algoritmus poté vezme spočítané pozice fragmentů a tam kde je fragment, je ve výsledném fingerprintu jednička a na ostatních místech zůstává nula.

Tento ECFP fingerprint je navržený tak, že je z většiny tvořen nulami, což dává i logicky smysl, protože jedna látka nikdy neobsahuje všechny strukturní motivy. [2]

2.2 Strojové učení

Strojové učení je podoblast informatiky, která se zabývá tvorbou modelů schopných nalézat souvislosti v datech a využívat je k predikci nebo rozhodování. U strojového učení není přesný způsob rozhodování určen předem, ale model si jej vytváří na základě trénovacích dat. Při učení získává vstupní data spolu se správnými výstupy a na jejich základě se snaží pochopit obecné vztahy, které lze následně využít i pro nová, dříve neznámá data.

2.3 Rozhodovací stromy

Každý rozhodovací strom má kořen, což je hlavní uzel. A následně vnitřní uzly na které se větví. Každý uzel má nějakou určitou otázku a podle toho co platí se rozhodne, na které z uzlů data dál pošle. V kontextu QSAR modelů by otázka například mohla být jestli je

bit 315 jednička nebo nula. Podle její odpovědi se tedy rozhodne na který z dalších uzlů se data dále pošlou a takto se pokračuje až dokud se data nedostanou na uzel, který už nemá žádné poduzly.

Cílem rozhodovacího stromu je, aby podobná data vždy skončila na stejných cílových uzlech, protože poté může těmto cílovým uzlům přiřadit cílovou hodnotu podle výsledků z trénovacích dat. A následně pokud do stromu vložíme data, která nikdy předtím neviděl, tak podle toho kde skončí dokáže předpovědět kterým z dat, které už viděl se nejvíce podobají a dokáže předpovědět jaký by měl být výsledek pro tato data.

Jeden z největších problémů rozhodovacích stromů je overfitting, což je jev při kterém se stromy naučí šum trénovacích dat, ale nejsou schopny se přizpůsobit na nová data, která ještě neznají. V praxi se stane to, že pokud strom roste neomezeně do hloubky tak si vytvoří specifické pravidlo pro každý jednotlivý vzorek v trénovacích datech. Tím pádem se data naučí nazpaměť, ale na nových molekulách poté selže. Právě kvůli tomuto nikdy nepoužíváme samotný rozhodovací strom, ale místo toho se využívají takzvané souborové (ensemble) metody, které spojují sílu mnoha menších, obecnějších stromů dohromady ve snaze zachytit obecné trendy a nesestupovat na úroveň overfitování konkrétních datových bodů.

2.3.1 Učící metody Random Forest a XGBoost

2.3.1.1 Random Forest

Random Forest funguje na principu toho, že natrénuje vysoký počet rozhodovacích stromů. Každý z nich poté natrénuje na náhodném vzorku dat a při větvení vybírá z náhodné podmnožiny vlastností. Následně při samotném rozhodování se data vloží do všech stromů současně a jednotlivé stromy hlasují jaký bude výsledek. Toto fungování pomáhá k předejití již zmiňovaného overfittingu. [3]

2.3.1.2 XGBoost

XGBoost rovněž funguje tak, že trénuje spoustu rozhodovacích stromů, ale samotné vyhodnocení probíhá jinak než u Random Forestu. XGBoost totiž pracuje tak, že vloží data do prvního stromu, ten nám vrátí nějaký výsledek a další strom se následně neučí předpovídat samotný výsledek, ale snaží se předpovědět a opravit chyby, které udělal strom před ním. Tímto způsobem poté projdou data sekvenčně skrz všechny stromy.

Ve věci overfittingu je už XGBoost o něco sofistikovanější a řeší ho především pomocí takzvané regularizace, což je jedna z jeho silných výhod oproti běžným rozhodovacím stromům. Toto funguje tak, že XGBoost dává penaltu za příliš složité stromy. Pokud by se tedy strom chtěl větvit až příliš do hloubky, je algoritmicky upozaděn oproti jiným, obecnějším stromům.

Další z principů, které XGBoost používá je takzvaný Shrinkage. Ten pracuje tak, že se model učí po malých částech a nový strom opraví chyby toho předchozího jen částečně. Toto zabraňuje tomu, aby se model vydal příliš rychle špatným směrem na základě šumu v datech. [4, 5]

2.3.2 Modely QSAR a QSPR

Modely QSAR se zaměřují na predikci bioaktivity chemických látek vůči konkrétnímu biologickému cíli, jako je například interakce s buněčným receptorem. V literatuře se můžeme setkat s tím, že QSAR může být i regresní problém, ale v kontextu práce je tento přístup definován jako binární klasifikace. Model v podstatě odpovídá na otázku zda je, nebo není chemická látka vůči danému biologickému cíli jakkoliv aktivní, jakoukoliv formou aktivace nebo inhibice. Použití regrese by vynucovalo omezení modelů na specifickou formu aktivity vyjádřenou specifickým způsobem (IC50, EC50, Ki, atd.), což by modely nevyhnutelně roztříštilo a vybočilo by příliš daleko ze zaměření této práce.

Hlavní rozdíl mezi QSPR a QSAR spočívá v tom, co přesně se modely snaží předpovídat. Zatímco QSPR zkoumá fyzikálně-chemické vlastnosti látky (jako je rozpustnost logS), QSAR se zaměřuje přímo na to, jak bude látka účinkovat na konkrétním receptoru.

V praxi se poté ale tyto metody velmi často používají dohromady. Jelikož látka může mít správnou reakci s receptorem, ale pokud se nerozpustí ve vodě, pak ji není možné použít na výrobu léčiv, protože by se neměla do těla jak vstřebat.

2.4 Použité technologie

Jako náš hlavní programovací jazyk jsme se rozhodli zvolit Python (ve verzi 3.11). Jedním z hlavních důvodů bylo, že tento jazyk má velice široký ekosystém knihoven pro strojové učení. Rovněž se hodí na rychlé a efektivní prototypování.

Dále jsme využili knihovnu RDKit, kterou používáme především na parsování molekul ze SMILES formátu, jejich validaci a standardizaci a následně k vytváření samotných

ECFP fingerprintů. Je to rovněž nejpoužívanější a nejvíce rozšířená knihovna pro tento účel. Pro efektivní načítání dat a práci s nimi jsme použili knihovny NumPy a Pandas.

Poté pro samotné strojové učení používáme knihovnu Scikit-learn, která poskytuje nástroje pro cross-validaci, měření výkonnosti modelu a jednoduché propojení přípravy dat a tréninku modelu. Dále zajišťuje export a uložení natrénovaných modelů pro pozdější predikce. Následně na vizualizaci výsledků do grafů používáme knihovnu Matplotlib.

Pro dohledatelnost změn je využit verzovací systém Git. Ten používáme v kombinaci s virtuálním prostředím a balíčkovacím systémem pip, který uchovává přesné verze použitých knihoven v souboru requirements.txt, aby bylo zajištěno, že se projekt bude chovat na všech zařízeních stejně.

3 Praktická část

3.1 Tvorba modelů strojového učení

3.1.1 Příprava datasetů, generování ECFP Fingerprintů

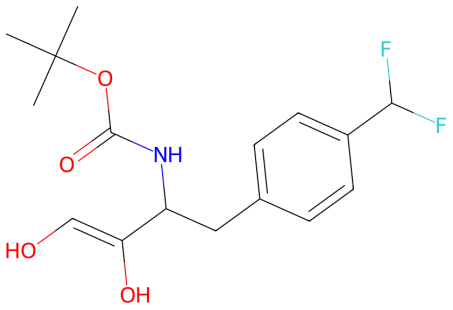
Před samotným trénováním modelů je nutné připravit vhodný dataset. Námi použité datasety jsou pro QSAR ChEMBL a pro QSPR AqSolPred. Oba datasety obsahují chemické látky reprezentované SMILES řetezci a k nim odpovídající hodnotu sledované vlastnosti, u AqSolDB to je rozpustnost ve vodě v jednotce log(S), u ChEMBL to jsou receptory na které daná chemická látka reaguje.

SMILES	Rozpustnost ve vodě, $log(S)$
CCO	1.2336682179
C=C	-0.4
...	
C=CCOCC=C	-0.0206

Tabulka 1: Ukázka části datasetu AqSolDB obsahujícího SMILES reprezentaci molekul a jejich experimentální hodnoty rozpustnosti ve vodě ($\log(S)$). [6]

Tyto vstupní údaje je nejprve potřeba převést do číselné reprezentace, které mohou algoritmy strojového učení rozumět. V této práci je k tomuto účelu použito ECFP, který převádí strukturu molekuly do binárního vektoru reprezentujícího přítomnost různých strukturních motivů.

ECFP může vypadat například takto: `[0,0,0,0,0,0,0,1,0,1,...]`. [2]



Obrázek 1: Vizualizace SMILES řetězce `CC(C(C)OC(=O)NC(Cc1ccc(cc1)C(F)F)C(O)=CO`

Dalším krokem je rozdělení datasetu na trénovací a testovací část. Trénovací data slouží k samotnému učení modelu, zatímco testovací data se používají k vyhodnocení jeho schopnosti zobecnění na dosud neviděná data. Toto rozdělení je důležité, protože umožňuje ověřit, zda model pouze nezapamatoval trénovací data, ale skutečně se naučil obecné vztahy mezi strukturou molekul a sledovanou vlastností.

3.1.2 Mapování bitů fingerprintu na strukturní motivy pro pozdější vizualizaci

3.1.2.1 Princip mapování strukturních motivů

Pro následnou interpretaci modelů a vizualizaci důležitých strukturních motivů je potřeba převést jednotlivé bity ECFP zpět na chemicky srozumitelnou reprezentaci. Přestože ECFP fingerprint představuje molekulu ve formě binárního vektoru, jednotlivé aktivované bity odpovídají lokálním atomovým okolím v molekule, která jsou vytvářena podle zvoleného poloměru fingerprintu. U Morganových fingerprintů tedy každý bit reprezentuje určitý strukturní motiv odvozený z okolí vybraného atomu. [7, 8]

Abychom zjistili jaké strukturní motivy se pod daným bitem skrývají, využíváme možnost pro ukládání doplňujících informací. Díky tomu je možné získat metadata popisující, které části molekuly vedly k aktivaci jednotlivých bitů fingerprintu. Konkrétně je pro každý aktivovaný bit zaznamenán centrální atom a poloměr jeho okolí. Na základě těchto informací lze následně rekonstruovat odpovídající podstrukturu molekuly reprezentovanou daným bitem. [9, 8]

3.1.2.2 Technické provedení a zpracování dat

V této práci byl pro generování fingerprintů využit nástroj z knihovny RDKit. Pro všechny molekuly v datasetu byl vypočten fingerprint a z doplňujících metadat byla získána informace o aktivovaných bitech. Poté byla jednotlivá atomové okolí převedena do zápisu SMARTS, který umožňuje chemickou podstrukturu jednoznačně popsat a dále zobrazovat. Pro každý z bitů byla vytvořena množina odpovídajících podstruktur, aby bylo možné zaznamenat všechny různé motivy, které se mohou pod stejným bitem vyskytovat v důsledku hashování fingerprintu. Zároveň byl veden slovník četností, který u každého SMARTS výrazu eviduje počet jeho výskytů v celém datasetu.

Výstupem této části je mapování, které ke každému bitu ukládá množinu odpovídajících SMARTS podstruktur, a rovněž jejich četnost v datasetu. Takto připravená data lze následně využít při interpretaci výsledků metod SHAP a Boruta-SHAP, protože významné bity fingerprintu lze převést na konkrétní chemické motivy a ty dále vizualizovat nebo chemicky interpretovat. [9, 10]

3.1.3 Trénování modelů

Po přípravě datasetu následovalo samotné trénování modelu. V tomto kroku algoritmus analyzuje trénovací data a snaží se nalézt vztah mezi vstupními proměnnými, tedy například hodnotami fingerprintu, a cílovou veličinou, kterou chceme predikovat. Výsledkem tohoto procesu je model, který je schopný odhadnout hodnotu sledované vlastnosti i pro chemické, které nebyly součástí trénovacích dat.

Při trénování modelů je důležité také správné nastavení jejich hyperparametrů, které mohou významně ovlivnit výslednou přesnost predikce. U metod založených na rozhodovacích stromech, jako jsou Random Forest nebo XGBoost, jde například o počet stromů v modelu nebo maximální hloubku jednotlivých stromů. Vhodná volba těchto parametrů může zlepšit schopnost modelu zachytit vztahy v datech a zároveň omezit riziko přeučení.

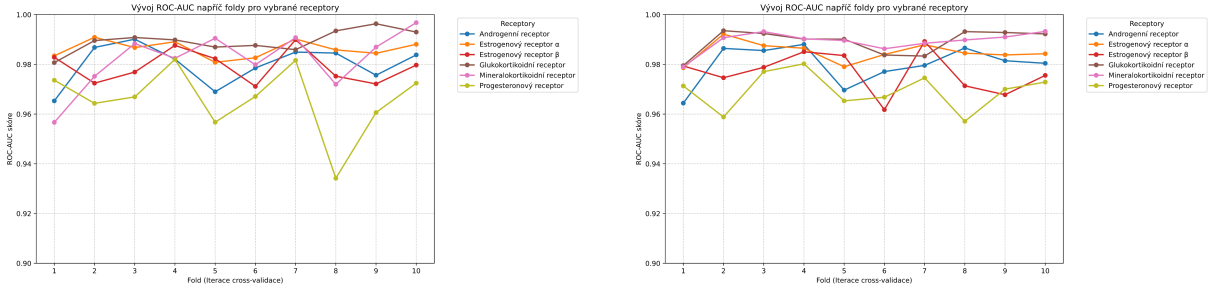
3.1.4 Cross-validace

Po natrénování modelu je nutné ověřit, jak dobře dokáže pracovat s novými daty. Jednoduché rozdělení datasetu na trénovací a testovací část může být někdy nedostatečné, protože výsledky mohou výrazně záviset na konkrétním rozdělení dat. Proto se v praxi často používá metoda zvaná **cross-validation**, která umožňuje spolehlivější vyhodnocení modelu. [11]

Nejčastěji používanou variantou je takzvaná **k-fold cross-validace**. Dataset je při ní rozdělen na k stejně velkých částí (tzv. foldů). Model je následně opakovaně trénován na $k - 1$ částech dat a testován na zbývající části. Tento proces se opakuje k-krát tak, aby každá část datasetu sloužila jednou jako testovací data. Výsledné hodnocení modelu je potom získáno zprůměrováním výsledků ze všech iterací. [12]

Vzhledem k tomu, že QSAR modely v této práci řeší klasifikační úlohy, u kterých bývá chemický dataset často nevyvážený, byla pro ně využita **stratifikovaná k-fold křížová validace** (Stratified k-fold). Tato varianta navíc dbá na to, aby byl v každém

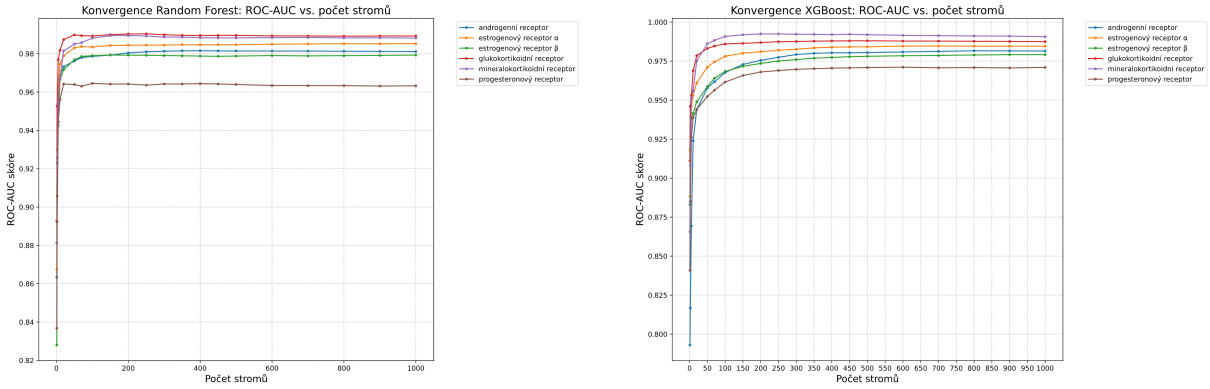
foldu zachován stejný poměr zastoupení jednotlivých tříd jako v původním datasetu, čímž se předchází vzniku foldů s extrémním nedostatkem jedné ze tříd. [13]



Obrázek 2: Cross-validace pro všechny receptory: (vlevo) u Random Forest QSAR modelu a (vpravo) u XGBoost QSAR modelu

3.1.5 Ladění hyperparametrů

V rámci této práce jsme se rovněž snažili nastavit náš model, tak aby byl co výpočetně nejefektivnější. Tohoto jsme docílili tím, že jsme testovali jak se přesnost modelu mění na základě počtu použitých estimátorů (počtu stromů). Při postupném zvyšování, na začátku přesnost modelu velmi rychle roste. Ale poté se dostane na pomyslné maximum, kde už přidávání stromů nemá na přesnost žádný pozitivní účinek. U QSAR jsme tohoto dosáhli, při 100 použitých stromech pro XGBoost a pro Random Forest stačilo pouze 50 stromů. Tímto jsme docílili toho, že modely nejsou zbytečně větší a výpočetně náročnější než by musely být.



Obrázek 3: Testování jak počet stromů ovlivňuje výkon pro všechny receptory: (vlevo) u Random Forest QSAR modelu a (vpravo) u XGBoost QSAR modelu

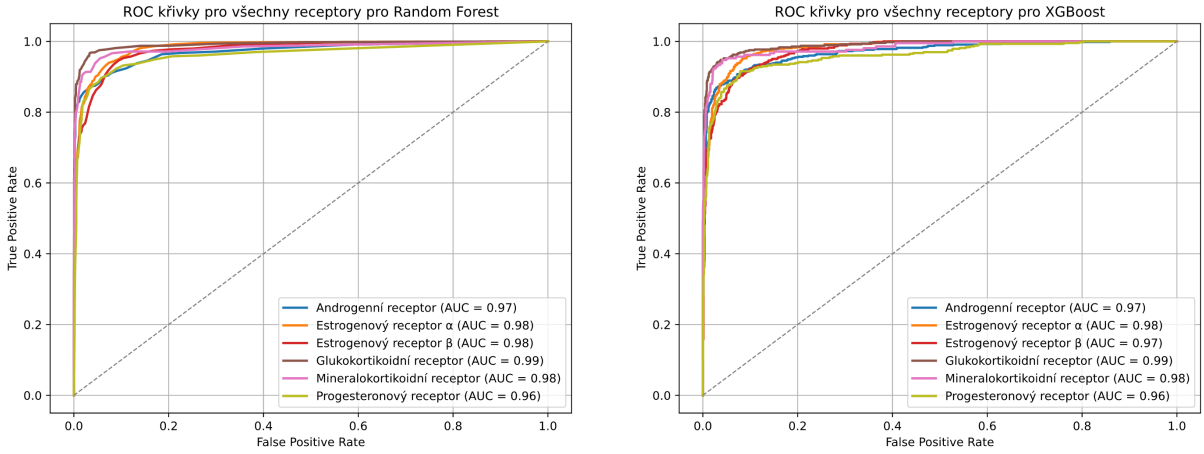
Další parametr, který jsme ladili, je použitý rádius u ECFP fingerprintu.

3.1.6 Vyhodnocení výkonu modelu

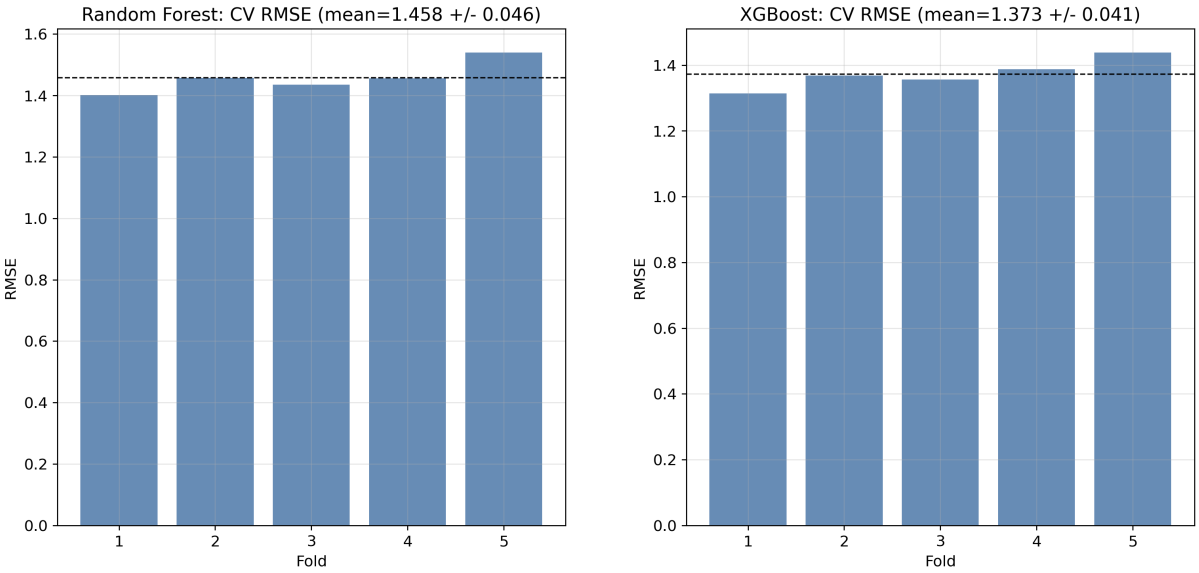
Po natrénování a otestování modelu je potřeba vyhodnotit jeho přesnost pomocí vhodných metrik. Volba konkrétní metriky závisí na typu predikce. V případě QSPR modelů se jedná o regresi, zatímco QSAR modely v naší práci řeší klasifikační problém. [14]

U regresních modelů, jako jsou QSPR modely predikující například rozpustnost, se často používají metriky jako odmocnina střední kvadratické chyby (Root Mean Squared Error, RMSE). RMSE udává průměrnou velikost chyby predikce. Tato metrika umožňuje posoudit, jak přesně model dokáže odhadovat hodnoty sledované vlastnosti. [15]

V případě klasifikačních QSAR modelů se často využívá ROC křivka (Receiver Operating Characteristic). Tato křivka znázorňuje vztah mezi hranicí pravdivé hodnoty modelu (true-positive rate) a mírou nesprávných výsledků (false-positive rate). Často se také používá plocha pod ROC křivkou (AUC), která shrnuje schopnost modelu správně rozlišovat mezi pozitivními a negativními případy. [16]



Obrázek 4: ROC křivky pro všechny receptory: (vlevo) u Random Forest QSAR modelu a (vpravo) u XGBoost QSAR modelu



Obrázek 5: RMSE při 5-fold cross-validaci, (vlevo) u Random Forest QSAR modelu a (vpravo) u XGBoost QSAR modelu

3.2 Identifikace klíčových strukturních motivů

Pro interpretaci rozhodování modelu lze využít metody vysvětlitelnosti, například SHAP nebo Boruta-SHAP. Tyto metody umožňují určit, které vstupní proměnné mají největší vliv na predikci modelu.

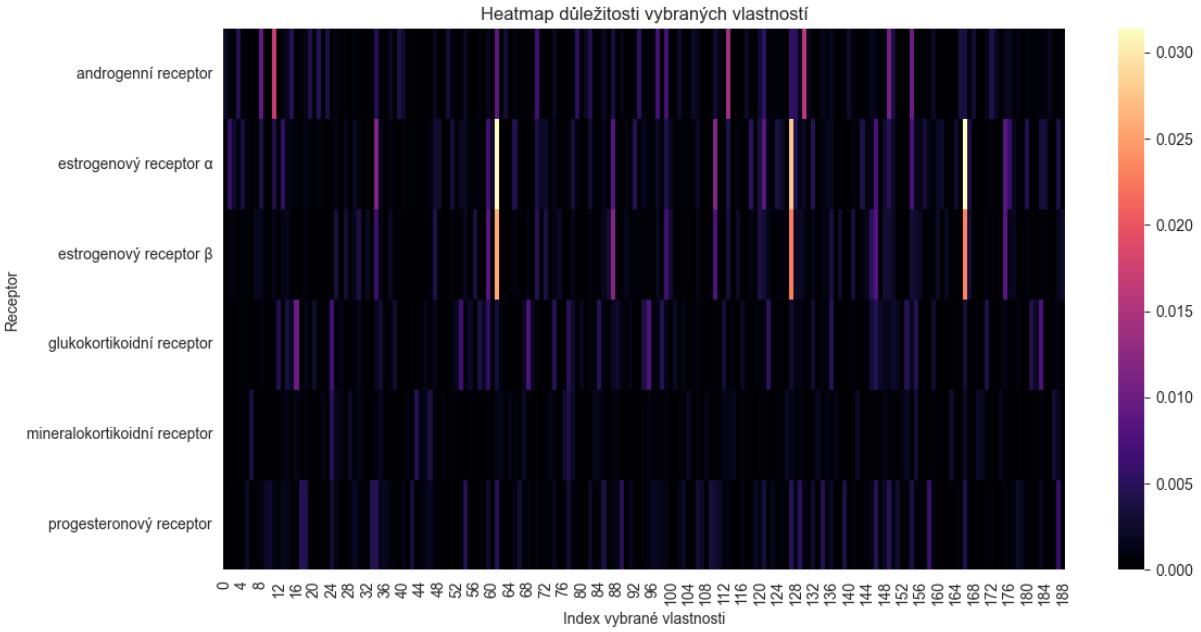
V této práci jednotlivé vstupní proměnné odpovídají bitům ECFP fingerprintu. Každý bit reprezentuje přítomnost určitého strukturního motivu chemické látky, a proto lze pomocí těchto metod identifikovat strukturní motivy, které mají největší vliv na predikci dané vlastnosti chemické látky. [2]

3.2.1 Metody SHAP a Boruta-SHAP

Metoda SHAP (SHapley Additive exPlanations) slouží k interpretaci predikcí modelu tím, že jednotlivým vstupním proměnným přiřazuje hodnotu vyjadřující jejich vliv na výslednou predikci. Umožňuje tak určit, které proměnné byly pro rozhodnutí modelu nejdůležitější, a to jak pro jednotlivé predikce, tak pro model jako celek. [17]

Boruta-SHAP je metoda určená pro výběr důležitých vstupních proměnných. Vychází z porovnání významu skutečných proměnných s náhodně vytvořenými proměnnými a využívá přitom hodnoty SHAP. Tím umožňuje určit, které proměnné mají pro rozhodování modelu skutečný význam a které naopak nepřinášejí podstatnou informaci. [18]

V případě této práce odpovídají jednotlivé proměnné bitům ECFP fingerprintu. Pomocí metod SHAP a Boruta-SHAP je proto možné určit, které bity mají největší vliv na predikci modelu, a následně je přiřadit ke konkrétním strukturním motivům chemických látek.



Obrázek 6: Heatmapa zobrazující důležitost jednotlivých bitů určena metodou SHAP při rozhodování QSAR modelů trénovaných pro jednotlivé receptory lidského těla

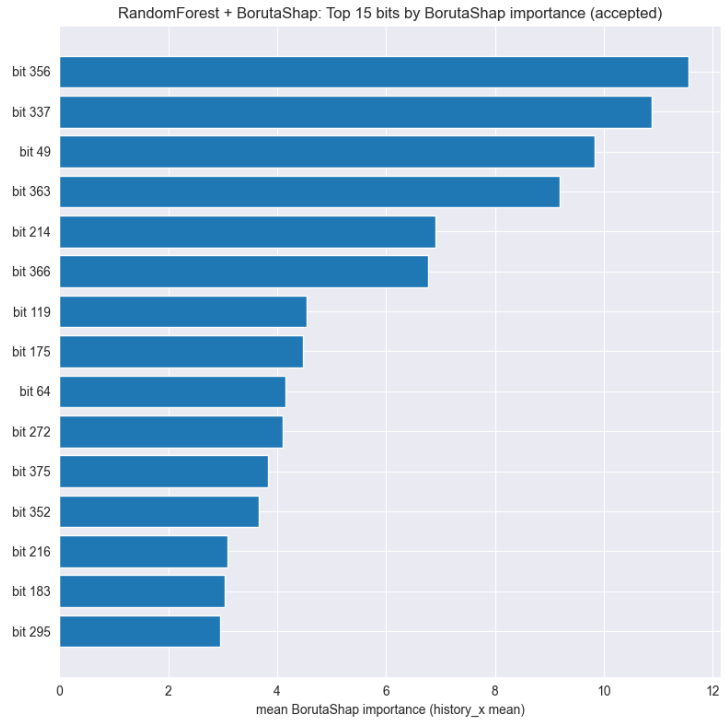
3.2.2 Identifikace významných fingerprint bitů

Na obrázku je znázorněna heatmapa zobrazující důležitost jednotlivých bitů ECFP fingerprintu pro rozhodování QSAR modelů typu Random Forest trénovaných na jednotlivých receptorech lidského těla. Důležitost jednotlivých bitů byla určena pomocí metody SHAP popsané v předchozí části. Na ose x je uveden index bitu fingerprintu a na ose y jednotlivé receptory. Světlejší zabarvení odpovídá vyšší důležitosti daného bitu pro predikci modelu pro konkrétní receptor.

3.3 Vizualizace identifikovaných klíčových strukturních motivů

Na základě důležitosti jednotlivých bitů určené metodou SHAP nebo Boruta-SHAP a s využitím bit-info mapy vytvořené při generování ECFP lze zpětně určit, kterým strukturním motivům tyto bity odpovídají. Tímto způsobem je možné lépe porozumět tomu, které části chemické látky mají největší vliv na rozhodování modelu.

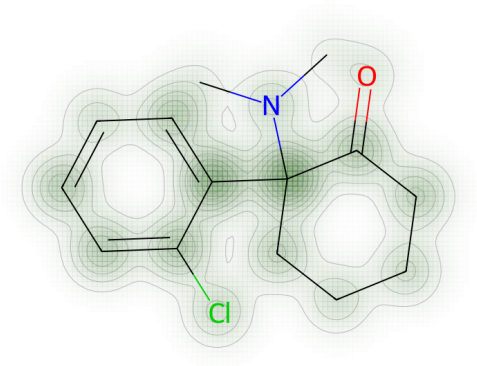
3.3.1 Výběr nejdůležitějších fingerprint bitů



Obrázek 7: Top 15 bitů zvolených jako nejdůležitější metodou Boruta-SHAP

Obrázek 7 zobrazuje nejdůležitější fingerprint bity identifikované metodou BorutaShap při trénování modelu Random Forest. Na vodorovné ose je zobrazena průměrná hodnota důležitosti jednotlivých bitů vypočtená během průběhu algoritmu.

Každý bit fingerprintu odpovídá určitému strukturnímu motivu v molekule. Bity s vyšší důležitostí mají větší vliv na rozhodování modelu při predikci rozpustnosti chemických látek. Tyto bity byly následně přiřazeny zpět ke konkrétním částem chemické struktury pomocí bit-info mapy, vygenerované při tvorbě ECFP, která umožňuje jejich interpretaci prostřednictvím odpovídajících SMARTS vzorů.



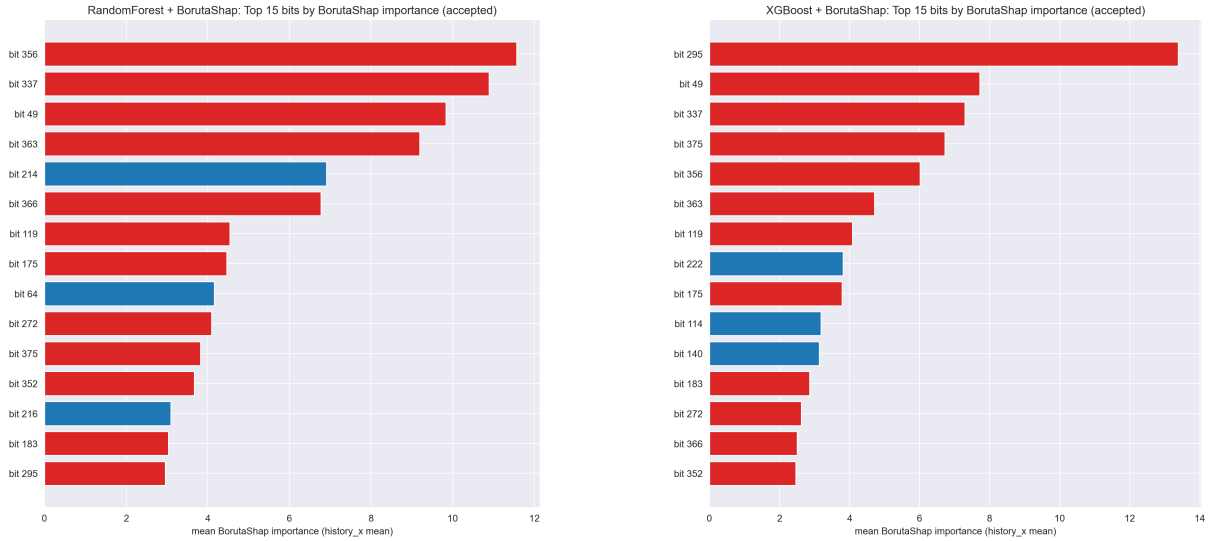
Obrázek 8: Ukázka vizualizace klíčových strukturních motivů

Na základě tohoto přiřazení bylo možné identifikovat strukturní fragmenty, které nejvíce ovlivňují výslednou predikci. Tyto fragmenty byly následně vizualizovány přímo ve struktuře molekuly, což umožňuje lépe pochopit, jak jednotlivé části molekuly přispívají k jejím vlastnostem.

4 Diskuze

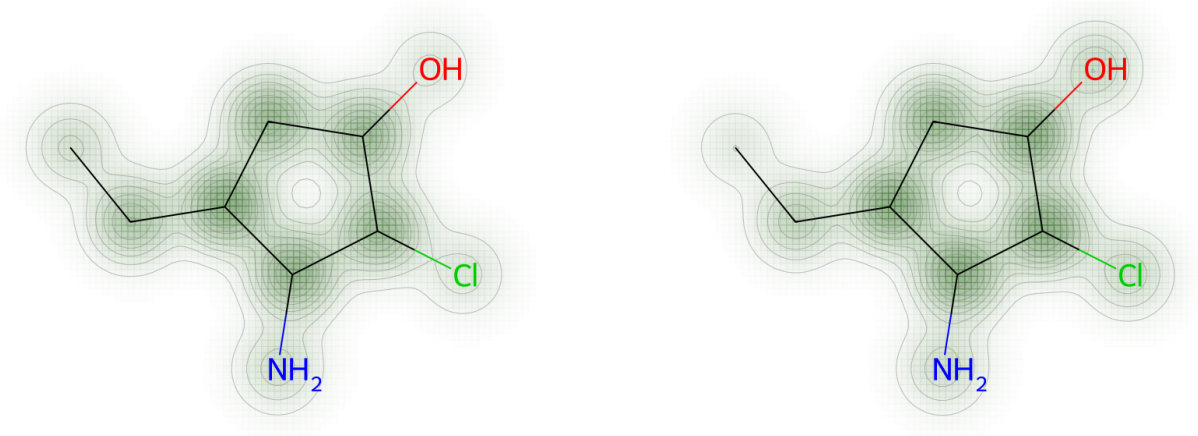
4.1 Porovnání rozhodování metod Random Forest a XGBoost

Při porovnání metod Random Forest a XGBoost se ukázalo, že oba modely se rozhodují velmi podobně. V patnácti nejdůležitějších bitech se lišily pouze ve čtyřech případech, což znamená, že oba přístupy považují za důležité téměř stejné strukturní motivy.



Obrázek 9: Top 15 bitů u QSPR modelů pro metody RandomForest a XGBoost, červeně jsou znázorněny společné bity, modře bity kterými se metody liší.

Tento výsledek je patrný i z grafu, kde je vidět, že velká část nejvýznamnějších bitů je u obou metod stejná. Rozdíly se objevují jen u menší části z nich, takže celkově lze říct, že Random Forest i XGBoost zachycují v datech velmi podobné vztahy.



Obrázek 10: Vizualizace důležitosti strukturních motivů pro rozhodování Random Forest (vlevo) a XGBoost (vpravo) QSPR modelů

Vizualizace ukazuje, že oba modely přisuzují největší význam stejným částem molekuly. Zvýraznění se v obou případech soustředí hlavně v oblasti centrálního uhlíkového kruhu a v okolí navázaných skupin NH_2 a Cl. Modely dosahují značné míry shody a rozdíly mezi metodami jsou jen velmi malé a projevují se především u postranních řetězců molekuly, tj. u hydroxylové skupiny (-OH), kterou XGBoost zohledňuje o něco více. Naopak Random Forest v tomtéž kontextu více přihlíží k ethylové skupině (-CH2-CH3).

Právě tato podobnost v rozhodování může být jedním z důvodů, proč oba modely dosahovaly podobné přesnosti. Tento příklad je samozřejmě jedna struktura, tímto způsobem lze interpretovat rozhodování modelů u jakéhokoliv jiného strukturního grafu.

4.2 Zvážení využití neuronových sítí

V průběhu práce byla zvažována také implementace modelů založených na neuronových sítích. Po provedení experimentů s metodami Random Forest a XGBoost však bylo dosaženo velmi dobrých výsledků. Z tohoto důvodu nebyla implementace neuronových sítí dále rozvíjena, protože by pravděpodobně nepřinesla výrazné zlepšení výsledků v rámci rozsahu této práce.

5 Závěr

Cílem této práce bylo prozkoumat možnosti využití a zejména interpretaci výpočetních metod pro predikci vlastností chemických látek na základě jejich struktury. V rámci práce se podařilo vytvořit a natrénovat modely založené na algoritmech Random Forest a XGBoost, které pro zvolená problém dosahovaly dobré přesnosti.

Důležitou součástí práce byla interpretace výsledků. Podařilo se nám kvantitativně určit klíčové strukturní motivy, které mají významný vliv na modelované vlastnosti chemických látek.

Dalším přínosem práce je využití vizualizačních metod, které umožňují názorně zobrazit vztah mezi strukturou molekuly a výslednou predikcí. Vizualizace tak usnadňuje interpretaci modelů a umožňuje lépe porozumět tomu, na základě jakých strukturních motivů modely svá rozhodnutí vytvářejí.

Pro demonstraci celého procesu predikce a interpretace je v následující části uveden konkrétní příklad chemické sloučeniny, na kterém je ukázán kompletní workflow od vstupní struktury až po výslednou predikci a její vizualizaci.

5.1 Využití navrženého workflow pro predikci vlastností a vizualizaci klíčových strukturních motivů chemické látky

5.1.1 Vstup

Na vstupu workflow je reprezentace molekuly ve formátu SMILES.

CC(=O)C1CCC2C3C(C(C4=CC(=O)C=CC4(C)C3CCC12C)O)O

5.1.2 Predikované hodnoty

Receptor	**Predikovaná pravděpodobnost**
Estrogenový receptor α (ERα)	0.984
Estrogenový receptor β (ERβ)	0.928
Glukokortikoidní receptor (GR)	0.730
Androgenní receptor (AR)	0.652
Progesteronový receptor (PR)	0.004
Mineralokortikoidní receptor (MR)	0.002
Vlastnost	**Hodnota**
Rozpustnost ve vodě	-3.6605 log(S)

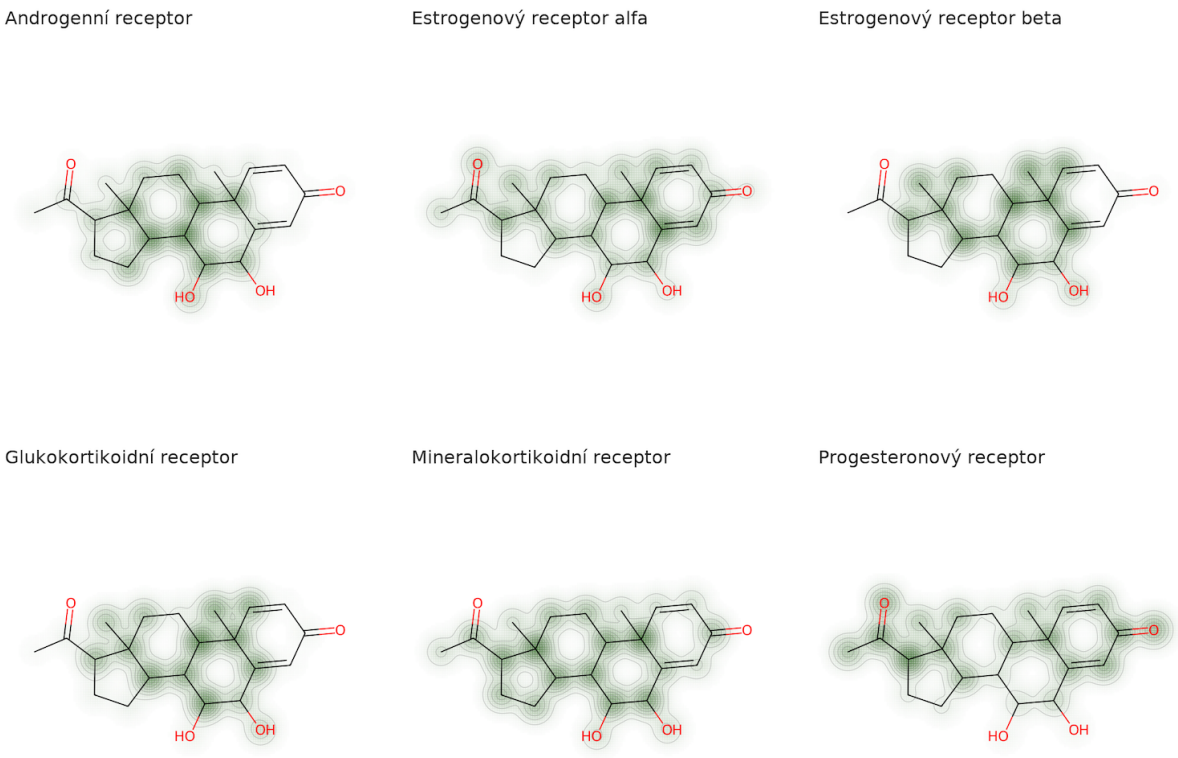
Tabulka 2: Predikované pravděpodobnosti vazby na jednotlivé receptory pro zvolenou molekulu.

5.1.3 Identifikace a vizualizace klíčových strukturních motivů

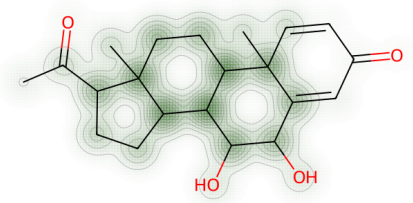
Identifikace klíčových strukturních motivů probíhá ve dvou krocích. Nejprve je pomocí metody Boruta-Shap vyhodnocena důležitost jednotlivých bitů ECFP. Tím je možné určit, které bity nejvíce ovlivňují výslednou predikci modelu.

Následně jsou tyto bity přiřazeny ke konkrétním částem molekuly pomocí dříve vygenerované bit-info mapy. Ta umožňuje zpětně dohledat, jakému strukturnímu fragmentu daný bit odpovídá. Z těchto fragmentů pak bylo možné vytvořit reprezentaci pomocí SMARTS a zobrazit je přímo ve struktuře molekuly.

Modely se pro odhad různých aktivit orientují na různé části téže molekuly. Například při odhadu aktivity na Progesteronový receptor se přihlíží k částem struktury téměř inverzním k Androgennímu a Glukokortikoidnímu receptoru. I u velmi příbuzných Estrogenových receptorů alfa a beta, které sdílí většinu strukturní pozornosti na centrálním kruhovém motivu, je rozdíl u uhlíku přilehlém na postranní oxo (=O) skupině, a na acetylové (-CH(=O)CH3) skupině na protilehlé části molekuly. Takovéto rozdíly právě mohou souviset se vzájemnou selektivitou těchto receptorů.



Obrázek 11: Vizualizace klíčových strukturních motivů modelů XGBoost QSAR pro jednotlivé receptory



Obrázek 12: Vizualizace klíčových strukturních motivů modelu XGBoost QSPR pro rozpustnost ve vodě

Seznam obrázků

Seznam tabulek

Použitá literatura

1. WEININGER, David. *3. SMILES - A Simplified Chemical Language* [online]. [cit. 2026-03-09]. Dostupné z: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.
2. KUMAR, Manish. *A Beginner's Guide for Understanding Extended-Connectivity Fingerprints (ECFPs)* [online]. ChemicBook, 2021-03-25. [cit. 2026-03-10]. Dostupné z: https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html.
3. BREIMAN, Leo. Random Forests. *Machine Learning*. 2001, roč. 45, č. 1, s. 5–32. Dostupné z DOI: 10.1023/A:1010933404324.
4. CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, s. 785–794. Dostupné z DOI: 10.1145/2939672.2939785.
5. XGBOOST DEVELOPERS. *XGBoost Documentation* [online]. 2026. [cit. 2026-03-22]. Dostupné z: https://xgboost.readthedocs.io/.
6. SORKUN, Murat Cihan; KHETAN, Abhishek; ER, Süleyman. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*. 2019, roč. 6, s. 143. Dostupné z DOI: 10.1038/s41597-019-0151-1.
7. ROGERS, David; HAHN, Mathew. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*. 2010, roč. 50, č. 5, s. 742–754. Dostupné z DOI: 10.1021/ci100050t.
8. RDKIT CONTRIBUTORS. *RDKit::MorganFingerprint Namespace Reference*. 2026. Dostupné také z: https://www.rdkit.org/docs/cppapi/namespaceRDKit_1_1MorganFingerprint.html. Online; accessed 2026-03-13.
9. RDKIT CONTRIBUTORS. *rdkit.Chem.rdFingerprintGenerator module*. 2025. Dostupné také z: https://www.rdkit.org/docs/source/rdkit.Chem.rdFingerprintGenerator.html. Online; accessed 2026-03-13.
10. RDKIT CONTRIBUTORS. *rdkit.Chem.Draw package*. 2025. Dostupné také z: https://www.rdkit.org/docs/source/rdkit.Chem.Draw.html. Online; accessed 2026-03-13.

11. GEEKSFORGEEKS. *K-Fold Cross Validation in Machine Learning* [online]. 2025. [cit. 2026-03-10]. Dostupné z: https://www.geeksforgeeks.org/machine-learning/k-fold-cross-validation-in-machine-learning/.

12. DIGITALOCEAN. *K-Fold Cross-Validation in Machine Learning* [online]. 2024. [cit. 2026-03-10]. Dostupné z: https://www.digitalocean.com/community/tutorials/k-fold-cross-validation-python.

13. GEEKSFORGEEKS. *Stratified K Fold Cross Validation* [online]. 2025. [cit. 2026-03-15]. Dostupné z: https://www.geeksforgeeks.org/machine-learning/stratified-k-fold-cross-validation/.

14. JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. *An Introduction to Statistical Learning*. New York: Springer, 2013.

15. SCIKIT-LEARN DEVELOPERS. *Regression metrics* [online]. 2024. [cit. 2026-03-10]. Dostupné z: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics.

16. GOOGLE DEVELOPERS. *Classification: ROC Curve and AUC* [online]. 2023. [cit. 2026-03-10]. Dostupné z: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

17. LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017. Dostupné také z: https://arxiv.org/abs/1705.07874.

18. KEANY, Eoghan. *Boruta-Shap: A Tree based feature selection tool which combines both the Boruta feature selection algorithm with shapley values* [online]. 2026. [cit. 2026-03-10]. Dostupné z: https://github.com/Ekeany/Boruta-Shap.