

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL

Bakalářská práce

BRNO 2019

RADKA SEDLÁKOVÁ

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, atomový typ, klasifikátor, EEM, PEOE, empirické metody |

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, atom type, EEM, PEOE, empirical methods |

Abstrakt

Parciální atomové náboje

Abstract

In this thesis we study

Místo tohoto listu vložte kopii oficiálního (podepsaného) zadání práce.

Poděkování

Na tomto místě bych chtěla poděkovat stránce LOTR University Memes za zlepšování nálady ve chvílích největší bezradnosti a skupině Dire Straits za příjemnou společnost ve večerních hodinách strávených psaním této práce.

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 | ix |
| Kapitola 1. Úvod | 1 |
| Kapitola 2. Teorie | 2 |
| 2.1 Parciální atomové náboje | 2 |
| 2.2 Kvantově-chemické metody | 3 |
| 2.2.1 Základy kvantové mechaniky | 3 |
| 2.2.2 Úvod do kvantově-chemických metod | 4 |
| 2.3 Empirické metody | 5 |
| 2.3.1 PEOE | 6 |
| 2.3.2 EEM | 6 |
| 2.3.3 Parametrizace | 8 |
| 2.3.4 Atomové typy | 8 |
| 2.4 Statistické pojmy | 9 |
| 2.4.1 Průměrná a maximální absolutní odchylka | 9 |
| 2.4.2 RMSD | 9 |
| 2.4.3 Pearsonův korelační koeficient | 9 |
| Kapitola 3. Metody | 11 |
| 3.1 Structure-Data file (SDF) | 11 |
| 3.2 PDB formát | 12 |
| 3.2.1 Nomenklatura atomových typů aminokyselin | 12 |
| 3.3 SMILES | 12 |
| 3.4 SMARTS | 12 |
| 3.5 RDKit | 13 |
| 3.6 MACH | 13 |
| Kapitola 4. Implementace | 14 |
| 4.1 Hledání strukturních motivů užitím SMARTS notace | 14 |
| 4.2 Rozšíření pro proteiny | 15 |
| 4.3 Výstup programu | 16 |
| Kapitola 5. Výsledky a diskuse | 17 |
| 5.1 Výsledky parametrizace | 17 |

| | |
|---|-----------|
| 5.1.1 Klasifikátor1 | 17 |
| 5.1.2 Klasifikátor2 | 17 |
| 5.1.3 Klasifikátor3 | 17 |
| 5.2 Srovnání navržených klasifikátorů | 17 |
| Kapitola 6. Závěr | 18 |
| Příloha | 19 |
| Seznam použité literatury | 20 |

Přehled použitého značení

Pro jednodušší orientaci čtenářů 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 |
| \mathbb{Z} | množina všech celých čísel |
| \mathbb{N} | množina všech přirozených čísel |

Kapitola 1

Úvod

Distribuce elektronů v m

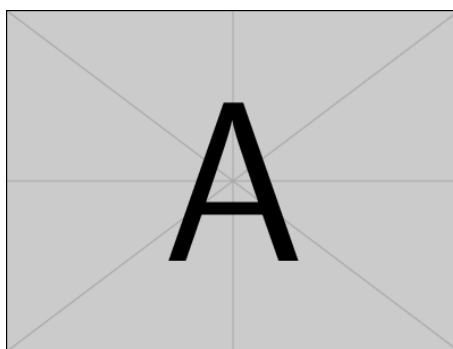
Kapitola 2

Teorie

Teoretická část práce seznamuje čtenáře 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 elektro-
pozitivnějšího vazebného partnera se elektronová hustota naopak snižuje a na atomu dochází ke vzniku parciálního kladného náboje (δ^+).



Obrázek 2.1: Rozložení elektronové hustoty v molekule ethanolu.

Koncept parciálních atomových nábojů je pouze teoretický, hodnoty nábojů proto nelze získat pomocí experimentu [2]. 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-chemi-

cké 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 pro výpočet parciálních atomových nábojů používají v rámci výpočtů parametrizovaná data a představují tak časově méně náročnou alternativu kvantově-chemického přístupu [3].

Aplikaci parciálních atomových nábojů lze nalézt ve výpočetní chemii a chemo-informaticce, kde slouží k predikci elektrostatických vlastností popisujících reaktivitu molekul. Uplatňují se v molekulových simulacích [4], ve virtuálním screeningu [5], při hledání vazebných míst proteinů nebo při návrhu farmakoforů [6]. Prokázaly se jako platné deskriptory v QSAR a QSPR modelech [7, 8]. V anorganické chemii se uplatňují při popisu toku elektronů v bateriích a katalyzátorech [9].

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 [10]. Vlnová funkce elektronu tak přispívá k popisu rozložení elektronové hustoty v molekule.

Základním úkolem kvantové mechaniky je výpočet vlnové funkce systému, z níž lze odvodit elektrostatický potenciál [11] nebo termodynamické vlastnosti molekuly [12]. Vlnová funkce je řešením Schrödingerovy rovnice

$$H\Psi = E\Psi \quad (2.1)$$

kde H je operátor hamiltonián a E je energie systému. Hamiltonián působí vlnovou funkcí 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 [13].

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 [14]. 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 [15]). 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 funkce, ale své poznatky staví na elektronové hustotě v molekule [2, 3, 16].

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

Příklady kvantově-chemických metod

Následující příklady popisují Hartree-Fockovu metodu, metody odvozené 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) [17].

- **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 [18]. 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 elektronovou hustotu, do množiny reálných čísel popisujících energii elektronů [19].

- **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 initio* výpočtů. Příkladem semi-empirické metody je metoda CNDO (*Complete*

Neglect of Differential Overlap) [20]. 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 [21], MNDO [22]) rozšířena o interakci orbitalů lokalizovaných na jiných jádrech a interakci volných elektronových párů.

Bázová sada

Bázová sada je soubor vlnových funkcí reprezentující atomové orbitály, jejichž vhodnou lineární kombinací (LCAO) lze následně vyjádřit vlnovou funkci molekuly. Pro popis funkcí reprezentujících atomové orbitály se používají orbitály 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 6-21*G [23].

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) [24] 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) [25], pracující s přirozenými atomovými orbitály. Přirozené atomové orbitály jsou odvozeny z bázové sady a jsou následně použity pro výpočet ortonormálních přirozených vazebných orbitalů (*Natural bonding orbitals*, NBO). Na základě NBO se poté provádí populační analýza.

Odlišný přístup finálního výpočtu parciálních atomových nábojů představuje metoda *Atoms-in-Molecules* (AIM) [26], která přiřazuje náboje atomům na základě integrace elektronové hustoty přes prostor příslušící danému atomu.

2.3 Empirické metody

Empirické metody výpočtu parciálních atomových nábojů byly vyvinuty v reakci na velkou výpočetní náročnost QM metod. V porovnání s QM metodami dosahují velmi přesných výsledků, a to ve výrazně kratším čase. Empirické metody se dělí na dvě hlavní skupiny, a to metody pracující s topologií molekuly (jinak řečeno s její 2D strukturou) 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*), známá také pod jménem autorů jako metoda Gasteiger-Marsili, byla poprvé publikována v roce 1980 [c]. Metodu lze aplikovat pouze na systémy obsahující σ vazby a nekonjugované π vazby, později však byla autory rozšířena i o výpočet systémů konjugovaných π vazeb. V rámci metody není uvažována 3D struktura molekuly, pracuje se pouze s její topologií.

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

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

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

$$\chi_{iv} = a_{iv} + b_{iv}Q_i + c_{iv}Q_i^2 \quad (2.3)$$

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. (za PEOE vložit tabulku parametrů abc, viz Gasteiger s. 330)

Při vzniku vazby dochází vlivem elektronegativity atomů k přesunu elektronů od elektropozitivnějšího atomu směrem k elektronegativnějšímu. Interagují spolu příslušné atomové orbitály a dochází k částečné ekvalizaci (vyrovnání) jejich nábojů. Množství přeneseného náboje mezi atomy A a B v k -té iteraci výpočtu, kde atom B má vyšší elektronegativitu, je definováno jako

$$Q^{(k)} = \frac{\chi_B^{(k)} - \chi_A^{(k)}}{\chi_A^+} \cdot \alpha^k \quad (2.4)$$

kde χ_A^+ označuje elektronegativitu kationtu atomu A . 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 šesti iteraci dochází k ustálení výpočtů.

Díky přesnosti a rychlosti výpočtů byla PEOE implementována do většiny programů pro molekulové modelování jako základní metoda výpočtu atomových nábojů. Reziduální elektronegativita atomů, získaná na základě PEOE, se prokázala vhodnou pro popis indukčního efektu v molekulách.

2.3.2 EEM

Autoři Mortier, Ghosh a Shankar publikovali metodu elektronové ekvalizace (*Electronegativity Equalization Method*, EEM) v r. 1986. Teoretický základ metody vychází z teorie funkcionálu hustoty, na němž je vystavěn matematický aparát pro výpočet atomových

nábojů. EEM využívá referenční náboje získané kvantově-mechanickými metodami, pomocí nichž parametrizuje vlastní výpočty. Díky nízké výpočetní náročnosti ($\Theta(N^3)$, kde N je počet atomů systému) a poměrně přesným výsledkům se stala metoda hojně využívanou a byla použita např. pro výpočet parciálních nábojů zeolitů, organických molekul nebo polypeptidů.

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

$$\bar{\chi} = \chi_1 = \chi_2 = \chi_3 = \dots = \chi_N \quad (2.5)$$

kde N je počet atomů, se elektronegativita každého atomu rovná průměrné elektronegativitě molekuly $\bar{\chi}$. Sandersonův postulát je potvrzen principy DFT.

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

$$\sum_i q_i = Q \quad (2.6)$$

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

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

V rovnici 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.

$$A_i = \chi_i^0 + \Delta\chi_i \quad (2.8)$$

$$B_i = 2(\eta_i^0 + \Delta\eta_i) \quad (2.9)$$

Cílem metody je empiricky nalézt hodnoty korekcí pro definované atomové typy a zajistit tak znovupoužitelnost uvedených parametrů.

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 ($q_1, q_2, q_3, \dots, q_N, \bar{\chi}$).

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

Po předchozí parametrizaci (sekce 2.3.3) získáváme z matice hodnoty parciálních atomových nábojů, které jsou následně porovnány s referenčními QM náboji. Proces uzavírá statistické vyhodnocení vypočtených dat.

Na principu EEM byly později vyvinuty další empirické metody jako např. *Atom-Bond Electronegativity Equalization Method* (ABEEM) nebo *General Bond Electronegativity Equalization Method* (GBEEM).

2.3.3 Parametrizace

Parametrizace je základním nástrojem empirických modelů, jejichž cílem je reprodukce experimentálních dat. Parametrizovat lze silová pole pro výpočty molekulové mechaniky nebo také *ab initio* výpočty zahrnující korelační energii elektronů. Cílem parametrizace je nalezení hodnot parametrů, po jejichž integraci do empirického modelu je dosaženo co nejlepší shody experimentálních a empirických výpočtů.

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

1. Výběr tréninkové sady molekul obsahující atomy, které v dostatečné míře reprezentují atomové typy, jež chceme parametrizovat
2. Výpočet parciálních atomových nábojů tréninkové sady pomocí QM
3. Parametrizace tréninkové sady na základě nábojů získaných v kroku 2
4. Výpočet nábojů testovací sady molekul kvantově-mechanickou a parametrizovanou empirickou metodou
5. Statistické vyhodnocení parametrizace

Cílem EEM je na základě znalosti parciálních nábojů odvozených z QM určit hodnoty parametrů A_i , B_i a κ (viz 2.10) pro každý atom molekuly. Pro zjednodušení výpočtu jsou odpovídající řádky matice (atomy molekuly) sloučeny pod jeden atomový typ, pro který jsou počítány výše zmíněné parametry. Ty jsou na základě znalosti dílčích atomových nábojů ($q_1, q_2, q_3, \dots, q_N$) a elektronegativity molekuly $\bar{\chi}$ získány z rovnice 2.7 upravené na tvar

$$A_i + B_i \cdot q_i = \bar{\chi} - \kappa \sum_{j \neq i}^N \frac{q_j}{R_{ij}} \quad (2.11)$$

Pro každou hodnotu parametru κ ze zvoleného intervalu jsou hledány vhodné hodnoty parametrů A_i a B_i . Každému atomovému typu je pak přiřazena trojice hodnot A_i , B_i a κ , která nejlépe reprodukuje náboje získané QM výpočty.

2.3.4 Atomové typy

Atomové typy jsou nástrojem parametrizace empirických metod výpočtu parciálních atomových nábojů. Definice atomových typů souvisí s charakteristickými chemickými vlastnostmi atomů (hybridizace, vazebný partner, nejvyšší řád vazby), které tyto atomové typy popisují. Každý atomový typ je definován takovou charakteristikou atomu, na základě které atom vykazuje odlišné chemické vlastnosti od jiných atomů, a je tedy vhodné pro něj definovat samostatný atomový typ. Názorným příkladem může být separace atomu uhlíku do tří atomových typů na základě jeho hybridizace. Prostorové

uspořádání orbitalů uhlíku je v hybridizacích sp^3 , sp^2 , sp diametrálně odlišné a předurčuje tak tvorbu odlišných typů vazeb s vazebnými partnery.

Napříč parametrizacemi empirických metod zaznamenáváme různé úrovně návrhu atomových typů, od triviálních klasifikací definujících atomové typy na základě protonových čísel až po komplexní rozdělení zahrnující nabitě funkční skupiny či příslušnost k větším atomovým celkům (aromatické systémy, postranní řetězce aminokyselin). Detailní dělení atomových typů nalézáme zejména v publikacích orientujících se na parametrizaci metod pro výpočet parciálních nábojů komplexních celků, např. polypeptidů.

2.4 Statistické pojmy

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).

$$MAE(X, Y) = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (2.12)$$

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

$$ABSMAX(X, Y) = \max |x_i - y_i| \quad (2.13)$$

2.4.2 RMSD

Veličina RMSD (z anglického *Root Mean Square Deviation*, někdy uváděná též jako *Root Mean Square Error*) popisuje míru odlišnosti dvojic odpovídajících hodnot (x_i, y_i) náhodných veličin X a Y 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.

$$RMSD(X, Y) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (2.14)$$

2.4.3 Pearsonův korelační koeficient

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

$$r(X, Y) = \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.15)$$

kde hodnoty $x_i, \dots, x_n, y_i, \dots, y_n$ jsou i -té prvky dvourozměrného náhodného vektoru o velikosti n realizovaného dvěma náhodnými veličinami X a Y , a kde \bar{x}, \bar{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 x.x.x), 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

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 [27, 28] 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 blocku* 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

3.2.1 Nomenklatura atomových typů aminokyselin

3.3 SMILES

SMILES [29, 30, 31] (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 jednoznačná SMILES reprezentace 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 geometrická 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 [32].

3.4 SMARTS

SMARTS notace [33, 34] (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 'c0c', 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 závis-

losti na jeho chemickém okolí. Příkladem je výraz '[O-;!\$([O-]C(=O))]', který definuje kyslíkový anion, jenž zároveň není součástí aniontu karboxylové skupiny COO-.

| | |
|------------------------------|---|
| <chem>[c,n&H1]</chem> | aromatický uhlík nebo (aromatický dusík s jedním vodíkem) |
| <chem>[c,n;H1]</chem> | (aromatický uhlík nebo aromatický dusík) s jedním vodíkem |
| <chem>[!#6][O][H]</chem> | hydroxylová skupina navázaná na jakýkoli atom kromě dusíku |
| <chem>[CX4,c][O][H]</chem> | alkoholová skupina navázaná na alifatický či aromatický uhlík |
| <chem>[NX3;!\$(NC=O)]</chem> | 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 [35] 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. 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.

Software MACH byl vyvinut s cílem optimalizovat parametrizaci empirických metod výpočtu parciálních nábojů za použití optimalizační metody Guided Minimization [36]. Software je vyvinut v jazyce Python za využití specializovaných knihoven pro práci s vědeckými daty (NumPy, SciPy, NLOpt). V softwaru byla úspěšně implementována parametrizace empirických metod EEM, PEOE, SFKEEM [37], QEq [4], ACKS2 [38] a MGC [39]. 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 sekce popisuje implementaci knihovny ATTYC (*ATom TYpe Classification*), která přiřazuje atomům molekulového souboru atomové typy na základě vybraného klasifikátoru. Pro formát PDB je určen klasifikátor `protein`, ostatní klasifikátory se vztahují k formátu SDF. Knihovna je implementována v jazyce Python 3.7 a obsahuje následující soubory:

`__init__.py` obsahuje funkci `classify_atoms(input_sdf, classifier)` spouštějící klasifikaci atomů molekulové sady za užití zvoleného klasifikátoru
`classifier.py` definuje rozhraní pro klasifikátory implementované ve složce `\classifiers`
`exceptions.py` definuje výjimky (potomky třídy `Exception`) pro řízení běhu programu
`io.py` zpracovává I/O všech částí programu včetně kontroly vstupních argumentů
`SMARTS_atom_types.txt` obsahuje SMARTS výrazy a odpovídající atomové typy pro vyhledání specifických strukturních motivů a funkčních skupin (viz sekci 4.1)
`\classifiers` obsahuje klasifikátory, které rozdělují atomy do atomových typů na základě:

`__init__.py`

`hbo.py` nejvyššího řádu vazby (angl. *highest bond order*)

`hybrid.py` hybridizace

`partners.py` vazebných partnerů

`protein.py` pozice atomu v rámci aminokyseliny

`substruct.py` příslušnosti ke strukturnímu motivu nebo funkční skupině

4.1 Hledání strukturních motivů užitím SMARTS notace

Klasifikátor `substruct` rozděluje atomy do atomových typů 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 [40, 41].

Strukturní motivy a funkční skupiny jsou v molekulách detekovány pomocí výrazů SMARTS, a to metodou třídy Mol `GetSubstructMatches(SMARTS_pattern)` implementovanou v knihovně RDKit. Příkaz `molecule.GetSubstructMatches(SMARTS_pattern)`, kde `molecule` je instance třídy Mol, vrátí n-tici (datový typ *tuple*) n-tic, které obsahují prvky typu integer. Tato čísla označují atomy molekuly vyhovující danému SMARTS dotazu. Čísla atomů vycházejí z pořadí, ve kterém jsou atomy definovány v SDF souboru. Příklad lze vidět na obr. 4.1.

```
# molecule je instance třídy Mol
SMARTS_pattern = "[SX4](=[OX1])(=[OX1])[OX2,OX1-]"
pattern_atoms = molecule.GetSubstructMatches(SMARTS_pattern)
for atom_tuple in pattern_atoms:
    print(atom_tuple, get_elements(atom_tuple))
print(pattern_atoms)
```

```
Python 3.7.0 (default, Apr 9 2019, 10:31:47)
>>> (5, 7, 8, 11) ('S', 'O', 'O', 'O')
>>> (21, 22, 26, 29) ('S', 'O', 'O', 'O')
>>> ((5, 7, 8, 11), (21, 22, 26, 29))
```

Obrázek 4.1: Ukázka funkce `GetSubstructMatches(SMARTS_pattern)`. Pomocí SMARTS výrazu jsou v molekule vyhledány sulfonové skupiny $\text{S(=O)}_2\text{OH}$ nebo její anionty. Indexy atomů jsou v n-ticích zapsány v pořadí odpovídajícímu prvkům ve SMARTS výrazu. Reálná implementace metody pro získání uvedeného výstupu, `molecule.GetSubstructMatches(Chem.MolFromSmarts(SMARTS_pattern))`, je v textu i v obrázku pro názornost syntakticky zjednodušena.

Klíčovým prvkem klasifikace je - na základě znalosti výstupů funkce `GetSubstructMatches(SMARTS_pattern)` pro jsou v souboru `SMARTS_atom_types.txt` definovány atomové typy

- validace volně dostupných SMARTS výrazů užitím online nástroje LiteMol a jejich úprava -> zpřesnění dotazů, odstranění redundantních výstupů - strukturní motivy hledány v pořadí, v jakém jsou definovány SMARTS vstupy v souboru `SMARTS_atom_types.txt`
- pořadí SMARTS vstupů ovlivňuje výsledný atomový typ, který je atomu přiřazen - SMARTS vstupy definovány od nejzvláštnějších k nejvíce obecným - atomům, které nebyly vyhledáváním strukturních motivů označeny, je přiřazen atomový typ 'plain'

4.2 Rozšíření pro proteiny

navržené atomové typy vychází z notace atomů aminokyselin dle IUPAC [42], která je použita pro specifikaci atomů v PDB souborech. - IUPAC 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 této nomenklatuře nahrazena velkými písmeny latinské abecedy (CA, CB, CG, CD, CE, CZ, CH). Navržená klasifikace přiřazuje atomové typy obdobně jako výše uvedený klasifikátor `substruct.py`, tedy na základě příslušnosti atomů ke strukturálním celkům.

Po analýze atomových typů aminokyselin dle IUPAC Soubor AA_atom_types.txt

4.3 Výstup programu

- spíše do přílohy?

Kapitola 5

Výsledky a diskuse

5.1 Výsledky parametrizace

Jedním z cílů této bakalářské práce je otestovat, jak jemná či hrubá klasifikace atomových typů je postačující pro parametrizaci výpočetních metod poskytujících ve srovnání s QM metodami kvalitní výsledky. V souvislosti s parametrizací je též testováno, zda jsou některé charakteristiky atomu (např. hybridizace, nejvyšší řád vazby, vazebný partner) pro definici atomových typů vhodnější než charakteristiky jiné.

5.1.1 Klasifikátor1

5.1.2 Klasifikátor2

5.1.3 Klasifikátor3

5.2 Srovnání navržených klasifikátorů

Kapitola 6

Závěr

[illegible][illegible]

Příloha

Sem můžete přidat přílohu. Pokud chcete “Přílohy”, tak upravte definici záhlaví v souboru `sci.muni.thesis.sty`, viz příkaz `\HlavickaPriloha`.

Seznam použité literatury

- [1] ATKINS, P. W. & DE PAULA, J. *Atkins' physical chemistry*. 9th ed. Oxford: Oxford University Press, c2010. ISBN 978-0-19-954337-3.
- [2] LEACH, A. R. *Molecular modelling: principles and applications*. 2nd ed. New York: Prentice Hall, 2001. ISBN 0-582-38210-6.
- [3] GASTEIGER, J. & ENGEL, T. *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH, c2003. ISBN 978-3-527-30681-7.
- [4] 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. ISSN: 0022-3654. URL: <http://pubs.acs.org/doi/abs/10.1021/j100161a070>
- [5] 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. ISSN: 13596446. URL: <http://linkinghub.elsevier.com/retrieve/pii/S1359644602024832>
- [6] 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. ISSN: 0009-2665.
- [7] 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. ISSN: 00223573.
- [8] VAŘKOVÁ, 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. ISSN: 1758-2946.
- [9] 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. ISSN: 1549-9618. URL: <http://pubs.acs.org/doi/10.1021/ct500790p>

- [10] CELÝ, J. *Základy kvantové mechaniky pro chemiky: I. Principy*. Brno: Rektorát UJEP Brno, 1986.
- [11] HÖLTJE, H.-D. & FOLKERS, G. *Molecular Modeling: Basic Principles and Applications*. Volume 5. Weinheim: Wiley-VCH, 2008. ISBN 978-3-527-61476-9.
- [12] 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. ISBN 9780470125922.
- [13] JEAN, Y., VOLATRON, F. & BURDETT, J. K. *An introduction to molecular orbitals*. New York: Oxford University Press, 1993. ISBN 0-19-506918-8.
- [14] PILAR, F. L. *Elementary quantum chemistry*. Dover ed. Mineola, N.Y.: Dover Publications, 2001. ISBN 0-486-41464-7.
- [15] „Hierarchy of *ab initio* Post-HF methods“. Přetisknuto s laskavým svolením Prof. Martina Kauppa, TU Berlín. In SEMRÁD, H. „Studium mechanismu bromoborární 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/
- [16] CRAMER, C. J. *Essentials of Computational Chemistry: Theories and Models*. Chichester: John Wiley, 2002. ISBN 0-471-48552-7.
- [17] LEVINE, I. N. *Quantum Chemistry*. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2000.
- [18] SEFZIK, T. H., TURCO, D., IULIUCCHI, 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.
- [19] KOCH, W. & HOLTHAUSEN, M. C. *A chemist's guide to density functional theory*. New York: Wiley-VCH, 2000. ISBN 3527299181.
- [20] POPLÉ, 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.
- [21] POPLÉ, 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.
- [22] 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.

- [23] DAVIDSON, E. R. & FELLER, D. „Basis set selection for molecular calculations“. *Chemical Reviews*. 1986, **86**(4), 681-696. DOI: 10.1021/cr00074a002.
- [24] 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.
- [25] 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.
- [26] BADER, R. F. W. *Atoms in molecules: a quantum theory*. New York: Oxford University Press, 1994. ISBN 978-0198558651.
- [27] *CTFile Formats* [online]. [cit. 2019-04-18]. URL: <http://c4.cabrillo.edu/404/ctfile.pdf>
- [28] 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. ISSN 1549-9596. URL: <http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00007a012>
- [29] 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. ISSN 1549-9596.
- [30] BUNIN, B. A. *Chemoinformatics: Theory, Practice, & Products*. Dordrecht: Springer, 2007. ISBN 987-1-4020-5000-8
- [31] LEACH, A. R. *An Introduction to Chemoinformatics*. Dordrecht: Springer, 2007. ISBN 978-1-4020-6290-2
- [32] *DAYLIGHT: SMILES - A Simplified Chemical Language* [online]. 2008 [cit. 2019-04-20]. URL: <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>
- [33] *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>
- [34] *DAYLIGHT: SMARTS Examples* [online]. 2008 [cit. 2019-04-20]. URL: http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html#X
- [35] *RDKit: An overview of the RDKit* [online]. 2018 [cit. 2019-04-20]. URL: <https://www.rdkit.org/docs/Overview.html>

- [36] 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
- [37] 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. ISSN 1549-9596. URL: <https://pubs.acs.org/doi/10.1021/ci050505e>
- [38] 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. ISSN 0021-9606. URL: <http://aip.scitation.org/doi/10.1063/1.4791569>
- [39] 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. ISSN 08943230. URL: <http://doi.wiley.com/10.1002/poc.378>
- [40] 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. ISSN 0095-2338.
- [41] 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. ISSN 01928651.
- [42] 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. ISSN 0014-2956.

