MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Národní centrum pro výzkum biomolekul

Bakalářská práce

Brno 2019 **Radka Sedláková**

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL

Atomové typy v metodách pro výpočet parciálních atomových nábojů

Bakalářská práce

Radka Sedláková

Vedoucí práce: RNDr. Tomáš Raček Brno 2019

Bibliografický záznam

Autor:	Radka Sedláková Přírodovědecká fakulta, Masarykova univerzita Národní centrum pro výzkum biomolekul
Název práce:	Atomové typy v metodách pro výpočet parciálních atomových nábojů
Studijní program:	Biochemie
Studijní obor:	Chemoinformatika a bioinformatika
Vedoucí práce:	RNDr. Tomáš Raček
Akademický rok:	2018/2019
Počet stran:	počet stran od druhé stránky dokumentu po poslední stránku obsahu + počet stránek práce od první strany Úvodu
Klíčová slova:	parciální atomové náboje, empirické metody, EEM, PEOE, parametrizace, atomový typ, klasifikátor

Bibliographic Entry

Author:	Radka Sedláková Faculty of Science, Masaryk University National Centre for Biomolecular Research
Title of Thesis:	Atom types in methods for calculation of partial atomic charges
Degree Programme:	Biochemistry
Field of Study:	Chemoinformatics and Bioinformatics
Supervisor:	RNDr. Tomáš Raček
Academic Year:	2018/2019
Number of Pages:	počet stran od druhé stránky dokumentu po poslední stránku obsahu + počet stránek práce od první strany Úvodu
Keywords:	partial atomic charges, empirical methods, EEM, PEOE, parameterization, atom type

Abstrakt

Parciální atomové náboje jsou vhodnou charakteristikou pro popis elektrostatických vlastností molekul. Pro jejich výpočet bylo vyvinuto množství kvantově-chemických a empirických metod. Ve výpočtech vybraných empirických metod vystupují neznámé (parametry), které lze získat z experimentálních dat nebo pomocí procesu tzv. parametrizace.

Parametrizace empirických metod je netriviální výpočetní problém, který je mimo jiné ovlivněn dělením atomů do atomových typů. Bakalářská práce se zaměřuje na optimalizaci parametrizace empirických metod, a to skrze analýzu rozdělení atomů do atomových typů na základě vybraných charakteristik (hybridizace, nejvyšší řád vazby, příslušnost ke strukturnímu celku, vazební partneři atomu).

Výsledky klasifikací byly užity pro parametrizaci empirických metod EEM a PEOE. Rozdělení atomů do atomových typů atomů dle příslušnosti ke strukturním celkům poskytly relevantní výsledky, ovšem za cenu velké časové náročnosti. S ohledem na efektivitu a přesnost parametrizace se klasifikace atomů dle nejvyššího řádu vazby a hybridizace prokázaly jako nejvhodnější, a to u obou uvedených metod.

Abstract

Partial atomic charges are fit for describing electrostatic properties of molecules. Several methods have been proposed for their calculation. These divide into methods based on quantum chemistry computations and empirical methods. Empirical calculations depend on unknown variables (parameters) which can be either derived from experimental data or gained through the process of parameterization.

Parameterization of empirical methods is a complex computational problem depending on number of factors, atom type classification namely. This thesis focuses on parameterization of empirical methods optimization through analysis of different atom type classification schemes (based on hybridization, highest bond order, structural motif membership and bonding partners of atom).

Results of proposed classifications were used for parameterization of empirical methods EEM and PEOE. Atom classifications based on structural motif membership resulted relevant, yet being computationally demanding. Regarding the cost-effective solution, atom type classification based on highest bond orded and atom hybridization gave the best results with respect to both of methods mentioned.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2018/2019

Ústav: Národní centrum pro výzkum biomolekul

Studentka: Radka Sedláková

Program: Biochemie

Obor: Chemoinformatika a bioinformatika

Ředitel *Národního centra pro výzkum biomolekul* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s názvem:

Název práce: Atomové typy v metodách pro výpočet parciálních atomových nábojů

Název práce anglicky: Atom types in methods for calculation of partial atomic charges

Oficiální zadání:
Parciální atomové náboje jsou reálná čísla, jejichž cílem je popsat rozložení elektronové hustoty v molekule. Nacházejí mnohá uplatnění ve výpočetní chemii nebo chemoinformatice, konkrétně např. v oblastech jako je QSAR/QSPR modelování či jiné aplikace atomových deskriptorů, molekulový docking nebo virtuální screening. Jelikož se jedná pouze o teoretický koncept, nelze je určit experimentálně a existuje tak velké množství jejich definic. Standardní metody výpočtu vycházejí z kvantové mechaniky, nicméně výpočetní náročnost limituje jejich využití pouze na malé systémy. Empirické metody využívající naměřená nebo parametrizovaná data naproti tomu představují výrazně rychlejší alternativu. Parametrizace těchto metod je však netriviální proces, který je ovlivněn řadou faktorů. Jedním z nich je rozdělení atomů do atomových typů, např. dle nejvyššího řádu vazby, hybridizace, chemického okolí ap.

Cíle bakalářské práce jsou:

1. Seznámit se s problematikou parciálních atomových nábojů a metodami jejich výpočtu.
2. Provést rešerši atomových typů objevujících se v literatuře popisující empirické metody.
3. Definovat klasifikátory, které nalezené atomové typy popisují, a implementovat je jako knihovnu jazyka Python.
4. S využitím externího nástroje provést a vyhodnotit parametrizaci metod EEM [1] a PEOE [2] při zavedení různých klasifikací atomů.

[1] Mortier, Wilfried J., Swapan K. Ghosh, and S. Shankar. "Electronegativity-equalization method for the calculation of atomic charges in molecules." Journal of the American Chemical Society 108.15 (1986): 4315-4320.
[2] Gasteiger, Johann, and Mario Marsili. "Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges." Tetrahedron 36.22 (1980): 3219-3228.

Jazyk závěrečné práce: čeština

Vedoucí práce: RNDr. Tomáš Raček

Datum zadání práce: 10. 10. 2018

V Brně dne: 30. 4. 2019

Souhlasím se zadáním (podpis, datum):

.....	.....	.....
Radka Sedláková	RNDr. Tomáš Raček	doc. RNDr. Oldřich Janiczek, CSc.
studentka	**vedoucí práce**	**zástupce ředitele Národního centra pro výzkum biomolekul pro pedagogické záležitosti**

Poděkování

Na tomto místě bych chtěla poděkovat vedoucímu mé bakalářské práce RNDr. Tomáši Račkovi, a to za cenné rady, ochotu a trpělivost. Děkuji Bc. Ondřeji Schindlerovi za pomoc s technickými záležitostmi spojenými s tvorbou této práce, kterou mi byl ochoten poskytnout i ve chvílích velkého časového vytížení. Velké díky patří mé rodině a blízkým, kteří mi jsou oporou nejen při studiu. Na závěr děkuji Daniele Gachové za sdílení radostí i strastí spojených s psaním této práce a organizaci MetaCentrum za umožnění potřebných výpočtů.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracovala samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 15. května 2019

.....
Radka Sedláková

Obsah

Přehled použitého značení

Pro jednodušší orientaci v textu bakalářské práce uvádím seznam zkratek, které jsou v textu použity.

QM	Quantum Mechanics
SE	Schrödinger's Equation
QC	Quantum Chemistry
DFT	Density Functional Theory
EEM	Electronegativity Equalization Method
PEOE	Partial Equalization of Orbital Electronegativity
MAE	Maximum Average Deviation
RMSD	Root Mean Square Deviation
PCC	Pearson Correlation Coeffecient

Kapitola 1

Úvod

Vzájemná interakce molekul úzce souvisí s jejich reaktivitou. Nejběžnější způsob popisu reaktivity molekul je skrze elektrostatické vlastnosti, které jsou přímo odvozené z rozložení elektronů v molekule. Užitečným nástrojem popisu rozložení elektronové hustoty jsou parciální atomové náboje [1], neboť aproximují elektronovou hustotu v okolí každého atomu v molekule na reálné číslo.

Jelikož jsou parciální náboje pouze teoretickým konceptem a nelze je získat pomocí experimentu, jsou jejich hodnoty stanoveny pomocí výpočetních metod. Standardní přístup představují kvantově-chemické metody, které vycházejí z exaktního řešení Schrödingerovy rovnice [2] a pro velké systémy jsou z hlediska náročnosti výpočtů prakticky nepoužitelné. Empirické metody výpočtu parciálních atomových nábojů naopak představují rychlejší a zároveň poměrně přesnou alternativu kvantově-chemického přístupu, optimalizace těchto metod je proto aktuálním tématem výpočetní chemie.

Cílem vybraných empirických metod (*Electronegativity Equalization Method*, EEM [3], *Partial Equalization of Orbital Electronegativity*, PEOE [4]) je reprodukovat hodnoty nábojů získané kvantově-chemickými metodami, a to za použití vhodných empirických parametrů. Parametry empirických výpočtů lze odvodit z experimentálních dat, v některých případech je však nutno jejich hodnoty určit explicitně. Parametrizace empirických metod je proces, v rámci kterého jsou hledány hodnoty těchto empirických parametrů s cílem reprodukovat za jejich použití referenční náboje vypočtené vybranou kvantově-chemickou metodou. Počet hledaných parametrů se odvíjí od počtu atomových typů definovaných pro popis zvolené molekulové sady.

Cílem této práce je analyzovat výsledky parametrizací metod EEM a PEOE v závislosti na aplikaci různých dělení atomů do atomových typů. Dílčími úkoly práce jsou:

- seznámit se s problematikou parciálních atomových nábojů a s metodami jejich výpočtu
- provést rešerši atomových typů objevujících se v literatuře popisující empirické metody
- definovat klasifikátory, které nalezené atomové typy popisují, a implementovat je jako knihovnu jazyka Python
- s využitím externího nástroje provést a vyhodnotit parametrizaci metod EEM a PEOE při zavedení různých klasifikací atomů

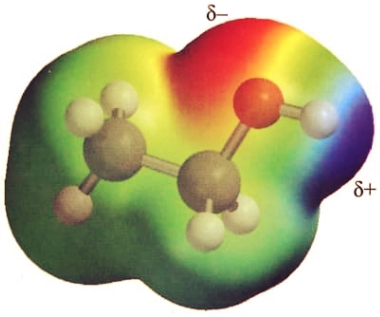
Kapitola 2

Teorie

Teoretická část práce seznamuje s konceptem parciálních atomových nábojů a s metodami jejich výpočtu. Blíže popisuje teoretické základy empirických metod EEM a PEOE, které byly použity v praktické části práce, podobně jako uvedené statistické veličiny.

2.1 Parciální atomové náboje

Parciální atomové náboje jsou reálná čísla, která popisují asymetrické rozložení elektronové hustoty na chemické vazbě [1]. Vznikají v důsledku rozdílných elektronegativit vazebných partnerů. Pokud v chemické vazbě figuruje vysoce elektronegativní atom, pak tento k sobě přitahuje vazebný elektronový pár, čímž se zvyšuje elektronová hustota v jeho okolí a dochází ke vzniku parciálního záporného náboje (δ-). V okolí elektropozitivnějšího vazebného partnera se elektronová hustota naopak snižuje a na atomu dochází ke vzniku parciálního kladného náboje (δ+).



Obrázek 2.1: Parciální atomové náboje v molekule ethanolu. Červená barva značí zvýšenou elektronovou hustotu v blízkosti atomu kyslíku.

Koncept parciálních atomových nábojů je pouze teoretický, hodnoty nábojů proto nelze získat pomocí experimentu [5]. Jelikož se parciální atomové náboje uplatňují při predikci fyzikálních, chemických a biologických vlastností molekul, bylo pro jejich stanovení vyvinuto množství výpočetních metod. Tyto se dělí na metody kvantově-chemické a metody empirické. Kvantově-chemické metody představují standardní přístup

výpočtu parciálních atomových nábojů, avšak použití těchto metod může být limitováno jejich velkou časovou náročností. Empirické metody používají v rámci výpočtů parametrizovaná nebo exprimentálně naměřená data a představují tak časově méně náročnou alternativu kvantově-chemického přístupu [6].

Aplikaci parciálních atomových nábojů lze nalézt ve výpočetní chemii a chemoinformatice, kde slouží k predikci elektrostatických vlastností popisujících reaktivitu molekul. Uplatňují se v molekulových simulacích [7], ve virtuálním screeningu [8], při hledání vazebných míst proteinů nebo při návrhu farmakoforů [9]. Prokázaly se jako platné deskriptory v QSAR a QSPR modelech [10, 11]. V anorganické chemii se uplatňují při popisu toku elektronů v bateriích a katalyzátorech [12].

2.2 Kvantově-chemické metody

V této podkapitole je popsán teoretický aparát kvantové mechaniky, na kterém jsou založeny kvantově-chemické metody, které lze použít pro výpočet parciálních atomových nábojů. Tyto metody jsou dále popsány.

2.2.1 Základy kvantové mechaniky

Kvantová mechanika se rozvinula ve 20. letech 20. století v reakci na newtonovskou mechaniku, jejíž aparát již nepostačoval pro popis mikrosvěta. Základním principem QM je vlnově-korpuskulární dualismus, který mikročástici připisuje jak charakteristiky hmoty (hybnost), tak charakteristiky elektromagnetické vlny šířící se prostorem. Vlna, popisující částici, je v kvantové mechanice reprezentována matematickou funkcí Ψ, tzv. vlnovou funkcí. Tato funkce popisuje dynamický stav částice a nese informaci o výskytu částice v prostoru [13]. Vlnová funkce elektronu tak popisuje rozložení elektronové hustoty v molekule.

Základním úkolem kvantové mechaniky je výpočet vlnové funkce systému, z níž lze odvodit elektrostatický potenciál [14] nebo termodynamické vlastnosti molekuly [15]. Vlnová funkce je řešením Schrödingerovy rovnice (*Schrödinger's Equation*, SE)

$$H\Psi = E\Psi \tag{2.1}$$

kde H je operátor hamiltonián a E je energie systému. Hamiltonián působí na vlnovou funkci Ψ a transformuje ji na funkci jinou. Řešením Schrödingerovy rovnice je soubor funkcí, které lze po aplikaci hamiltoniánu zapsat jako součin původní funkce a skaláru E. Takovéto funkce označujeme jako *vlastní funkce* a odpovídající skaláry jako *vlastní hodnoty* operátoru [2].

Schrödingerova rovnice je exaktně řešitelná pouze pro vybrané problémy, např. pro atom vodíku. Pro víceelektronové systémy je nutno do výpočtu zavádět velké množství aproximací, z nichž nejznámější je Born-Oppenheimerova aproximace [16]. Jejím základním konceptem je oddělení řešení SE pro jádra od řešení rovnice elektronů. Tento postup vychází z předpokladu, že jádra atomů, mnohonásobně těžší než elektrony, se pohybují výrazně pomaleji než elektrony samotné, a jejich polohu lze tedy pro řešení SE elektronů pokládat za fixní.

2.2.2 Úvod do kvantově-chemických metod

Cílem kvantově-chemických (QC) metod je exaktní popis chemických vlastností systému. Jednou ze zkoumaných vlastností je rozložení elektronové hustoty, které je odvozeno z vlnové funkce molekuly. Znalost rozložení elektronové hustoty v molekule je klíčovým faktorem pro odvození parciálních atomových nábojů.

Kvantově-chemické metody se dělí na tři hlavní skupiny, a to metody semi-empirické, metody odvozené od teorie funkcionálu hustoty a metody *ab initio*. *Ab initio* metody (lat. *ab initio* - od počátku) staví výpočty na teoretickém aparátu a Schrödingerovu rovnici řeší exaktně, pouze za použití fyzikálních konstant, z čehož vyplývá jejich velká výpočetní náročnost (výpočty těchto metod škálují až s n^7, kde n je počet elektronů systému [17]). Metody semi-empirické jsou stejně jako metody *ab initio* založeny na řešení SE a pro urychlení výpočtů využívají kromě aproximací také data z experimentu. Metody odvozené z DFT nevycházejí z řešení vlnové fukce, ale své poznatky staví na elektronové hustotě v molekule [5, 6, 18].

Postup kvantově-chemické výpočtů pro popis rozložení elektronové hustoty se typicky skládá z následujících kroků: výběr metody, výběr bázové sady a provedení populační analýzy.

Příklady kvantově-chemických metod

Následující příklady popisují Hartree-Fockovu metodu, metody odvozených z teorie funkcionálu hustoty a metody semi-empirické.

- **Hartree-Fockova metoda**

  Tato raná *ab initio* metoda řeší SE rozložením původní n-elektronové vlnové funkce na řešení n jednoelektronových rovnic. Tyto rovnice jsou určeny předpisem $\hat{F}\chi_i = \varepsilon_i\chi_i$, kde $\hat{F}$ je Fockův operátor (Hartreeho-Fockův hamiltonián) aplikovaný na jednoelektronový orbital χ_i. Metoda pracuje iterativním způsobem, a to až do ustálení výsledných hodnot vlnových funkcí. Tento přístup se označuje termínem selfkonzistentní pole (angl. *self-consistent field*, SCF) [19].

- **DFT - Density Functional Theory**

  Metody založené na teorii funkcionálu hustoty nenásledují schéma kvantově-chemických metod, ale své výpočty staví na rozložení elektronové hustoty, ze kterého odvozují energii systému a další vlastnosti molekuly [20]. Elektronová hustota je funkcí souřadnic x, y, z. V porovnání s řešením Schrödingerovy rovnice, která pro systém s n elektrony obsahuje $4n$ neznámých, je její výpočet výrazně jednodušší. Termín 'funkcionál' je v kontextu DFT chápán jako zobrazení, které zobrazuje funkci, představující elektronovu hustotu, do množiny reálných čísel popisujících energii elektronů [21].

- **Semi-empirické metody**

  Semi-empirické metody část výpočtů aproximují a při řešení rovnic aplikují parametry odvozené z experimentálních dat, a to s cílem reprodukovat výsledky *ab intio* výpočtů. Příkladem semi-empirické metody je metoda CNDO (*Complete*

Neglect of Differential Overlap) [22]. Metoda zamítá interakce atomových orbitalů lokalizovaných na různých atomech molekuly a pracuje pouze s interakcemi atomových orbitalů stejného typu lokalizovaných na stejném atomu. Tato aproximace byla v metodách navazujících na CNDO (INDO [23], MNDO [24]) rozšířena o interakci orbitalů lokalizovaných na jiných jádrech a interakci volných elektronových párů.

Bázová sada

Bázová sada je soubor vlnových funkcí reprezentující atomové orbitaly, jejichž vhodnou lineární kombinací (LCAO) lze následně vyjádřit vlnovou funkci molekuly. Pro popis funkcí reprezentujících atomové orbitaly se používají orbitaly Gaussova typu (GTO). Kombinace několika Gaussových orbitalů přibližuje tzv. Slaterův orbital (STO), který je pro výpočet vlnové funkce molekuly méně vhodný z důvodu složitosti výpočtů. Příkladem bázových sad jsou sady STO-3G, STO-4G či obecně STO-*n*G, kde *n* je počet orbitalů Gaussova typu reprezentujících jeden atomový orbital. Dalšími bázovými sadami jsou např. 6-31G nebo 6-21G* [25].

Populační analýza

Pomocí populační analýzy je získáno zastoupení elektronů v molekulových orbitalech (rozložení elektronové hustoty), ze kterého lze odvodit parciální atomové náboje.

V rámci Mullikenovy populační analýzy (*Mulliken Population Analysis*, MPA) [26] není brána v potaz rozdílnost elektronegativit vazebných partnerů a elektronová hustota na vytvořené chemické vazbě je tedy rovnoměrně rozdělena mezi příslušné atomy. Výsledky MPA jsou silně závislé na zvolené kvantově-chemické metodě a na velikosti bázové sady. Nevýhody MPA, zejména nepřesnost výsledků související s rozšiřováním bázové sady, řeší přirozená populační analýza (*Natural Population Analysis*, NPA) [27], pracující s přirozenými atomovými orbitaly. Přirozené atomové orbitaly jsou odvozeny z bázové sady a jsou následně použity pro výpočet ortonormálních přirozených vazebných orbitalů (*Natural bonding orbitals*, NBO). Na základě NBO se poté provadí populační analýza.

Odlišný přístup finálního výpočtu parciálních atomových nábojů představuje metoda *Atoms-in-Molecules* (AIM) [28], která přiřazuje náboje atomům na základě integrace elektronové hustoty přes prostor příslušící danému atomu.

2.3 Empirické metody

Empirické metody výpočtu parciálních atomových nábojů nevycházejí z řešení vlnové funkce systému, ale definují vlastní postupy výpočtu. Vybrané empirické metody se snaží reprodukovat náboje získané kvantově-chemickými přístupy skrze parametrizaci vůči referenční sadě nábojů, jiné definují parciální atomové náboje vlastním způsobem. Narozdíl od kvantově-chemických metod se metody empirické vyznačují nízkou výpočetní náročností.

Empirické metody se dělí na dvě hlavní skupiny, a to metody pracující s topologií molekuly (popisuje počet a typ vazebných partnerů, násobnost vazeb apod.) a metody pracující s prostorovým uspořádáním molekuly. Metody zastupující obě uvedené skupiny, jmenovitě metoda PEOE a metoda EEM, jsou popsány v odstavcích níže.

2.3.1 PEOE

Metoda PEOE (*Partial Equalization of Orbital Electronegativity*) je také známá pod jménem autorů jako metoda Gasteiger-Marsili [4]. V rámci výpočtu neuvažuje 3D strukturu molekuly a pracuje pouze s její topologií. Metoda byla navržena pouze pro systémy obsahující σ vazby a nekonjugované π vazby, metody odvozené z PEOE však původní metodu výrazně rozšířily, a to např. o aplikaci na halogenované a aromatické sloučeniny [29] nebo o výpočty založené na prostorovém uspořádání molekuly [30].

Koncept elektronegativity atomových orbitalů, na němž je metoda PEOE založena, vychází z Mullikenovy definice elektronegativity χ_A atomu A

$$\chi_A = \frac{1}{2}(I_A + E_A) \tag{2.2}$$

Dle Mullikena je elektronegativita atomu určena hodnotami elektronových afinit E_A a ionizačních potenciálů I_A jeho valenčních stavů. PEOE připisuje na základě hodnot I_A a E_A elektronegativitu každému orbitalu valenčního stavu atomu. Elektronegativita χ_{iv} orbitalu iv na atomu i

$$\chi_{iv} = a_{iv} + b_{iv}Q_i + c_{iv}Q_i^2 \tag{2.3}$$

je ovlivněna náboji ostatních orbitalů a tedy i celkovým nábojem příslušného atomu Q_i. Koeficienty a_{iv}, b_{iv} a c_{iv} jsou empirické parametry vypočtené z ionizačních potenciálů a elektronových afinit neutrálního, kationtového a aniontového stavu příslušného orbitalu.

Při vzniku vazby dochází vlivem elektronegativity atomů k přesunu elektronů od elektropozitivnějšího atomu směrem k elektronegativnějšímu. Interagují spolu příslušné atomové orbitaly a dochází k částečné ekvalizaci (vyrovnání) jejich nábojů. Metoda PEOE pracuje iterativním způsobem. Množství přeneseného náboje mezi atomy A a B, pokud má atom B vyšší elektronegativitu, je definováno jako

$$Q^{\langle k \rangle} = \frac{\chi_B^{\langle k \rangle} - \chi_A^{\langle k \rangle}}{\chi_A^+} \cdot \left(\frac{1}{2}\right)^k \tag{2.4}$$

kde χ_A^+ označuje elektronegativitu kationtu atomu A a k iteraci výpočtu. Iniciální výpočet elektronegativity orbitalu (2.3) pracuje s formálním nábojem atomu. Po výpočtu příspěvků přenesených nábojů (2.4) všech vazebných partnerů atomu je náboj daného atomu přepočítán a použit v další iteraci. Množství přeneseného náboje mezi dvěma atomy se v každé iteraci výpočtu snižuje a vypočtené hodnoty nábojů atomů postupně konvergují. Přibližně po šesté iteraci dochází k ustálení výsledných hodnot.

Parciální atomové náboje vypočtené metodou PEOE se prokázaly vhodné pro predikci chemických posunů v rámci elektronové spektroskopie (*Electron Spectroscopy for Chemical Analysis*, ESCA) [4] a ^{13}C NMR [31].

2.3.2 EEM

Metoda vyrovnání elektronegativity (*Electronegativity Equalization Method*, EEM) [3] byla publikována v r. 1986. Teoretický základ metody vychází z teorie funkcionálu hustoty, na němž je vystavěn matematický aparát pro výpočet atomových nábojů. Cílem metody je přiblížit vypočtené hodnoty parciálních atomových nábojů hodnotám získaným pomocí kvantově-chemických metod. Metoda škáluje s n^3, kde n je počet atomů v systému. Na principu EEM byly vyvinuty další empirické metody, např. *Selfconstistent Functional Kernel Equalized Electronegativity Method* (SFKEEM) [32] nebo *Atom-Bond Electronegativity Equalization Method* (ABEEM) [33, 34].

Výchozím konceptem metody je Sandersonův princip ekvalizace elektronegativity. Dle něj je každému atomu přiřazena stejná elektronegativita jako je elektronegativita ostatních atomů molekuly. Podle rovnice

$$\overline{\chi} = \chi_1 = \chi_2 = \chi_3 = ... = \chi_N \tag{2.5}$$

kde N je počet atomů, se elektronegativita každého atomu rovná průměrné elektronegativitě molekuly $\overline{\chi}$.

Další základním principem metody je princip zachování náboje. Celkový náboj molekuly Q odpovídá součtu dílčích atomových nábojů q_i.

$$\sum_i q_i = Q \tag{2.6}$$

Třetí základní princip představuje efektivní elektronegativita χ_i atomu i. Jelikož metoda pracuje s prostorovým uspořádáním molekuly, je při určování elektronegativity atomu i bráno v potaz jeho molekulové okolí. Sumace ve vzorci reprezentuje elektrostatickou interakci atomu i s atomy j v závislosti na jejich vzdálenosti R_{ij}. N značí počet atomů v molekule.

$$\chi_i = A_i + B_i \cdot q_i + \kappa \sum_{j \neq i}^{N} \frac{q_j}{R_{ij}} \tag{2.7}$$

V rovnici 2.7 kromě nábojů q_i, q_j interagujících atomů vystupují empirické parametry A_i, B_i a κ. Parametry A_i a B_i zahrnují elektronegativitu χ_i^0 a tvrdost η_i^0 neutrálního izolovaného atomu a korekce $\Delta\chi_i$, $\Delta\eta_i$, které upravují výslednou elektronegativitu χ_i atomu na základě jeho interakce s prostředím molekuly.

$$A_i = \chi_i^0 + \Delta\chi_i \tag{2.8}$$

$$B_i = 2(\eta_i^0 + \Delta\eta_i) \tag{2.9}$$

Řešení systému s N atomy vede po kombinaci vztahů 2.5, 2.6 a 2.7 na systém N+1 lineárních rovnic o N+1 neznámých (rov. 2.10). Z této matice jsou vypočteny hodnoty nábojů q_1, q_2, ..., q_N a elektronegativita molekuly $\overline{\chi}$. Pro výpočet nábojů je nutná znalost parametrů A_i, B_i a κ. Tyto jsou získány pomocí parametrizace (2.3.3) nebo jsou použity hodnoty již vypočtených parametrů.

$$\begin{pmatrix} B_1 & \frac{\kappa}{r_{1,2}} & \cdots & \frac{\kappa}{r_{1,n}} & -1 \\ \frac{\kappa}{r_{1,2}} & B_2 & \cdots & \frac{\kappa}{r_{2,n}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{r_{1,n}} & \frac{\kappa}{r_{2,n}} & \cdots & B_n & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \\ \overline{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_n \\ Q \end{pmatrix} \tag{2.10}$$

Metoda EEM byla úspěšně použita např. pro výpočet parciálních atomových nábojů zeolitů [35] nebo proteinů [36]. Parciální atomové náboje vypočetené metodu EEM byly použity pro predikci pK_a [11].

2.3.3 Parametrizace empirických metod

Cílem parametrizace empirických metod pro výpočet parciálních atomových nábojů je nalézt hodnoty empirických parametrů (v kontextu EEM mluvíme o parametrech A_i, B_i a κ) užitých v průběhu výpočtu tak, aby hodnoty nábojů získané empirickou metodou co nejlépe reprodukovaly náboje získané kvantově-chemickým přístupem. Hodnoty parametrů jsou hledány pro každý definovaný atomový typ v molekulové sadě.

Parametrizace empirických metod pro výpočet parciálních atomových nábojů se skládá z následujících kroků:

1. výběr tréninkové a testovací sady molekul obsahující atomy, které v dostatečné míře reprezentující atomové typy, pro něž hledáme hodnoty empirických parametrů
2. výpočet parciálních atomových nábojů tréninkové sady pomocí kvantově-chemické metody (náboje vypočtené QC metodami se označují jako náboje referenční)
3. výpočet parametrů empirické metody na základě nábojů získaných v kroku 2
4. výpočet nábojů testovací sady molekul kvantově-chemickou a parametrizovanou empirickou metodou

Parametrizace empirických metod je optimalizačním problémem. Nalezení parametrů, pomocí nichž jsou vypočteny náboje korespondující s QC metodami, odpovídá nalezení globálního minima objektivní funkce, která popisuje podobnost mezi referenčními náboji a náboji vypočtenými pomocí empirické metody. Příkladem globální optimalizační metody pro parametrizaci empirických metod je metoda Guided Minimization [37].

2.3.4 Atomové typy

Pro každý atomový typ jsou v rámci parametrizace hledány vhodné hodnoty parametrů vystupujících ve výpočtech parciálních atomových nábojů. Napříč parametrizacemi empirických metod zaznamenáváme různé úrovně návrhu atomových typů, od triviálních klasifikací definujících atomové typy na základě protonových čísel [32] až po komplexní rozdělení zahrnující funkční skupiny či příslušnost k větším strukturním celkům (aromatické systémy, postranní řetězce aminokyselin [30]). Detailní dělení atomových

typů lze nalézt zejména v publikacích orientujících se na parametrizaci metod pro výpočet parciálních nábojů komplexních celků, např. polypeptidů [38]. Tabulka 2.1 ukazuje atomové typy definované v publikaci „Kirchhoff Atomic Charges Fitted to Multipole Moments: Implementation for a Virtual Screening System“ [39].

Hrubé klasifikace atomů do atomových typů, např. na základě protonového čísla, nesou riziko nepřesných výpočtů atomových nábojů, neboť neberou v potaz chemické okolí atomu. Detailní dělení atomových typů naopak způsobuje špatnou přenositelnost klasifikace mezi tréninkovou a validační sadou molekul, neboť nese riziko přetrénování tréninkové sady. Parametrizace vybrané empirické metody může být detailním dělením atomů do atomových typů výrazně ztížena, neboť v přímé závislosti na množství atomových typů roste počet hledaných empirických parametrů.

atomový typ	**elektronegativita** (V)	**tvrdost** (V/e)
C sp^3	26,31	9,50
C sp^2 H_2C=	29,43	12,54
C arom. cycle 6	28,98	9,49
C sp	31,02	9,30
N sp^3	31,44	9,48
N pyridine	33,07	11,05
N cyano	42,45	11,36
O sp^3	33,34	15,48
O sp^3 in carboxyl	32,85	10,97
O^- phenol	29.60	49.14
O sp^2	37,56	5,17
S sp^3	30,40	7,92
S sulfoxide	30,40	8,30
H (O sp^3)	26,32	1,03
H (O resonance)	26,66	8,35
H (N sp^3)	26,65	5,16
H (N resonance)	26,58	8,52
H (S sp^3)	26,70	8,98
H (S sp^2)	26,68	9,62

Tabulka 2.1: Atomové typy definované v rámci empirické metody *Kirchoff charge model* (KCM). Parametry metody jsou elektronegativita a tvrdost atomu. Pro ilustraci jsou uvedeny pouze vybrané atomové typy a vypočtené parametry.

2.4 Statistické pojmy

Úspěšnost parametrizace empirických metod se hodnotí na základě srovnání hodnot empiricky vypočtených nábojů s referenčními hodnotami, a to na tréninkové nebo validační sadě. K tomuto srovnání slouží uvedené statistické veličiny. Náhodné veličiny X a Y ve vzorcích reprezentují sadu referenčních a empirických nábojů, proměnné x_i a y_i označují odpovídající hodnoty nábojů srovnávané dvojice. Proměnná n značí velikost datového souboru.

2.4.1 Průměrná a maximální absolutní odchylka

Průměrná absolutní odchylka (ang. *Mean Absolute Error*, MAE) je dána aritmetickým průměrem absolutních hodnot rozdílů hodnot x_i a y_i příslušných náhodných veličin. Po odstranění absolutních hodnot by vzorec popisoval tzv. *Mean Bias Error* (MBE).

$$\mathrm{MAE}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{2.11}$$

Maximální absolutní odchylka popisuje největší rozdíl nalezený mezi hodnotami x_i a y_i náhodných veličin X a Y [40].

$$\mathrm{ABSMAX}(X,Y) = \max_{1\leq i\leq n}|x_i - y_i| \tag{2.12}$$

2.4.2 RMSD

Veličina RMSD (*Root Mean Square Deviation*, někdy uváděná též jako *Root Mean Square Error*) [41] popisuje míru odlišnosti dvojic hodnot (x_i, y_i) napříč datovým souborem. Je definována jako odmocnina ze střední kvadratické chyby (*Mean Square Deviation*, MSD). Stejně jako rozptyl je tato veličina kvůli kvadrátu rozdílu hodnot x_i a y_i citlivá na odlehlé a chybné hodnoty, které se promítají do vyšších výsledných hodnot RMSD srovnávaných datových sad.

$$\mathrm{RMSD}(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.13}$$

2.4.3 Pearsonův korelační koeficient

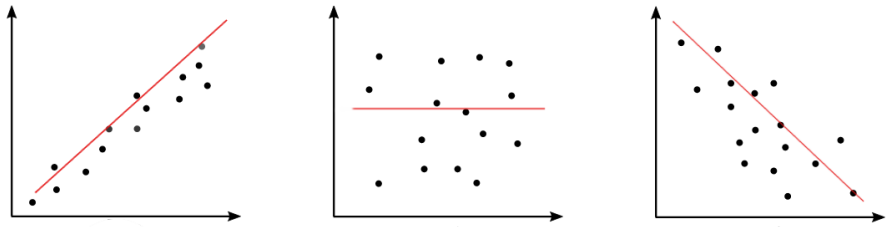
Pro kvantifikaci funkčního vztahu dvou sledovaných veličin užíváme tzv. Pearsonův korelační koeficient (*Pearson Correlation Coefficient*, PCC) [42]. PCC popisuje míru linearity závislosti veličiny Y na veličině X (lineární korelaci), pro popis jiných typů závislostí (např. kvadratických) není vhodný. Je definován jako

$$\mathrm{r}(X,Y) = \frac{\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.14}$$

kde $\overline{x}$ a $\overline{y}$ jsou aritmetické průměry naměřených hodnot veličin X a Y.

PCC nabývá hodnot z intervalu $\langle -1, 1 \rangle$, přičemž hodnoty koeficientu blízké číslu -1 nebo 1 indikují silnou lineární korelaci mezi pozorovanými veličinami. Linearita vztahu je dobře pozorovatelná v grafu (viz obr 2.2), kde jsou dvojice hodnot (x_i, y_i) znázorněny jako body v dvourozměrné soustavě souřadnic. Interpretace hodnoty k Pearsonova korelačního koeficientu je následující:

- pokud je k kladné, pak veličiny X a Y vykazují kladnou korelaci (pokud se hodnota veličiny Y zvětšuje, pak hodnota X roste)
- pokud je k záporné, pak veličiny X a Y vykazují zápornou korelaci (v závislosti na zvětšující se hodnotě veličiny Y hodnota X klesá)
- pokud je k rovno 0, pak veličiny X a Y nejsou lineárně korelované



Obrázek 2.2: Vizualizace lineární korelace dvou náhodných veličin. Směrem zleva zobrazují grafy lineární závislost veličin s hodnotami Pearsonova korelačního koeficientu 0,9; 0 a -0.8.

Kapitola 3

Metody

V následující sekci je popsán formát vstupních souborů vytvořeného programu a knihovny použité v rámci implementace. Zmíněn je též externí nástroj MACH, kterým byla pro klasifikované molekulové sady provedena a vyhodnocena parametrizace vybraných empirických metod.

3.1 Structure-Data file (SDF)

Formát SDF [43, 44] patří s formáty Molfile, RGfile, rxnfile, RDfile a XDfile mezi CTfile formáty (*Chemical Table file*) vyvinuté pro reprezentaci chemických dat. SDF je rozšířením formátu Molfile (zkr. MOL). Umožňuje zápis více záznamů do jednoho souboru, přičemž každý záznam ukončený sekvencí '$$$$' reprezentuje jednu molekulu.

Záznamy v SDF souboru mají pevně danou strukturu, odvozenou od struktury MOL souborů. Společnými částmi záznamu obou formátů jsou tzv. *header block* a tzv. *connection table* (viz obr. x). V SDF záznamu mohou být narozdíl od formátu MOL za řádek '`M END`' připojeny specifikace biologických či fyzikálně-chemických vlastností dané molekuly.

Struktura SDF záznamů je následující:

- *Header block* se skládá ze tří řádků obsahujících název molekuly, datum vytvoření záznamu, program použitý pro generování záznamu a komentář. Všechny tři řádky mohou být prázdné.
- *Counts line* obsahuje na definovaných indexech řádku počet atomů a vazeb popsaných v sekcích *Atom block* a *Bond block*, informaci o chiralitě molekuly a verzi molekulového záznamu (V2000 nebo V3000).
- *Atom block* obsahuje souřadnice atomů x, y, z a symbol prvku. Další indexy řádků, ve většině případů obsazené symbolem '0', slouží pro bližší specifikaci vlastností atomů. Na základě pořadí atomů v *Atom block*u jsou v sekci *Bond block* specifikováni vazební partneři a typ vytvořené vazby.
- *Data items* slouží pro doplňující záznamy vlastností molekuly. Řádek *Data header*, začínající znakem '`>`', obsahuje název dané vlastnosti nebo identifikační číslo molekuly v databázi MACCS-II. Následují řádky s příslušnými hodnotami.

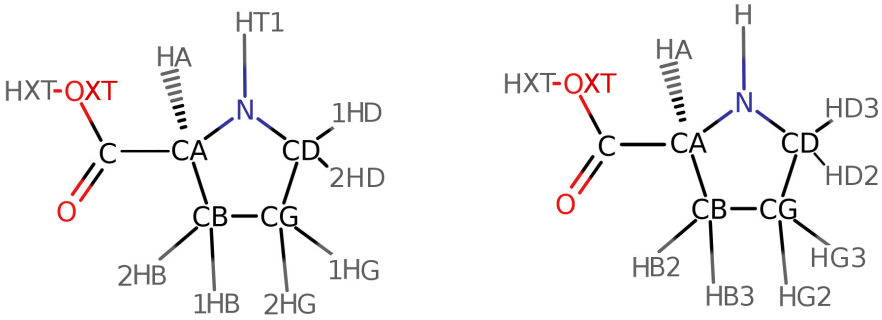
3.2 PDB formát

Formát PDB (*Protein Data Bank*) je molekulový formát vyvinutý pro počítačovou reprezentaci biologicky aktivních makromolekul. Existuje ve dvou verzích, v původní verzi 2.0 a v novější verzi 3.0. V PDB souboru jsou obsaženy informace o prostorovém uspořádání molekuly, strukturní faktory pro rentgenovou krystalografii a data z NMR experimentu. PDB formát je v současnosti nahrazován standardizovaným formátem mmCIF.

PDB formát nese kromě dat o prostorovém uspořádání struktury informace o primární a sekundární struktuře nukleových kyselin a proteinů. Informace ohledně 3D struktury nesou záznamy (tj. řádky PDB souboru) začínající sekvencí 'ATOM'. Tyto záznamy obsahují na definovaných pozicích sériové číslo atomu, zkratku residua, jemuž atom přísluší, x, y a z souřadnice, symbol prvku apod. Záznamy 'HETATM' označují nestandardní residua (např. ligandy nebo kofaktory enzymů) a mají stejnou strukturu jako záznamy 'ATOM'. Záznamy 'HELIX' a 'SHEET' popisují sekundární strukturu proteinů. 'SSBOND' je záznam vyhrazený pro specifikaci disulfidických můstků mezi cysteinovými residui [45, 46].

3.2.1 Nomenklatura atomových typů aminokyselin

PDB soubory používají pro specifikaci atomů proteinogenních aminokyselin atomové typy deklarované nomenklaturou IUPAC [47]. Tato nomenklatura označuje atomy vedlejšího řetězce aminokyseliny dle vzdálenosti od uhlíku s navázanou karboxylovou skupinou a aminoskupinou. Řecká písmena v označení atomů (C^{α}, C^{β}, C^{γ}, C^{δ}, C^{ε}, C^{ζ}, C^{η}) jsou v nomenklatuře nahrazena velkými písmeny latinské abecedy (CA, CB, CG, CD, CE, CZ, CH). Pro specifické atomy aminokyselin (vodík navázaný na C^{α}, kyslík COO^{-} skupiny hydrolyzovaný při tvorbě peptidické páteře) jsou definovány speciální atomové typy (HA, OXT), atomy aromatických jader však nejsou odlišeny. Mezi verzemi 2.0 a 3.0 PDB formátu došlo ke změně názvů vybraných atomových typů, popsaná nomenklatura tak není napříč PDB soubory plně konzistentní.



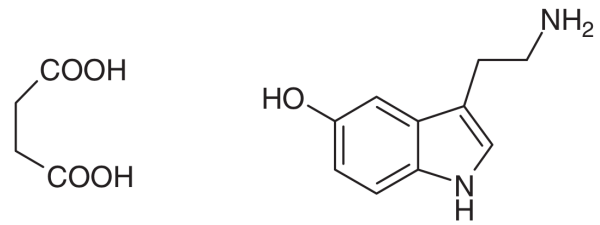
Obrázek 3.1: Porovnání názvů atomových typů molekuly prolinu v jednotlivých verzích PDB formátu. Vlevo je zobrazena verze 2.0, vpravo verze 3.0. Změny v názvech atomů se týkají atomů vodíku.

3.3 SMILES

SMILES [48, 49, 50] (angl. *Simplified Molecular Input Line Entry System*) je počítačová notace molekul či molekulových reakcí definovaná pomocí ASCII symbolů. SMILES reprezentace byla vyvinuta v 80. letech pro usnadnění práce s chemickými daty a zvýšení efektivity jejich zpracování (např. prohledávání molekulových databází, vyhledávání podstruktur v molekulách). Zápis SMILES vychází z teorie grafů. Molekula je v tomto kontextu chápána jako graf, tzn. uspořádaná dvojice množiny vrcholů a množiny hran $G(V, E)$. Průchodem molekulového grafu vzniká jednoznačný SMILES zápis molekuly, kde je každý vrchol (atom) a každá hrana (vazba) navštívena pouze jednou.

Základními prvky SMILES notace jsou symboly atomů a vazeb. Atomy jsou reprezentovány symbolem příslušného prvku. Pokud se jedná o atom aromatický, je pro jeho specifikaci použito malé písmeno (SMILES notace benzenu je '`c1ccccc1`', cyklohexanu '`C1CCCCC1`'). Atomy vodíku jsou implicitně doplněny na základě valence základního stavu atomu, na který jsou navázány, a nemusí být zadány explicitně. Pro specifikaci počtu navázaných vodíků je třeba použít zápisu '`[AHX]`', kde `A` je symbol prvku a `X` počet navázaných vodíků. Vazba jednoduchá, dvojná, trojná a aromatická jsou reprezentovány symboly '–', '=', '#' a ' : '. Vazby jednoduché a aromatické nejsou ve většině SMILES výrazů explicitně zadány.

SMILES notace definuje zápis strukturních prvků sloučenin jako jsou cykly, větvení řetězců, chiralita a izomerie (E/Z a cis/trans izomerie). Přítomnost cyklu ve sloučenině indikují symboly atomů následované stejným číslem, viz např. výše uvedený SMILES pro benzen. Tyto atomy tvoří vazebný pár a cyklus tak uzavírají. Symboly atomů a vazeb uzavřené v kulatých závorkách značí vedlejší větve hlavního řetězce [51]. Příklad jednoduchých SMILES výrazů lze vidět na obr. 3.2.



Obrázek 3.2: Strukturní vzorce kyseliny šťavelové (vlevo) a serotoninu, doplněné o odpovídající SMILES reprezentace.

3.4 SMARTS

SMARTS notace [52, 53] (angl. *SMiles ARbitrary Target Specification*) je odvozena z notace SMILES, kterou rozšiřuje o další funkční prvky. Každý SMILES výraz je validním SMARTS výrazem, opačný přístup však vždy neplatí. Příčinou toho je fakt, že SMARTS

je narozdíl od notace SMILES, která slouží pro popis celých molekul, zaměřena na vyhledávání podstruktur. Validní SMARTS výraz 'cOc', popisující kyslík navázaný na dva aromatické uhlíky (můžeme najít např. v molekule difenyletheru), neodpovídá žádné reálné molekule, a není proto platným SMILES výrazem.

SMARTS rozšiřuje SMILES notaci o užití logických operátorů. Symboly '&' a ';' značí logické AND, přičemž symbol '&' má vyšší prioritu v kombinaci s ostatními logickými operátory. Symbol ',' značí logické OR. Pro negaci výrazů je použit symbol '!'. Užití jmenovaných logických operátorů je ilustrováno v tabulce 3.1 níže.

Dalším specifickým prvkem SMARTS výrazů je konstrukce '$(XY)', kde XY reprezentuje platný SMARTS výraz. Touto konstrukcí lze snadno specifikovat atom v závislosti na jeho chemickém okolí. Příkladem je výraz '[O-;!$([O-]C(=O))]', který definuje kyslíkový anion, jenž zároveň není součástí aniontu karboxylové skupiny COO^-.

[c,n&H1]	aromatický uhlík nebo (aromatický dusík s jedním vodíkem)
[c,n;H1]	(aromatický uhlík nebo aromatický dusík) s jedním vodíkem
[!#6][O][H]	hydroxylová skupina navázaná na jakýkoli atom kromě dusíku
[CX4,c][O][H]	alkoholová skupina navázaná na alifatický či aromatický uhlík
[NX3;!$(NC=O)]	trojvazný dusík, který není součástí amidové skupiny

Tabulka 3.1: Příklad užití logických operátorů v notaci SMARTS. V pravém sloupci jsou popsány vyhledané strukturní motivy.

3.5 RDKit

RDKit [54] je volně dostupná open-source knihovna pro práci s chemickými daty, určena pro jazyky Java a Python. RDKit poskytuje standardní funkce pro zpracování chemických dat, jako je načítání široké škály molekulových formátů, práce s 2D a 3D reprezentací molekul, zápis molekulových reakcí, hledání strukturních motivů molekul a vizualizace výstupů. Kromě jmenovaných funkcí umožňuje propojení s PostgreSQL databází [55]. Základní datové struktury a algoritmy RDKitu jsou implementovány v jazyce C++, což v porovnání s interpretovaným jazykem Python zvyšuje jejich výkonnost.

3.6 MACH

MACH je software pro parametrizaci empirických metod výpočtu parciálních atomových nábojů (viz kap. Parametrizace 2.3.3) vyvinutý v rámci diplomové práce Bc. Ondřeje Schindlera v Národním centru pro výzkum biomolekul v Brně [56].

Software MACH byl vyvinut s cílem optimalizovat parametrizaci empirických metod výpočtu parciálních nábojů za použití optimalizační metody Guided Minimization [37]. Software je vyvinut v jazyce Python za využití specializovaných knihoven pro práci s vědeckými daty (SciPy [57], NLopt [58]). V softwaru byla úspešně implementována parametrizace empirických metod EEM, PEOE, SFKEEM [32], QEq [7], ACKS2 [59] a MGC

[60]. Pro potřeby této bakalářské práce byly využity parametrizace prvních dvou jmenovaných metod. Kromě parametrizace MACH umožňuje výpočet parciálních nábojů molekul pomocí vybrané empirické metody, extrakci informací o vstupním SDF souboru či srovnání obsahu dvou nábojových sad.

Kapitola 4

Implementace

Tato kapitola popisuje implementaci knihovny ATTYC (*ATom TYpe Classification*), která přiřazuje atomům molekulového souboru atomové typy na základě zvolené vlastnosti atomu. Kromě klasifikátoru `peptide`, který je určen výhradně pro formát PDB, lze zbylé klasifikátory libovolně použít pro formát PDB i SDF. Knihovna je implementována v jazyce Python ver. 3.7 a obsahuje následující soubory:

__init__.py obsahuje funkci `classify_atoms(input_file, classifier)`[1] spouštějící klasifikaci atomů molekulové sady za užití zvoleného klasifikátoru

classifier.py definuje rozhraní pro klasifikátory implementované v adresáři **\classifiers**

exceptions.py definuje výjimky (potomky třídy Exception) pro řízení běhu programu

io.py zpracovává všechny vstupy a výstupy (I/O) programu včetně kontroly vstupních argumentů

PDB_atom_types.txt obsahuje názvy atomů aminokyselin definovaných nomenklaturou IUPAC a přiřazené atomové typy pro klasifikaci atomů v PDB souborech

SMARTS_atom_types.txt obsahuje SMARTS výrazy a odpovídající atomové typy pro vyhledání specifických strukturních motivů a funkčních skupin

\classifiers obsahuje klasifikátory, které rozdělují atomy do atomových typů na základě:

__init__.py
hbo.py nejvyššího řádu vazby (*highest bond order*)
hybrid.py hybridizace
partners.py vazebných partnerů
peptide.py pozice atomu v rámci aminokyseliny
substruct.py příslušnosti ke strukturnímu motivu nebo funkční skupině

Výstup knihovny ATTYC je specifikován argumenty `file_output` a `screen_output` funkce `classify_atoms(...)`. Pokud má argument `file_output` hodnotu `True`, jsou přiřazené atomové typy molekul zapsány do textového souboru. V případě použití argumentu `screen_output` je na standardní výstup vypsána statistika přiřazených atomových typů.

[1] Argumenty `file_output` a `screen_output` jsou pro větší přehlednost v textu vynechány.

4.1 Spuštění klasifikace

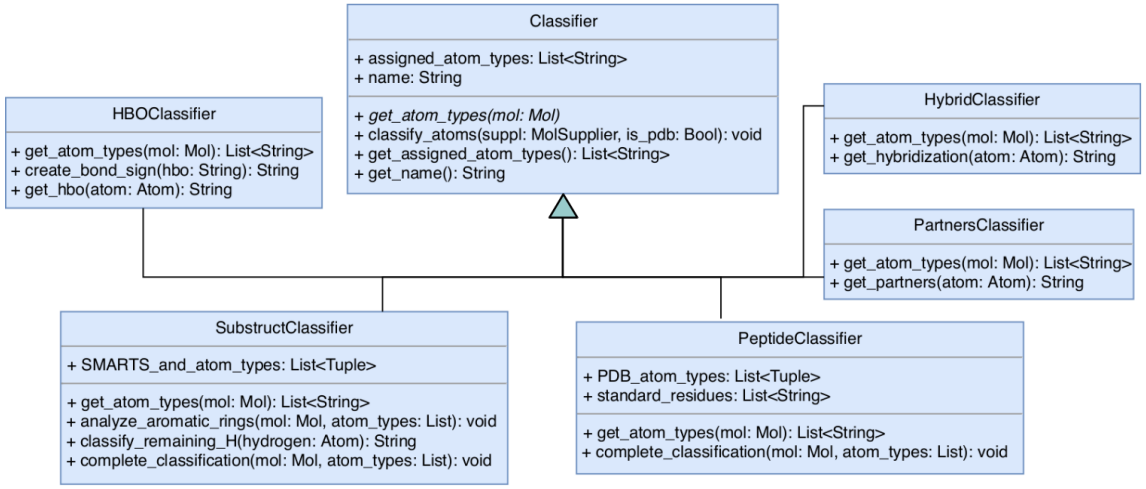
Klasifikace atomů do atomových typů vstupního molekulového souboru je spuštěna voláním funkce `classify_atoms(input_file, classifier)`, která je implementována v modulu **__init__.py** v adresáři **attyc**. Příkaz `attyc.classify_atoms(input_file, classifier)` spouští klasifikaci atomů z libovolného programu v jazyce Python, do kterého byla integrována knihovna ATTYC. Pro spuštění klasifikace je nutné do interpretu Pythonu daného programu nainstalovat knihovnu RDKit.

V rámci funkce `classify_atoms(input_file, classifier)` je kontrolována existence vstupního molekulového souboru a zvoleného klasifikátoru. Pokud je klasifikátor (tzn. podtřída třídy Classifier) implementován v adresáři **\classifiers** a je povolen pro daný molekulový formát, je vytvořena jeho instance. Tato instance přiřazuje atomům v molekulové sadě odpovídající atomový typ.

4.2 Dědičnost třídy Classifier

Třída Classifier definuje rozhraní pro podtřídy definované v modulech v adresáři **\classifiers**. Kromě getterů atributů `assigned_atom_types` a `name` dědí podtřídy metodu `classify_atoms_by_classifier(moleculeset, is_pdb)`. Tato metoda volá abstraktní metodu `get_atom_types(molecule)`, kterou každá podtřída třídy Classifier implementuje. Podtřídy, které představují klasifikátory `substruct` a `peptide`, obsahují kromě výše uvedených atributů také atributy pro uložení dat z externích souborů potřebných pro klasifikaci atomů.

Každý modul v adresáři **\classifiers** obsahuje právě jednu podtřídu třídy Classifier (dále jen 'klasifikátor'). Implementace třídy Classifier je na klasifikátorech definovaných v tomto adresáři nezávislá, moduly v adresáři spolu nijak neinteragují. Knihovna je tak snadno rozšiřitelná o další libovolné klasifikátory.



Obrázek 4.1: UML diagram tříd popisující třídu Classifier a její podtřídy.

4.3 Přiřazení atomových typů

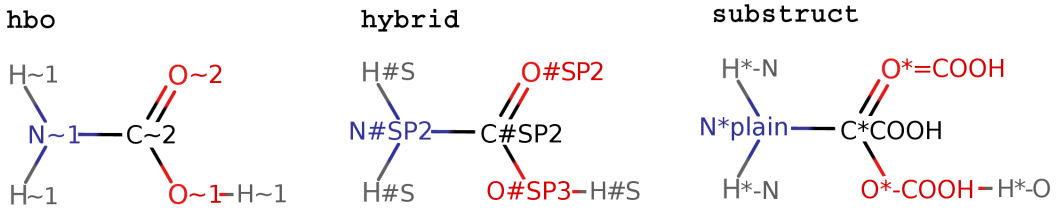
Každý klasifikátor v adresáři **\classifiers** implementuje abstraktní metodu nadtřídy `get_atom_types(molecule)`, která atomům molekuly přiřazuje odpovídající atomové typy. Klasifikátory `hbo`, `hybrid` a `partners` implementují tuto metodu triviálně, neboť využívají příslušné metody tříd Atom a Bond z knihovny RDKit (tab. 4.1).

Pro klasifikátory `substruct` a `peptide` nejsou v třídě Atom v knihovně RDKit implementovány triviální metody, atomy jsou proto těmito klasifikátory klasifikovány pomocí dat z externích souborů. Struktura vytvořených externích souborů a logika klasifikací je popsána v sekcích 4.3.1 a 4.3.2.

`hbo`	Atom.GetBonds(), Bond.GetBondTypeAsDouble()
`hybrid`	Atom.GetHybridization()
`partners`	Atom.GetNeighbors()

Tabulka 4.1: Klíčové metody z knihovny RDKit použité pro přiřazení atomových typů klasifikátory `hbo`, `hybrid` a `partners`.

Následující obrázky ilustrují výstup klasifikace atomů molekulu kyseliny karbamové za použití klasifikátorů `hbo`, `hybrid` a `substruct`.



Obrázek 4.2: Výstupy klasifikací atomů dle klasifikátorů `hbo`, `hybrid` a `substruct`.

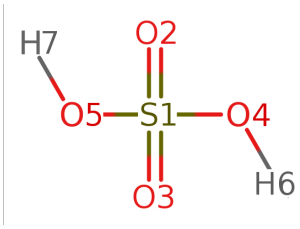
4.3.1 Klasifikátor 'substruct'

Klasifikátor `substruct` klasifikuje atomy na základě příslušnosti k charakteristickým strukturním celkům. Byl implementován s cílem reprodukovat atomové typy, které byly úspěšně použity pro parametrizaci empirických metod výpočtu parciálních atomových nábojů [39, 61].

Strukturní motivy jsou v molekulách detekovány pomocí výrazů SMARTS (kap. 3.4) metodou `GetSubstructMatches(SMARTS_pattern)`[2] třídy Mol z knihovny RDKit. Příkaz `molecule.GetSubstructMatches(SMARTS_pattern)` vrací n-tici (*tuple*) n-tic, jež obsahují prvky typu integer. Tato čísla označují atomy molekuly vyhovující danému

[2]Metoda `GetSubstructMatches(Chem.MolFromSmarts(SMARTS_pattern))` je v textu pro názornost syntakticky zjednodušena.

SMARTS dotazu. Obrázek 4.3 a kód níže demonstrují užití SMARTS výrazu k vyhledání sulfonové skupiny $S(=O)_2OH$ nebo jejích aniontů.



Obrázek 4.3: Molekula kyseliny sírové s číselným označením atomů.

```
>>> SMARTS_pattern = "[SX4](=[OX1])(=[OX1])[OX2,OX1-]"
>>> pattern_atms = molecule.GetSubstructMatches(SMARTS_pattern)
>>> for atom_tuple in pattern_atms:
...     print(atom_tuple, get_elements(atom_tuple))
>>> print(pattern_atms)

Python 3.7.0 (default, Apr 9 2019, 10:31:47)
(1, 2, 3, 4) ('S', 'O', 'O', 'O')
(1, 2, 3, 5) ('S', 'O', 'O', 'O')
((1, 2, 3, 4), (1, 2, 3, 5))
```

Obrázek 4.4: Ukázka kódu klasifikátoru `substruct`. Čísla v n-tici odpovídají atomům molekuly na obrázku 4.3.

Pro klasifikaci atomů je klíčovým souborem **SMARTS_atom_types.txt**. Soubor obsahuje SMARTS výrazy, pomocí nichž jsou v molekule hledány strukturní motivy, a atomové typy, které jsou atomům strukturního motivu explicitně přiřazeny. Pořadí detekce strukturních motivů odpovídá pořadí SMARTS vstupů v souboru. Některé atomy molekuly mohou příslušet více strukturním motivům, které SMARTS vstupy detekují. Uspořádání SMARTS vstupů v souboru tak ovlivňuje výsledný atomový typ, který je atomu přiřazen. Atomové typy přiřazené jednotlivým atomům lze v souboru předefinovat, stejně jako změnit pořadí SMARTS výrazů, a upravit tak logiku klasifikace a optimalizovat výsledky parametrizace.

```
[CX3](=O)[OX2][H1]      COOH,=COOH,-COOH,-O     carboxylic acid
[CX4,c][OX2][H]         -O,OH,-O                alcohol group
[NX3+](=O)[O-]          nitro,nitro,nitro       nitro group
[C,c]=[OX1]             carbonyl,carbonyl       carbonyl group
[NX1]#[CX2]             nitrile,nitrile         nitrile group
[c,C]=[NX2][H]          =N,=C,-N                imine group
```

Obrázek 4.5: Struktura souboru **SMARTS_atom_types.txt**. Prostřední sloupec definuje atomové typy atomů dané funkční skupiny.

SMARTS výrazy v souboru **SMARTS_atom_types.txt** byly převzaty ze stránek společnosti Daylight Chemical Information System, Inc. [53], pro účely klasifikace atomů však musely být ve většině případů dodatečně upraveny. Vybrané SMARTS výrazy byly rozšířeny o detekci většího počtu atomů nebo byly upraveny tak, aby se výsledky jednotlivých SMARTS dotazů nepřekrývaly.

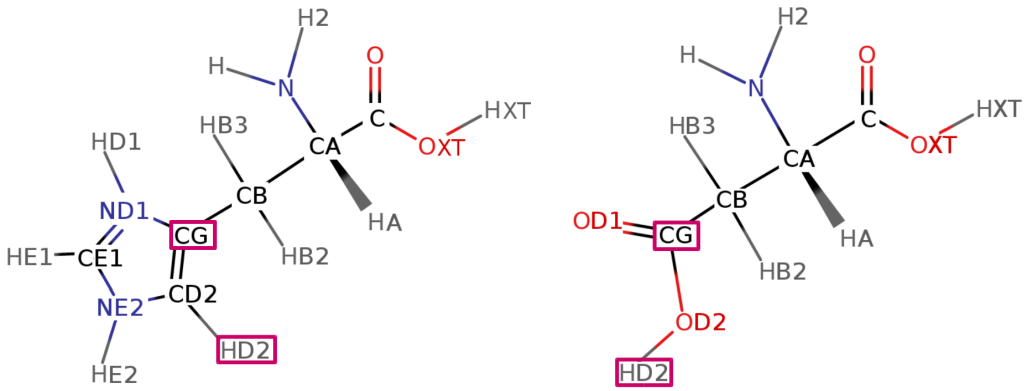
Úpravu pro snížení redundance výsledků ilustruje následující příklad. Vyhledání karboxylové skupiny za užití výrazu `[CX3](=O)[OX2][H]` je v externím souboru následováno detekcí alkoholové skupiny pomocí výrazu `[CX4,c][OX2][H]`. Pokud by nebyla specifikována vaznost uhlíku (’CX4’), vyhledal by daný SMARTS výraz i OH skupiny, které by byly součástí dříve určených karboxylových skupin.

Rozšíření SMARTS dotazů o detekci více atomů se ve většině případů týkalo detekce vodíků. Pro vyhledání primárních aminů je v uvedeném online zdroji specifikován výraz `[NX3;H2!$(NC=O)]`. Tento výraz byl upraven na dotaz `[NX3;!\$(NC=O)]([H])[H]`, kterým jsou narozdíl od původního výrazu vodíky již detekovány.

4.3.2 Klasifikátor ’peptide’

Klasikace atomů peptidových řetězců přiřazuje atomové typy obdobně na základě příslušnosti atomů ke strukturním celkům. Logika vstupního souboru klasifikátoru `peptide` se od struktury externího souboru klasifikátoru `substruct` liší, neboť nepracuje s vyhledáváním strukturních motivů pomocí SMARTS výrazů.

Symboly atomových typů aminokyselin popsané nomenklaturou IUPAC (CA, CB, HXT, OXT,...) popisují atomy všech proteinogenních residuí. Jednotlivé atomové typy však nijak nereflektují vlastnosti atomů, které popisují, chemická okolí atomů označených stejným atomovým typem se tak často liší (srov. obr. 4.6). Pro implementaci klasifikátoru `peptide` tak bylo klíčové provést rešerši existujících atomových typů aminokyselin a každé dvojici ’atomový typ nomenklatury IUPAC - aminokyselina’ přiřadit atomový typ reflektující chemické okolí daného atomu. Navržená klasifikace atomů aminokyselin byla implementována s cílem reprodukovat atomové typy, které byly použity pro úspěšnou parametrizaci empirických metod [30, 38].



Obrázek 4.6: Strukturní vzorce histidinu a kyseliny asparagové s atomovými typy nomenklatury IUPAC verze 3.0. Atomy CG a HD2 jsou v molekulách součástí odlišných strukturních celků, je tedy vhodné každému z nich přiřadit jiný atomový typ.

Kapitola 5

Výsledky a diskuse

Kapitola shrnuje a interpretuje výsledky parametrizací metod EEM a PEOE za užití navržených klasifikací atomů s cílem zhodnotit, zda jsou pro parametrizaci postačující základní klasifikace atomů nebo je pro tento problém vhodné aplikovat detailnější dělení. Veškeré výsledky parametrizací lze nalézt v externí příloze bakalářské práce nebo na adrese `https://lcc.ncbr.muni.cz/~raduse19/`.

5.1 Vstupní data

Pro parametrizaci empirických metod byla použita molekulová sada CCD_gen a sada Protein (viz tab. 5.1). Pro analýzu přenositelnosti parametrů byly molekuly v molekulových sadách rozděleny na sadu tréninkovou a validační v poměru 9:1. Software MACH zajistil, aby validační sada vždy obsahovala identické atomové typy jako sada tréninková. Empirické metody byly pro obě molekulové sady parametrizovány vůči referenčním nábojům typu B3LYP/6-311G*/NPA.

sada molekul	**CCD_gen** [62]	**Protein**
počet molekul	4 443	32
počet atomů	204 760	29 107
počet atomů v molekule	3-305	166-1 174
zdroj struktur	software CORINA	RTG krystalografie
molekuly	ligandy z databáze PDB pouze validní struktury	malé proteiny neobsahují ligandy ani nestandardní residua

Tabulka 5.1: Souhrn informací o molekulových sadách užitých pro parametrizaci. V obou sadách byly doplněny atomy vodíku a byla optimalizována geometrie molekul.

5.2 Výsledky parametrizace

Pro popis výsledků parametrizací empirických metod za užití různých klasifikátorů je zavedeno značení 'molekulová sada/empirická metoda/užitý klasifikátor'. Výpočty parametrizací probíhaly na serverech virtuální organizace MetaCentrum.

Triviální referenční klasifikace `plain`, vůči které jsou implementované klasifikátory porovnávány, rozděluje atomy do atomových typů na základě hodnoty protonového čísla. Klasifikátory `hbo`, `hybrid`, `substruct` a `peptide` se po vyhodnocení statistik a korelačních grafů prokázaly pro parametrizaci metod EEM a PEOE jako platné. Klasifikátor `partners` generuje pro obsáhlé molekulové sady extrémní množství atomových typů (pro sadu CCD_gen konkrétně 235), jeho použití proto nelze obecně doporučit.

PCC^2 získaných empirických a referenčních nábojů nabývá napříč parametrizacemi minimální hodnoty 0,9768 pro CCD_gen/EEM/plain parametrizaci, hodnota RMSD je maximálně 0,0641, a to u téže parametrizace. Statistiky parametrizací sady CCD_gen lze nalézt v následující tabulce 5.2. Statistiky parametrizací sady Protein jsou umístěny v příloze B.

klasifikátor	***n***	**metoda**	**RMSD**	**PCC^2**	**MAE**	**ABSMAX**	**doba výpočtu**
`plain`	9	EEM	0,0641	0,9768	0,0440	0,7455	1:01:08
		PEOE	0,0516	0,9849	0,0343	0,5228	0:55:06
`hbo`	19	EEM	0,0582	0,9809	0,0405	0,7133	1:59:29
		PEOE	0,0394	0,9912	0,0259	0,7313	2:53:04
`hybrid`	15	EEM	0,0620	0,9781	0,0424	0,7475	1:15:40
		PEOE	0,0479	0,9870	0,0310	1,0619	1:42:28
`substruct`	64	EEM	0,0479	0,9870	0,0300	0,6593	17:46:49
		PEOE	0,0464	0,9878	0,0293	0,9175	6:56:11

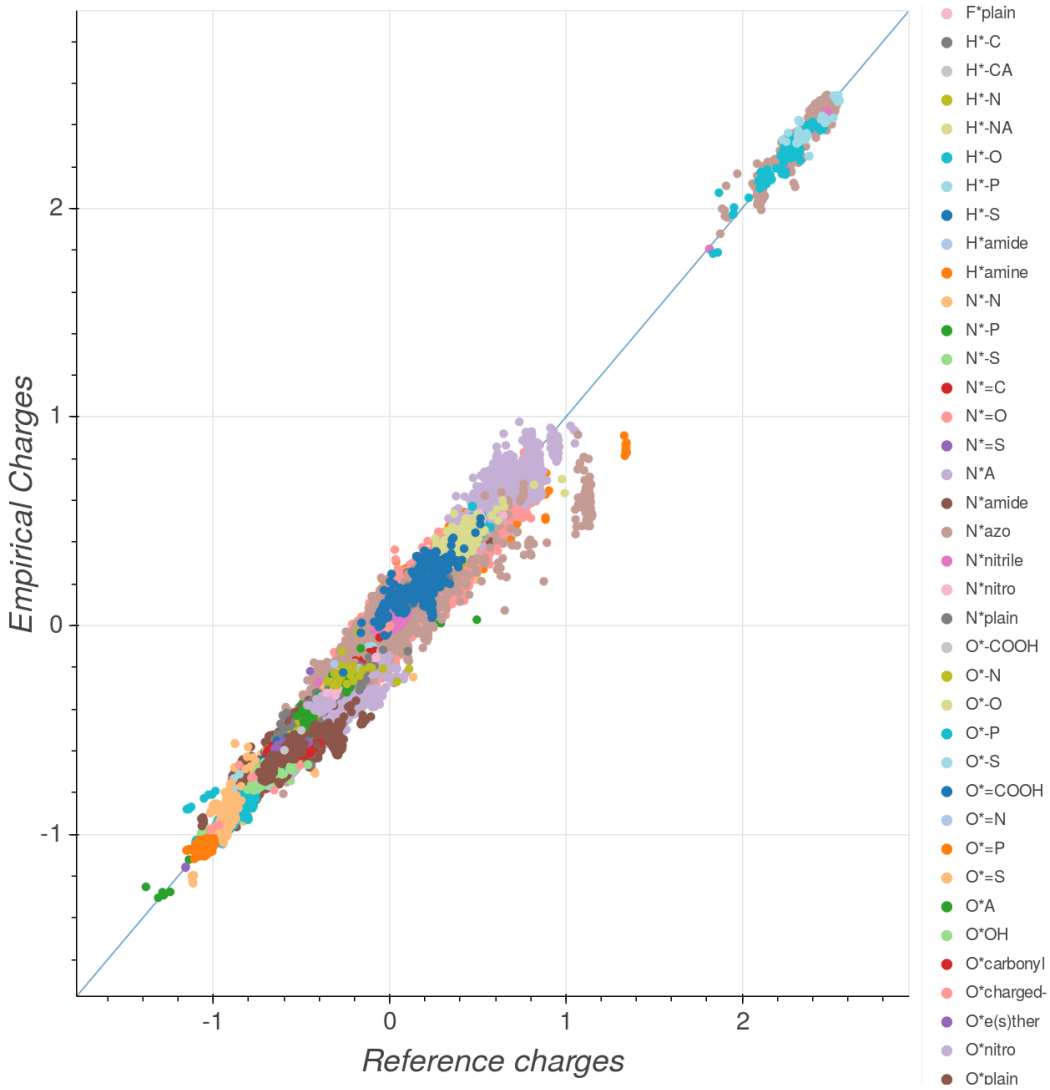
Tabulka 5.2: Výsledky vybraných statistik parametrizace CCD_gen sady, *n* značí počet užitých atomových typů. Jsou zaznamenány pouze statistiky tréninkových sad.

Hodnota PCC^2 napříč parametrizacemi za užití obou molekulových sad neklesá pod hodnotu 0,97; lze tedy vyvodit, že náboje atomů získané skrze vypočtené parametry silně lineárně korelují s referenčními náboji (viz obr. 5.1). Hodnoty RMSD pro vypočtené empirické a referenční náboje se pohybují nejčastěji v rozmezí 0,040-0,060; u PDB sady v některých případech klesají až k hodnotě 0,023. Uvedená rozmezí hodnot statistik platí jak pro tréninkové, tak validační sady molekul.

Rozdíl hodnot RMSD tréninkových a validačních sad je nejčastěji v řádu desetitisícin, maximální hodnota tohoto rozdílu je 0,018. Hodnot PCC^2 zmíněných sad se liší v řádu tisícin či desetitisícin, pro MAE a ABSMAX rozdíl proniká až do řádu setin. Na základě těchto údajů lze parametry získané užitím implementovaných klasifikátorů pokládat za přenositelné. Hodnoty statistik tréninkové a validační sady klasifikátoru `hbo` ilustruje tabulka 5.3.

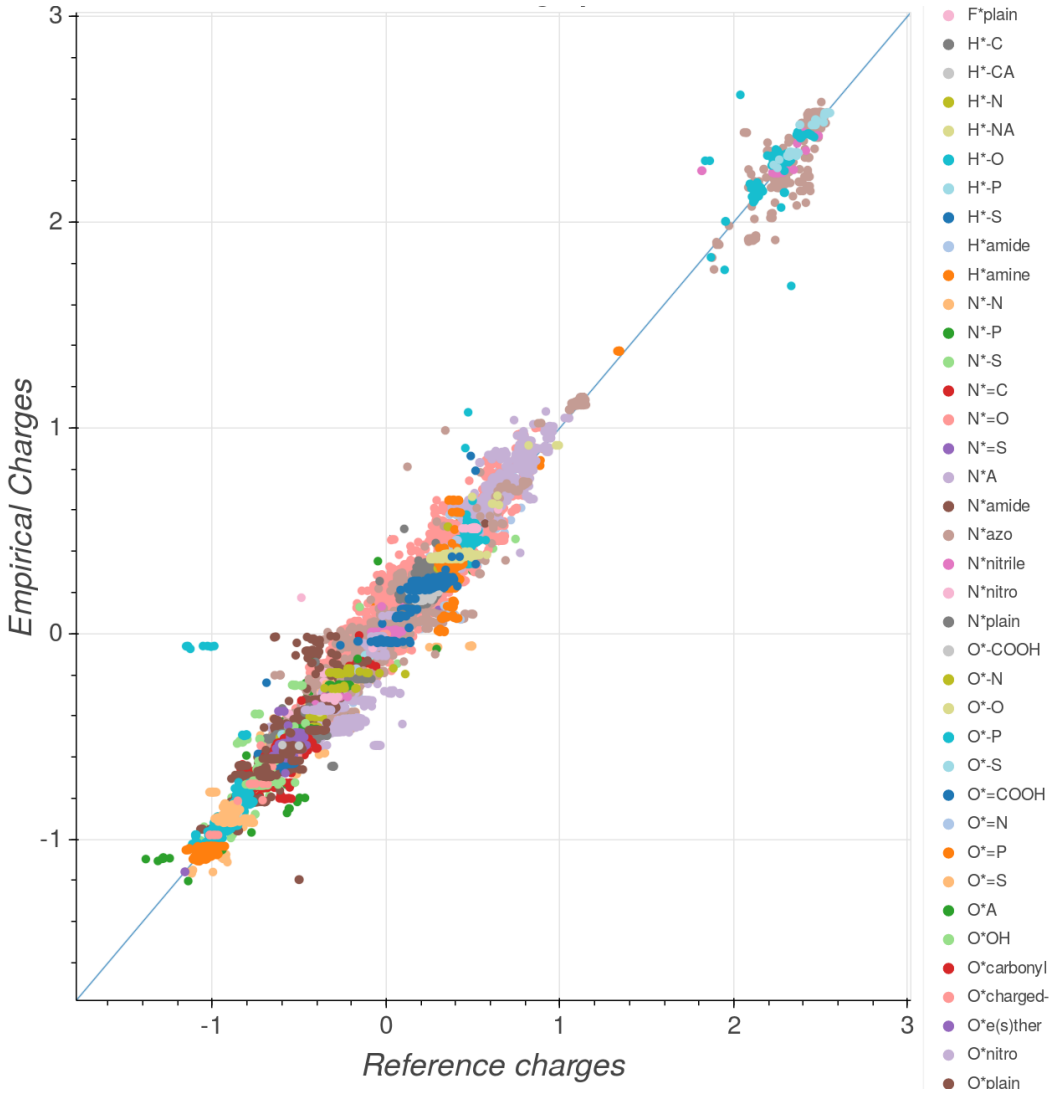
metoda	**sada**	**RMSD**	**PCC^2**	**MAE**	**ABSMAX**
EEM	train	0,0582	0,9809	0.0405	0,7133
	test	0,0566	0,9819	0,0397	0,7045
PEOE	train	0,0394	0,9912	0.0259	0,7313
	test	0,0386	0,9916	0,0258	0,4765

Tabulka 5.3: Porovnání statistik tréninkových a validačních sad klasifikátoru `hbo`. Statistiky vykazují pro tréninkové a validační sady minimální rozdíly. Uvedené hodnoty se vztahují ke parametrizaci sady CCD_gen.



Obrázek 5.1: Korelační graf parametrizace CCD_gen/EEM/substruct. PCC^2=0,9870; RMSD=0,0479. Legenda ve skutečnosti obsahuje 64 atomových typů, v obrázku je oříznuta na velikost grafu.

Z korelačních grafů empirických a referenčních nábojů je patrné, že metoda PEOE přiřazuje v určitých případech atomům náboje o konstatní hodnotě, což v grafu odpovídá vodorovným shlukům bodů (viz obr. 5.2 nebo obr. C.2 v příloze). Přiřazení konstaních či velmi podobných hodnot parciálních nábojů vyplývá z iterativního způsobu výpočtu. Atomu je iniciálně přiřazena hodnota jeho formálního náboje. V každé iteraci výpočtu dochází k přesunu elektronové hustoty mezi vazebnými partnery, přičemž se toto množství exponenciálně snižuje, a to až do ustálení výsledné hodnoty parciálního náboje. Metoda PEOE navíc pracuje pouze s topologií molekuly a atomům, které mají stejné chemické okolí, implicitně přiřazuje stejné náboje.



Obrázek 5.2: Korelační graf parametrizace CCD_gen/PEOE/substruct. PCC^2=0,9878; RMSD=0,0464. Legenda ve skutečnosti obsahuje 64 atomových typů, v obrázku je oříznuta na velikost grafu.

5.2.1 Úprava klasifikátorů 'substruct' a 'peptide'

Časová náročnost parametrizací za použití klasifikátorů `substruct` a `peptide` se v porovnání s ostatními klasifikátory ukázala jako netriviální z důvodu velkého množství definovaných atomových typů. Po analýze korelačních grafů byly vybrané atomové typy sloučeny, nejčastěji na základě nejvyššího řádu vazby daného atomu (atomové typy popisující jednovaznou síru navázanou na dusík, fosfor a síru byly sloučeny pod atomový typ 'S*1bond'), v případě vodíků byly sloučeny atomové typy popisující vodíky navázané na atomy o stejném protonovém čísle. Počet atomových typů byl u obou klasifikátorů zredukován přibližně na polovinu. Korelační grafy parametrizací za užití redukovaného klasifikátoru `peptide` jsou zobrazeny v příloze C.

klasifikátor	metoda	RMSD	PCC^2	MAE	ABSMAX	doba výpočtu
`substruct`	EEM	0,0334	0,9948	0,0229	0,3880	23:01:14
	PEOE	0,0493	0,9887	0,0292	0,4484	0:19:51
`substruct simplified`	EEM	0,0388	0,9930	0,0265	0,3198	14:30:45
	PEOE	0,0514	0,9876	0,0319	0,5083	0:13:02
`peptide`	EEM	0,0359	0,9940	0,0228	0,3558	1 den, 15:03:35
	PEOE	0,0456	0,9902	0,0293	0,3147	0:29:01
`peptide simplified`	EEM	0,0389	0,9929	0,0266	0,3647	18:28:26
	PEOE	0,0512	0,9877	0,0302	0,2536	0:15:12

Tabulka 5.4: Srovnání statistik původních a zjednodušených klasifikátorů `peptide` a `substruct` aplikovaných na sadě Protein. V tabulce jsou zaznamenány pouze statistiky tréninkových sad.

Statistiky zjednodušených klasifikací se od statistik původních klasifikátorů liší pouze minimálně, rozdíly hodnot jednotlivých statistik se týkají řádu tisícin, ve výjimečných případech setin. Zjednodušením klasifikace se docílilo snížení časové náročnosti parametrizace zhruba na polovinu, a to s minimálním vlivem na kvalitu vypočtených parametrů.

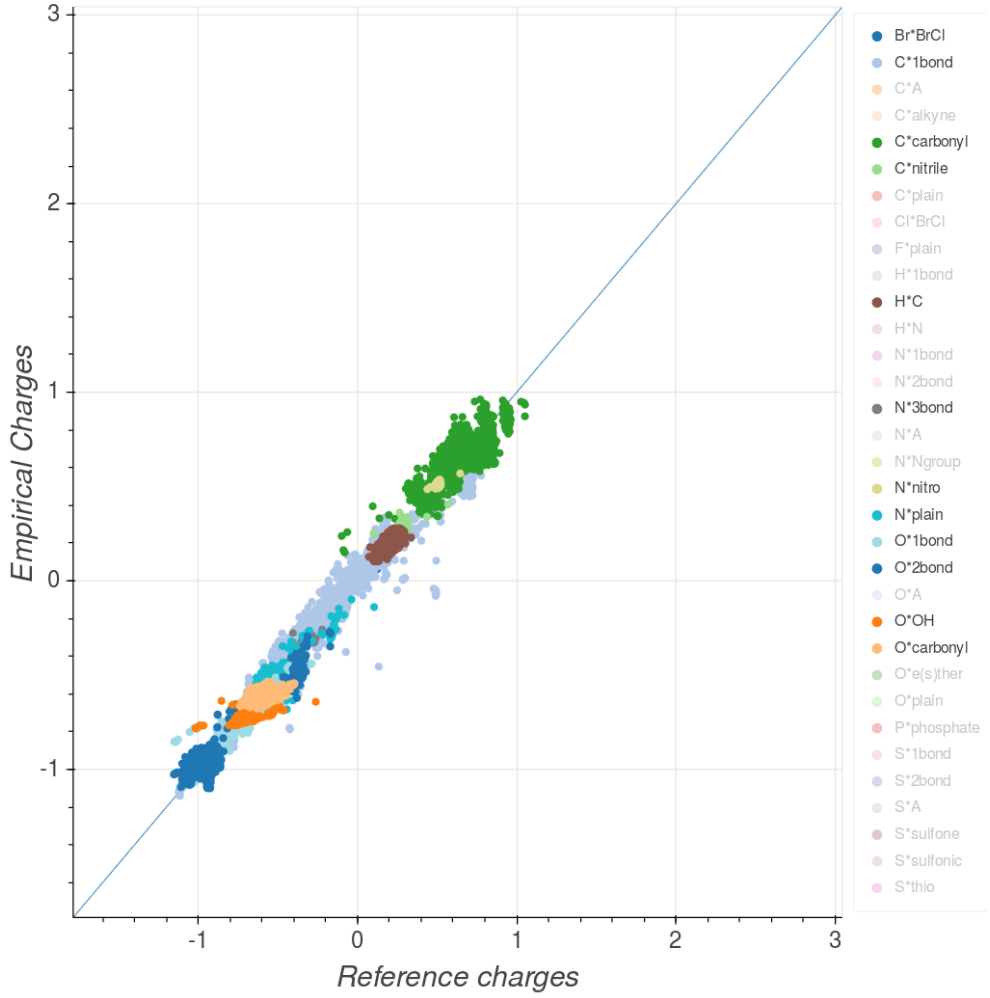
5.2.2 Statistiky vybraných atomových typů

Vysoká hodnota PCC^2 molekulové sady popisuje míru lineární závislosti hodnot empirických a referenčních nábojů, nevypovídá však o lineární korelaci mezi empirickými a referenčními náboji pro jednotlivé atomové typy.

Pro vybrané atomové typy nabývá PCC^2 hodnot blízkých nule, mezi atomy, které jsou daným atomovým typem klasifikované, je tedy detekována nulová lineární závisost mezi empirickými a referenčními náboji. Nízká hodnota kvadrátu Pearsonova korelačního koeficientu vybraného atomového typu může být způsobená přiřazenou konstantní hodnotou empirických nábojů nebo malým počtem atomů, který atomový typ reprezentují. I minimální odchylky nábojů od přímky $y = x$ korelačního grafu se projeví

do výsledné hodnoty PCC^2 atomového typu, zejména v případě, kdy je rozsah hodnot referenčních nábojů velmi malý.

Nízká hodnota PCC^2 atomového typu nepředstavuje problém, pokud se hodnoty empirických a referenčních nábojů liší pouze minimálně. Uvedené tvrzení platí pro atomové typy 'H*C', 'O*carbonyl' a 'O*OH' v obr. 5.3. Statistiky vybraných atomových typů lze nalézt v [příloze].



Obrázek 5.3: Korelační graf parametrizace CCD_gen/EEM/substruct_simplified. PCC^2 pro atomové typy 'H*C', 'O*carbonyl' a 'O*OH' nabývá hodnot 0,3702; 0,1501 a 0,3050.

Ze šesti klasifikátorů implementovaných v knihovně ATTYC se pět klasifikací atomů prokázalo platnými pro parametrizaci empirických metod EEM a PEOE. Nezanedbatelné rozdíly byly napříč klasifikátory nalezeny pouze v době výpočtu jednotlivých parametrizací. Analýza parametrizací za užití implementovaných klasifikátorů ve srovnání s referenční klasifikací prokázala, že základní klasifikace atomů využívající malý počet atomových typů jsou pro potřeby parametrizace plně dostačující.

Kapitola 6

Závěr

Parametrizace empirických metod pro výpočet parciálních atomových nábojů je netriviální výpočetní proces, který je ovlivněn řadou faktorů, mimo jiné množstvím definovaných atomových typů molekulové sady. S počtem atomových typů roste počet hledaných empirických parametrů, což ovlivňuje výpočetní náročnost a přesnost parametrizace.

Cílem této bakalářské práce bylo otestovat, jak jemné či hrubé klasifikace atomových typů postačují pro parametrizaci empirických metod poskytujících ve srovnání s kvantově-chemickými metodami kvalitní výsledky. V souvislosti s parametrizací bylo také testováno, zda jsou některé charakteristiky atomu pro definici atomových typů vhodnější než charakteristiky jiné.

Výsledky parametrizací za užití implementovaných klasifikátorů byly srovnány s referenční klasifikací, která atomům přiřazuje atomové typy na základě jejich protonového čísla. Kromě klasifikátoru `partners` byly implementované klasifikátory úspěšně použity pro parametrizace empirických metod EEM a PEOE, a to u obou molekulových sad. Statistiky vykazují pro úspěšně implementované klasifikátory vysokou korelaci empiricky vypočtených a referenčních nábojů s odchylkami v tolerovaných hodnotách, hodnoty těchto statistik jsou napříč klasifikátory pro obě empirické metody velmi podobné.

Markantní rozdíly byly pro klasifikátory nalezeny v době výpočtu jednotlivých parametrizací. Klasifikace používající jemné dělení atomových typů se ukázaly zhruba 8x více časové náročné než klasifikace triviální, statistiky empirických a referenčních nábojů zůstaly vůči méně detailním klasifikacím téměř nezměněny. Redukce počtu atomových typů detailních klasifikací, následná parametrizace za použití redukovaného počtu atomových typů a porovnání statistik obou přístupů pozorování potvrdily.

Klasifikace atomů dle nejvyššího řádu vazby, hybridizace a hodnoty protonového čísla atomu se pro potřeby parametrizace empirických metod EEM a PEOE prokázaly jako plně dostačující. Ve srovnání s detailními klasifikacemi poskytují kvalitní výsledky, a to s menší výpočetní náročností.

Seznam použité literatury

[1] ATKINS, P. W. & DE PAULA, J. *Atkins' physical chemistry*. 9th ed. Oxford: Oxford University Press, 2010. ISBN 978-0-19-954337-3.

[2] JEAN, Y., VOLATRON, F. & BURDETT, J. K. *An introduction to molecular orbitals*. New York: Oxford University Press, 1993. ISBN 0-19-506918-8.

[3] MORTIER, W. J., GHOSH, S. K. & SHANKAR, S. „Electronegativity-equalization method for the calculation of atomic charges in molecules". *Journal of the American Chemical Society*. 1986, **108**(15), 4315-4320. DOI: 10.1021/ja00275a013.

[4] GASTEIGER, J. & MARSILI, M. „Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges". *Tetrahedron*. 1980, **36**(22), 3219-3228. DOI: 10.1016/0040-4020(80)80168-2.

[5] LEACH, A. R. *Molecular modelling: principles and applications*. 2nd ed. New York: Prentice Hall, 2001. ISBN 0-582-38210-6.

[6] GASTEIGER, J. & ENGEL, T. *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH, c2003. ISBN 978-3-527-30681-7.

[7] RAPPE, A. K. & GODDARD, W. A. „Charge equilibration for molecular dynamics simulations". *The Journal of Physical Chemistry*. 1991, **95**(8), 3358-3363. DOI: 10.1021/j100161a070.

[8] LYNE, P. D. „Structure-based virtual screening: an overview". *Drug Discovery Today*. 2002, **7**(20), 1047-1055. DOI: 10.1016/S1359-6446(02)02483-2.

[9] KARELSON, M., LOBANOV, V. S. & KATRITZKY, A. R. „Quantum-Chemical Descriptors in QSAR/QSPR studies". *Chemical Reviews*. 1996, **96**(3), 1027-1044. DOI: 10.1021/cr950202r.

[10] GHAFOURIAN, T. & DEARDEN, J. C. „The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods". *Journal of Pharmacy and Pharmacology*. 2000, **52**(6), 603-610. DOI: 10.1211/0022357001774435.

[11] VAŘEKOVÁ, R. S., GEIDL, S., IONESCU, C.-M., SKŘEHOTA, O., BOUCHAL, T., SEHNAL, D., ABAGYAN, R. & KOČA, J. „Predicting pK a values from EEM atomic charges". *Journal of Cheminformatics*. 2013, **5**(1). DOI: 10.1186/1758-2946-5-18.

[12] WANG, B., LI, S. L. & TRUHLAR. D. G. „Modeling the Partial Atomic Charges in Inorganometallic Molecules and Solids and Charge Redistribution in Lithium-Ion Cathodes“. *Journal of Chemical Theory and Computation*. 2014, **10**(12), 5640-5650. DOI: 10.1021/ct500790p.

[13] CELÝ, J. *Základy kvantové mechaniky pro chemiky: I. Principy*. Brno: Rektorát UJEP Brno, 1986.

[14] HÖLTJE, H.-D. & FOLKERS, G. *Molecular Modeling: Basic Principles and Applications*. Volume 5. Weinheim: Wiley-VCH, 2008. ISBN 978-3-527-61476-9.

[15] CURTISS, L. A., REDFERN, P. C. & FRURIP, D. J. „Theoretical Methods for Computing Enthalpies of Formation of Gaseous Compounds“. LIPKOWITZ, K. B. & BOYD, D. B., ed. *Reviews in Computational Chemistry*. Hoboken, NJ, USA: John Wiley & Sons, 2000, s. 147-211. *Reviews in Computational Chemistry*. DOI: 10.1002/9780470125922.ch3.

[16] PILAR, F. L. *Elementary quantum chemistry*. Dover ed. Mineola, N.Y.: Dover Publications, 2001. ISBN 0-486-41464-7.

[17] „Hierarchy of *ab initio* Post-HF metods“. Přetisknuto s laskavým svolením Prof. Martina Kauppa, TU Berlín. *In* SEMRÁD, H. „Studium mechanismu bromoborační reakce metodami kvantové chemie“. Brno: Masarykova univerzita, Přírodovědecká fakulta, Brno, 2016 [online]. [cit. 2019-04-25]. URL: `https://is.muni.cz/th/375827/prif_m/`

[18] CRAMER, C. J. *Essentials of Computational Chemistry: Theories and Models*. Chichester: John Wiley, 2002. ISBN 0-471-48552-7.

[19] LEVINE, I. N. *Quantum Chemistry*. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2000. CHYBÍ ISBN

[20] SEFZIK, T. H., TURCO, D., IULIUCCI, R. J. & FACELLI, J. C. „Modeling NMR Chemical Shift: A Survey of Density Functional Theory Approaches for Calculating Tensor Properties“. *The Journal of Physical Chemistry*. 2005, **109**(6), 1180-1187. CHYBÍ DOI

[21] KOCH, W. & HOLTHAUSEN, M. C. *A chemist's guide to density functional theory*. New York: Wiley-VCH, 2000. ISBN 3527299181.

[22] POPLE, J. A. & SEGAL, G. A. „Approximate Self-Consistent Molecular Orbital Theory. II. Calculations with Complete Neglect of Differential Overlap“. *The Journal of Chemical Physics*. 1965, **43**(10), 136-151. DOI: 10.1063/1.1701476.

[23] POPLE, J. A., BEVERIDGE, D. L. & DOBOSH, P. A. „Approximate Self-Consistent Molecular-Orbital Theory. V. Intermediate Neglect of Differential Overlap“. *The Journal of Chemical Physics*. 1967, **47**(6), 2026-2033. DOI: 10.1063/1.1712233.

[24] DEWAR, M. J. S. & THIEL, W. „Ground states of molecules. 38. The MNDO method. Approximations and parameters“. *Journal of the American Chemical Society*. 1977, **99**(15), 4899-4907. DOI: 10.1021/ja00457a004.

[25] DAVIDSON, E. R. & FELLER. D. „Basis set selection for molecular calculations“. *Chemical Reviews*. 1986, **86**(4), 681-696. DOI: 10.1021/cr00074a002.

[26] MULLIKEN, R. S. „Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I“. *The Journal of Chemical Physics*. 1955, **23**(10), 1833-1840. DOI: 10.1063/1.1740588.

[27] REED, A. E., WEINSTOCK, R. B. & WEINHOLD, F. „Natural population analysis“. *The Journal of Chemical Physics*. 1985, **83**(2), 735-746. DOI: 10.1063/1.449486.

[28] BADER, R. F. W. *Atoms in molecules: a quantum theory*. New York: Oxford University Press, 1994. ISBN 978-0198558651.

[29] PARK, J. M., NO, K. T., JHON, M. S. & SCHERAGA, H. A. „Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. III. Application to halogenated and aromatic molecules“. *Journal of Computational Chemistry*. 1993, **14**(12), 1482–1490. DOI:10.1002/jcc.540141210.

[30] CHO, K.-H., KANG, Y. K., NO, K. T. & SCHERAGA, H. A. „A Fast Method for Calculating Geometry-Dependent Net Atomic Charges for Polypeptides“. *The Journal of Physical Chemistry B*. 2001, **105**(17), 3624-3634. DOI: 10.1021/jp0023213.

[31] GASTEIGER, J. & SURYANARAYANA, I. „A quantitative empirical treatment of 13C NMR chemical shift variations on successive substitution of methane by halogen atoms“. *Magnetic Resonance in Chemistry*. 1985, **23**(3), 156-157. DOI: 10.1002/mrc.1260230304.

[32] CHAVES, J., BARROSO, J. M., BULTINCK, P. & CARBÓ-DORCA, R. „Toward an Alternative Hardness Kernel Matrix Structure in the Electronegativity Equalization Method (EEM)“. *Journal of Chemical Information and Modeling*. 2006, **46**(4), 1657-1665. DOI: 10.1021/ci050505e.

[33] YANG, Z.-Z. & WANG, C.-S. „AtomBond Electronegativity Equalization Method. 1. Calculation of the Charge Distribution in Large Molecules“. *The Journal of Physical Chemistry A*. 1997, **101**(35), 6315-6321. DOI: 10.1021/jp9711048.

[34] WANG, C.-S. & YANG, Z. Z. „Atom–Bond Electronegativity Equalization Method. II. Lone-pair electron model“. *The Journal of Chemical Physics*. 1999, **110**(13), 6189-6197. DOI: 10.1063/1.478524.

[35] HEIDLER, R., JANSSENS, G. O. A., MORTIER, W. J. & SCHOONHEYDT, R. A. „Charge Sensitivity Analysis of Intrinsic Basicity of Faujasite-Type Zeolites Using the Electronegativity Equalization Method (EEM)“. *The Journal of Physical Chemistry*. 1996, **100**(50), 19728-19734. DOI: 10.1021/jp9615619.

[36] IONESCU, C.-M., GEIDL, S., VAŘEKOVÁ, R. S. & KOČA, J. „Rapid Calculation of Accurate Atomic Charges for Proteins via the Electronegativity Equalization Method“. *Journal of Chemical Information and Modeling*. 2013, **53**(10), 2548-2558. DOI: 10.1021/ci400448n.

[37] PAZÚRIKOVÁ, J., KŘENEK, A. & MATYSKA, L. „Guided Optimization Method for Fast and Accurate Atomic Charges Computation“. In ÉVORA-GÓMEZ, J. & HERNANDÉZ-CABRERA, J. J. *Proceedings of the 2016 European Simulation and Modelling Conference*. Ghent: EUROSIS - ETI, 2016. 267-274. ISBN 978-90-77381-95-3.

[38] KANG, Y. K. & SCHERAGA, H. A. „An Efficient Method for Calculating Atomic Charges of Peptides and Proteins from Electronic Populations“. *The Journal of Physical Chemistry B*. 2008, **112**(17), 5470-5478. DOI: 10.1021/jp711484f.

[39] YAKOVENKO, O., OLIFERENKO, A. A., BDZHOLA, V. G., PALYULIN, V. A. & ZEFIROV, N. S. „Kirchhoff atomic charges fitted to multipole moments: Implementation for a virtual screening system“. *Journal of Computational Chemistry*. 2008, **29**(8), 1332-1343. DOI: 10.1002/jcc.20892.

[40] DODGE, Y. *The Oxford Dictionary of Statistical Terms*. Oxford: Oxford University Press, 2003. ISBN 978-0199206131.

[41] MAIOROV, V. N. & CRIPPEN, G. M. „Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins“. *Journal of Molecular Biology*. 1994, **235**(2), 625-634. DOI: 10.1006/jmbi.1994.1017.

[42] PAVLÍK, T. & DUŠEK, L. *Biostatistika*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-782-6.

[43] *CTFile Formats* [online]. [cit. 2019-04-18]. URL: `http://c4.cabrillo.edu/404/ctfile.pdf`

[44] DALBY, A., NOURSE, J. G., HOUNSHELL, W. D., GUSHURST, A. K. I., GRIER, D. L., LELAND, B. A. & LAUFER, J. „Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited“. *Journal of Chemical Information and Modeling*. 1992, **32**(3), 244-255. DOI: 10.1021/ci00007a012.

[45] Worldwide Protein Data Bank: *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description, Version 3.30* [online]. [cit. 2019-04-30]. URL: `ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_Letter.pdf`

[46] UCSF: Resource for Biocomputing, Visualization, and Informatics: *Introduction to Protein Data Bank Format* [online]. [cit. 2019-04-30]. URL: `https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html`

[47] MARKLEY, J. L., BAX, A., ARATA, Y., HILBERS, C. W., KAPTEIN, R., SYKES, B. D., WRIGHT, P. E. & WUTHRICH, K. „Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy“. *European Journal of Biochemistry*. 1998, **256**(1), 1-15. DOI: 10.1046/j.1432-1327.1998.2560001.x.

[48] WEININGER, D. „SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules“. *Journal of Chemical Information and Modeling*. 1988, **28**(1), 31-36. DOI: 10.1021/ci00057a005.

[49] BUNIN, B. A. *Chemoinformatics: Theory, Practice, & Products*. Dordrecht: Springer, 2007. ISBN 987-1-4020-5000-8.

[50] LEACH, A. R. *An Introduction to Chemoinformatics*. Dordrecht: Springer, 2007. ISBN 978-1-4020-6290-2.

[51] DAYLIGHT: *SMILES - A Simplified Chemical Language* [online]. 2008 [cit. 2019-04-20]. URL: `http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`

[52] DAYLIGHT: *SMARTS - A Language for Describing Molecular Patterns* [online]. 2008 [cit. 2019-04-20]. URL: `http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`

[53] DAYLIGHT: *SMARTS Examples* [online]. 2008 [cit. 2019-04-20]. URL: `http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html#X`

[54] RDKit: *An overview of the RDKit* [online]. 2018 [cit. 2019-04-20]. URL: `https://www.rdkit.org/docs/Overview.html`

[55] PostgreSQL: The World's Most Advanced Open Source Relational Database [online]. [cit 2019-05-02]. URL: `https://www.postgresql.org/`

[56] RAČEK, T., SCHINDLER, O., SVOBODOVÁ VAŘEKOVÁ, R. & KOČA, J. „Empirical methods for calculation of partial atomic charges - applicability for proteins?“. In *ENBIK2018*. 2018. ISBN 978-80-7592-017-1.

[57] SciPy.org [online]. [cit 2019-05-02]. URL: `https://www.scipy.org/`

[58] NLopt Documentation [online]. [cit 2019-05-02]. URL: `http://ab-initio.mit.edu/nlopt`

[59] VERSTRAELEN, T., AYERS, P. W., VAN SPEYBROECK, V. & WAROQUIER, M. „ACKS2: Atom-condensed Kohn-Sham DFT approximated to second order“. *The Journal of Chemical Physics* 2013, **138**(7). DOI: 10.1063/1.4791569.

[60] OLIFERENKO, A. A., PALYULIN, V. A., PISAREV, S. A., NEIMAN, A. V. & ZEFIROV, N. S. „Novel point charge models: reliable instruments for molecular electrostatic“. *Journal of Physical Organic Chemistry*. 2001, **14**(6), 355-369. DOI: 10.1002/poc.378.

[61] GILSON, M. K, GILSON, H. S. R. & POTTER, M. J. „Fast Assignment of Accurate Partial Atomic Charges: An Electronegativity Equalization Method that Accounts for Alternate Resonance Forms“. *Journal of Chemical Information and Computer Sciences*. 2003, **43**(6), 1982-1997. DOI: 10.1021/ci034148o.

[62] RAČEK, T., PAZÚRIKOVÁ, J., SVOBODOVÁ VAŘEKOVÁ, R., et al. „NEEMP: software for validation, accurate calculation and fast parameterization of EEM charges“. *Journal of Cheminformatics*. 2016, **8**(1). DOI: 10.1186/s13321-016-0171-1.

Příloha A

Obsah přiloženého archivu

K bakalářské práci je přiložen archiv **attachment.zip**, který obsahuje implementaci knihovny ATTYC a výstupy softwaru MACH jednotlivých parametrizací ve formě HTML souborů. Adresáře jsou označeny `strojovým písmem`. Struktura přiloženého archivu je následující:

- `program`
  - `attyc`: implementace knihovny ATTYC
  - `datasets`: klasifikované molekulové sady
  - **README.txt**: manuál knihovny ATTYC
  - **main.py**: ukázkový modul spouštějící vybranou klasifikaci atomů skrze knihovnu ATTYC
- `parameterization_results`
  Podadresáře adresářů `CCD_gen` a `Protein` obsahují výsledky parametrizací s užitím implementovaných klasifikátorů. Po názornost je uveden obsah podadresáře `hbo` první uvedené sady, zbylé podadresáře jsou pro přehlednost vynechány.
  - `CCD_gen`
    - `hbo`
      - **CDD_gen_hbo_EEM.html**: výsledky parametrizace metody EEM
      - **CDD_gen_hbo_PEOE.html**: výsledky parametrizace metody PEOE
  - `Protein`

Příloha B

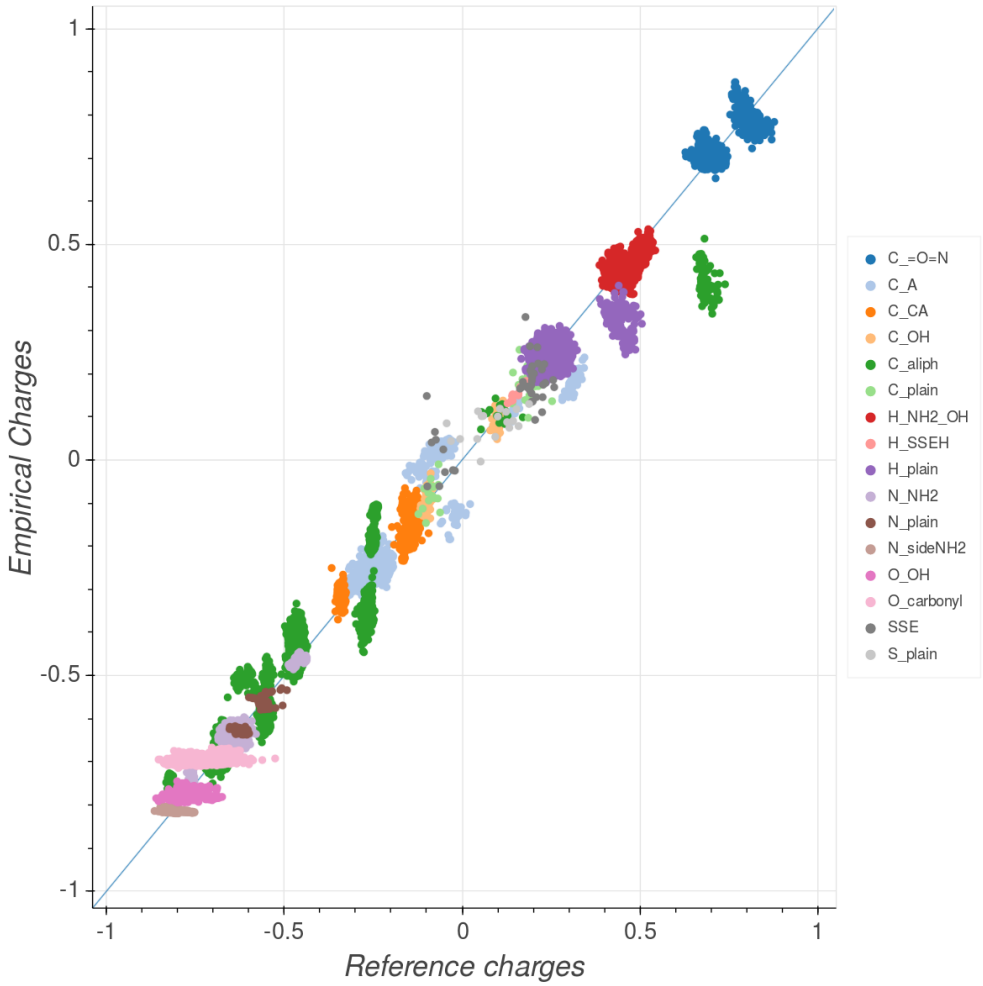
Statistiky parametrizací sady Protein

klasifikátor	metoda	RMSD	PCC^2	MAE	ABSMAX	doba výpočtu
`plain`	EEM	0,0671	0,9793	0,0498	0,3151	8:38:46
	PEOE	0,0619	0,9821	0,0339	0,4754	0:06:06
`hbo`	EEM	0,0498	0,9884	0,0373	0,3187	14:17:03
	PEOE	0,0531	0,9868	0,0307	0,451	0:10:55
`hybrid`	EEM	0,0527	0,9871	0,0394	0,3331	5:09:31
	PEOE	0,0585	0,9840	0,0377	0,4342	0:08:16
`substruct`	EEM	0,0334	0,9948	0,0229	0,3880	23:01:14
	PEOE	0,0493	0,9887	0,0292	0,4484	0:19:51
`substruct simplified`	EEM	0,0388	0,9930	0,0265	0,3198	14:30:45
	PEOE	0,0514	0,9876	0,0319	0,5083	0:13:02
`peptide`	EEM	0,0359	0,9940	0,0228	0,3558	1 den, 15:03:35
	PEOE	0,0456	0,9902	0,0293	0,3147	0:29:01
`peptide simplified`	EEM	0,0389	0,9929	0,0266	0,3647	18:28:26
	PEOE	0,0512	0,9877	0,0302	0,2536	0:15:12

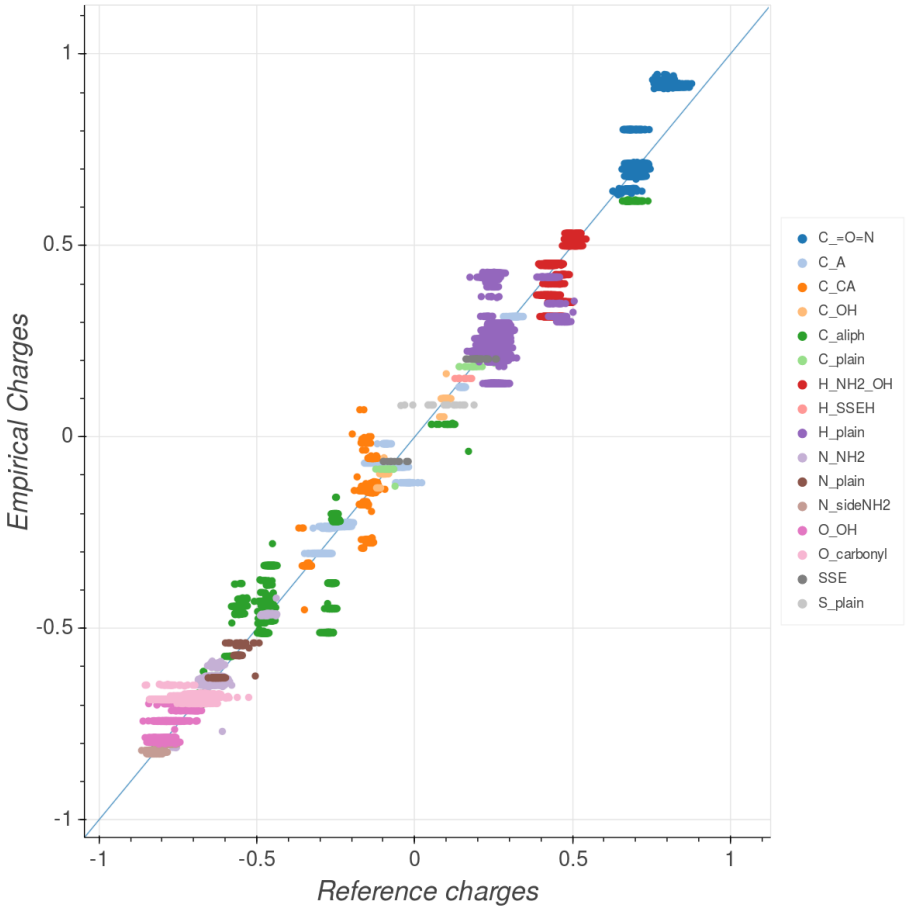
Tabulka B.1: Výsledky vybraných statistik parametrizace sady Protein. Jsou zaznamenány pouze statistiky tréninkových sad.

Příloha C

Klasifikátor 'peptide simplified': Výstupy



Obrázek C.1: Korelační graf parametrizace Protein/EEM/substruct_simplified. PCC^2=0,0.9929; RMSD=0,0389.



Obrázek C.2: Korelační graf parametrizace Protein/PEOE/substruct_simplified. PCC^2=0,9877; RMSD=0,0512.

atomový typ	**RMSD**	**PCC^2**	**MAE**	**ABSMAX**
C_A	0,0334	0,9948	0,0229	0,3880
N_plain	0,0388	0,9930	0,0265	0,3198

Tabulka C.1: Statistiky vybraných atomových typů užitých v rámci parametrizace Protein/PEOE/substruct_simplified. Korelační graf zmíněné parametrizace C.2 je zobrazen výše.