STŘEDNÍ PRŮMYSLOVÁ ŠKOLA CHEMICKÁ, BRNO
THE SECONDARY TECHNICAL SCHOOL OF CHEMISTRY, BRNO

PŘÍRODOVĚDECKÁ FAKULTA MASARYKOVY UNIVERZITY
NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL

FACULTY OF SCIENCE AT MASARYK UNIVERSITY
NATIONAL CENTRE FOR BIOMOLECULAR RESEARCH

PREDIKCE pK_a PRO NOVĚ NAVRŽENÉ MOLEKULY LÉKŮ

STŘEDOŠKOLSKÁ ODBORNÁ ČINNOST
OBOR: 3. CHEMIE

AUTOR PRÁCE ROMAN BERÁNEK
AUTHOR

STŘEDNÍ PRŮMYSLOVÁ ŠKOLA CHEMICKÁ, BRNO
THE SECONDARY TECHNICAL SCHOOL OF CHEMISTRY, BRNO

PŘÍRODOVĚDECKÁ FAKULTA MASARYKOVY UNIVERZITY
FACULTY OF SCIENCE AT MASARYK UNIVERSITY

NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL
NATIONAL CENTRE FOR BIOMOLECULAR RESEARCH

PREDIKCE pK_a PRO NOVĚ NAVRŽENÉ MOLEKULY LÉKŮ

pK_a PREDICTION FOR DRUG MOLECULE CANDIDATES

STŘEDOŠKOLSKÁ ODBORNÁ ČINNOST
OBOR: 3. CHEMIE

AUTOR PRÁCE — ROMAN BERÁNEK
AUTHOR

KONZULTANT — RNDr. RADKA SVOBODOVÁ VAŘEKOVÁ, Ph.D.
CONSULTANT

BRNO 2012

ABSTRAKT

Predikce hodnot disociačních konstant pro dosud nesyntetizované molekuly je oblastí, která má velký význam pro farmaceutický průmysl. Velmi slibnou metodikou pro predikci pK_a je aplikace QSPR modelů využívajících jako deskriptory parciální atomové náboje. Hodnoty nábojů je nutno vypočítat na základě 3D struktur molekul, přičemž tyto struktury lze generovat a optimalizovat různými metodami a softwarovými nástroji. Kvalita vygenerovaných struktur je klíčovým faktorem ovlivňujícím přesnost predikce pK_a.

V rámci své práce jsem nejdříve analyzoval vliv metod pro generování a optimalizaci 3D struktur na přesnost predikce pK_a, přičemž jsem zohlednil i vliv použitého typu nábojů. Konkrétně jsem sestavil tréninkové sady obsahující molekuly fenolů, anilinů a karboxylových kyselin, vygeneroval a optimalizoval pro ně 9 sad 3D struktur a pro každou 3D strukturu vypočítal 12 různých typů nábojů. Na základě těchto dat jsem vytvořil a parametrizoval 540 QSPR modelů a porovnal jejich přesnost. Výsledky těchto analýz potvrdily, že automaticky generované struktury jsou vhodnými vstupy pro predikci pK_a (37 % vytvořených QSPR modelů mělo hodnoty $R^2 > 0{,}9$). Dále jsem pak na základě uvedených analýz naleznul nejlepší metodiku pro predikci pK_a: Vygenerovat 3D struktury molekul pomocí software CORINA, tyto 3D struktury dále neoptimalizovat a vypočítat pro ně náboje pomocí HF/6–31G*/NPA.

Uvedenou metodiku jsem poté využil k predikci pK_a tří molekul léků (dronabinol, levorfanol a pentazocin), které nebyly součástí tréninkové sady. Hodnoty pK_a, predikované tímto způsobem, velmi přesně odpovídaly experimentálním hodnotám pK_a daných léků.

KLÍČOVÁ SLOVA

predikce pK_a, disociační konstanta, QSPR, molekulová mechanika, kvantová mechanika, Balloon, Corina, Open Babel, Gaussian

ABSTRACT

Prediction of dissociation constants for molecules, which were currently not synthesized, is a very important topic for pharmaceutical industry. A very promissing pK_a prediction method is an application of QSPR models employing partial atomic chares as descriptors. Values of charges are calculated from 3D structures of molecules. These 3D structures can be generated and optimized by various methods and software tools. A quality of the 3D structures strongly influences an accuracy of pK_a prediction.

In my work, first I analysed an influence of methods for 3D structure generation and optimization on an accuracy of pK_a prediction. An influence of atomic charges was also included into these analyses. Specifically, I prepared training sets containing molecules of phenols, anilines and benzoinc acids. Then, I generated and optimized 9 sets of 3D structures for each molecule. Afterwards, I calculated 12 different charge types for each 3D structure. Using these data, I created and parameterized 540 QSPR models and compared their accuracy. Results of these analyses confirmed that the automatically generated structures are very good inputs for pK_a prediction (37 % of our QSPR models have $R^2 > 0{,}9$). Next, based on these results, I found the best method for pK_a prediction: Generate 3D structures of molecules by CORINA, do not use any optimization and calculate charges using HF/6–31G*/NPA.

Afterwards, I used this method for pK_a prediction of three drug molecules (dronabinol, levorphanol and pentazocine), which were not a part of my training set. pK_a values calculated this way very precisely reflected experimental pK_a values of these drugs.

KEYWORDS

pK_a prediction, dissociantion constant, QSPR, molecular mechanics, quantum mechanics, Balloon, Corina, Open Babel, Gaussian

BERÁNEK, Roman *Predikce pKa pro nově navržené molekuly léků*: středoškolská odborná činnost. Brno: Střední průmyslová škola chemická, Brno, Vranovská 65, 55 s. Vedoucí práce byla RNDr. Radka Svobodová Vařeková, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou práci na téma „Predikce pKa pro nově navržené molekuly léků“ jsem vypracoval samostatně pod vedením vedoucího stáže a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené práce dále prohlašuji, že v souvislosti s jejím vytvořením jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

Brno ....

(podpis autora)

Děkuji vedoucí mé práce, RNDr. Radce Svobodové Vařekové, Ph.D. a odbornému konzultantovi Bc. Stanislavu Geidlovi za cenné rady a připomínky a veškerou pomoc.
Rád bych poděkoval rovněž všem svým přátelům, sestře a mamince, jež to se mnou nejen ve dnech intenzívní tvorby neměli vůbec jednoduché (a ani mít nebudou).
Moc rád bych poděkoval všem, kteří mi s touto prací pomohli, a bez kterých by vlastně tato práce ani nevznikla. Je škoda, že je nemohu všechny vyjmenovat – jejich seznam by tuto práci zvětšil o několik stran.

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Organická chemie přitahuje výzkumné pracovníky i studenty obrovským počtem již známých organických sloučenin a především ještě větším množstvím molekul dosud nesyntetizovaných. S drobnou aproximací se dá říci, že na svůj objev čeká takřka nekonečný počet organických molekul. Syntéza takového množství sloučenin by však jistě byla časově i finančně extrémně náročná. Proto je zapotřebí mechanismu, kterým se ze všech těchto molekul před samotnou syntézou vyberou ty, které jsou nejzajímavější.

Tato problematika je oblastí intenzivního výzkumu a investic například ve farmaceutickém průmyslu. Na základě několika struktur známých léků zaměřených na určité onemocnění lze totiž navrhnout tisíce či desetitisíce podobných molekul, které budou mít potenciálně vyšší účinnost či lepší vlastnosti [11]. Protože není možno syntetizovat a testovat všechny navržené molekuly, je nutno vlastnosti těchto molekul zjistit výpočetně (predikovat). Poté lze na základě predikcí vybrat k syntéze jen ty nejvhodnější molekuly.

Uvedené požadavky farmaceutického průmyslu spolu s obrovským nárůstem dostupných informací o strukturách organických molekul vedly v posledních letech ke vzniku vědního oboru chemoinformatika [12, 22, 39]. Tento obor využívá pro řešení chemických problémů informatických a algoritmických přístupů a aplikuje rovněž metodiky počítačové chemie a molekulového modelování [39]. Chemoinformatika se převážně zaměřuje na získání informací z databází malých nebo středně velkých molekul, predikci vlastností těchto molekul, návrh molekul s definovanými vlastnostmi apod.

Jednou z velkých výzev, kterými se vědečtí pracovníci v oblasti chemoinformatiky zabývají, je predikce disociačních konstant molekul [40]. Hodnoty disociačních konstant jsou velmi zajímavé pro chemický, biologický a environmentální výzkum, protože důležité fyzikálně-chemické vlastnosti látek – lipofilicita, rozpustnost a propustnost – jsou závislé na pK_a. Obzvláště velký význam má pak pK_a pro farmaceutický průmysl, konkrétně pro oblast vývoje léků. Hodnota pK_a je jedním z podstatných kritérií, které nám umožní eliminovat z množiny navrhovaných molekul léků nevhodné molekuly. Molekuly léků totiž nesmí být ani příliš silnými kyselinami, ani příliš silnými bazemi, protože jinak by poškozovaly organismus. Proto se jejich pK_a musí pohybovat v definovaném intervalu.

Velmi slibnou metodikou pro predikci pK_a je využití QSPR (Quantitative Structure Property Relationship) [39] modelů. Vstupem těchto modelů jsou číselné charakteristiky molekul (deskriptory), které jsou vypočítány na základě struktury molekul. Vlastní QSPR modely jsou pak matematické vztahy (lineární rovnice), které na základě těchto deskriptorů počítají hodnoty fyzikálně-chemických vlastností molekul (např. pK_a). Velice úspěšnými deskriptory pro výpočet pK_a jsou parciální náboje

na atomech v rámci molekuly [16, 56].

Pokud však chceme predikovat pK_a pro molekuly, které ještě nebyly syntetizovány, musíme nejdříve vyřešit jednu velmi závažnou otázku. Jak získat struktury těchto molekul? Tyto struktury jsou totiž nezbytné, abychom na jejich základě vypočítali náboje a ty pak využili k predikci pK_a pomocí QSPR modelů. Struktury nemůžeme získat experimentálně, protože dané molekuly nebyly syntetizovány. Je proto nutno tyto struktury konstruovat (generovat) pomocí vhodných softwarových nástrojů a poté dále zpřesňovat (optimalizovat). Softwarových nástrojů pro generování struktur molekul existuje několik (např. CORINA [22], Open Babel [48], Balloon [61]). Tyto nástroje používají různé aloritmy a jimi vytvořené struktury molekul se proto liší. K optimalizaci lze použít metody molekulové mechaniky [33, 48] nebo kvantové mechaniky [20, 33]. Kvalita generovaných struktur molekul (t.j., přesnost, s jakou popisují reálnou chemickou strukturu molekul a speciálně pak oblast disociace) je klíčovým faktorem ovlivňujícím přesnost QSPR modelů.

Protože predikce pK_a s využitím QSPR modelů je oblastí, kde stále probíhá intenzivní výzkum, nejsou dostupné studie analyzující vliv metody generování struktury molekul na přesnost predikce pK_a. Proto jsem se v rámci své práce zaměřil právě na tuto tématiku.

Konkrétní cíle mé práce jsou:

- Seznámení se s důležitými chemoinformatickými a počítačově chemickými pojmy a metodami – zápis molekuly v počítači, generování struktur molekul a jejich optimalizace, parciální atomové náboje a jejich metody výpočtu, QSPR modelování atd.
- Výběr konkrétních molekul, na kterých budu své analýzy realizovat (tzv. tréninková sada molekul). Jedná se o molekuly substituovaných fenolů, anilinů a karboxylových kyselin. Vyhledání experimentálních hodnot pK_a pro tyto molekuly.
- Vygenerování struktur pro všechny vybrané molekuly pomocí softwarových nástrojů CORINA, Open Babel a Balloon.
- Molekulově mechanická a kvantově mechanická optimalizace všech vytvořených struktur.
- Výpočet nábojových deskriptorů pro všechny získané struktury pomocí několika různých metodik pro výpočet nábojů.
- Vytvoření a parametrizace QSPR modelů pro všechny získané struktury a všechny typy nábojových deskriptorů.
- Výpočet kvalitativních kritérií QSPR modelů, jejich porovnání a diskuse vlivu různých faktorů na kvalitu QSPR modelů. Nalezení nejvhodnější metodiky pro predikci pK_a dosud nesyntetizovaných molekul.
- Ověření použitelnosti vytvořené metodiky na vybraných molekulách léků, které nebyly součástí tréninkové sady.

1 TEORIE

1.1 Disociační konstanta

Chemické reakce mnoha organických sloučenin se dají vysvětlit jako reakce kyselin a bází. Směr těchto reakcí je mimo jiné určen kyselostí a bazicitou reaktantů i produktů. Význam těchto dvou vlastností definuje Brønsted-Lowryho teorie kyselin a zásad [4].

Brønstedova kyselina je látka, která poskytuje proton a zásada je látka, která proton přijímá.

Při rozpouštění kyseliny (HA) ve vodě probíhá následující reakce:

$$\mathrm{HA}(aq) + \mathrm{H_2O}(l) \rightleftharpoons \mathrm{A^-}(aq) + \mathrm{H_3O^+}(aq) \tag{1.1}$$

Voda zde vystupuje jako zásada a po přijetí protonu přejde na oxoniový kation, tedy **konjugovanou kyselinu**. Anion kyseliny vzniklý po uvolnění protonu se nazývá **konjugovaná báze**.

Rovnováha v rámci této reakce je popsána rovnovážnou konstantou K, která je definována vztahem:

$$K = \frac{a_{\mathrm{H_3O^+}} a_{\mathrm{A^-}}}{a_{\mathrm{HA}} a_{\mathrm{H_2O}}} \tag{1.2}$$

kde $a_{\mathrm{H_3O^+}}$ je aktivita oxoniového katintu, a_{HA} je aktivita nedisociované kyseliny, $a_{\mathrm{A^-}}$ je aktivita její báze a $a_{\mathrm{H_2O}}$ je aktivita vody, která se ve zředěných roztocích blíží jedné. Rovnováhu lze pak vyjádřit konstantou acidity (kyselosti) neboli **disociační konstantou kyseliny** K_a.

$$K_a = \frac{a_{\mathrm{H_3O^+}} a_{\mathrm{A^-}}}{a_{\mathrm{HA}}} \tag{1.3}$$

Pro aktivitu a_i látky i je ve vztahu k její koncentraci c_i platí:

$$a_i = \gamma_{c,i} \frac{c_i}{c^\ominus} \tag{1.4}$$

kde $\gamma_{c,i}$ je aktivitní koeficient látky i a $c^\ominus$ je standardní koncentrace. U zředěných roztoků zanedbáváme aktivitní koeficienty. V praxi se tak vztah pro výpočet disociační konstanty aproximuje a aktivity přítomných složek jsou nahrazeny koncentracemi.

$$K_a \approx \frac{[\mathrm{H_3O^+}][\mathrm{A^-}]}{[\mathrm{HA}]} \tag{1.5}$$

Pro svůj vysoký řádový rozptyl se K_a zapisuje spíše jako její *záporně vzatý dekadický logaritmus*, tedy $\mathrm{p}K_a$.

$$\mathrm{p}K_a = -\log K_a \tag{1.6}$$

Podle výše zmíněných vztahů s klesajícím $\mathrm{p}K_a$ míra disociace kyseliny (a tím i síla kyseliny – acidita) roste a vice versa. Silné kyseliny mají $\mathrm{p}K_a$ menší než -1, například $\mathrm{p}K_a(\mathrm{HCl}) = -7$ a $\mathrm{p}K_a(\mathrm{HNO_3}) = -1{,}3$. Slabé kyseliny mají hodnotu $\mathrm{p}K_a$ v rozmezí od -1 až do 12, např. $\mathrm{p}K_a(\mathrm{CH_3COOH}) = 4{,}72$ a $\mathrm{p}K_a(\mathrm{C_6H_5OH}) = 9{,}95$ [43].

Mluví-li se o $\mathrm{p}K_a$ zásad (např. později diskutovaných anilinů), myslí se tím obvykle $\mathrm{p}K_a$ jejich **konjugovaných kyselin**. Ty disociují podle následující rovnice:

$$\mathrm{BH^+}(aq) + \mathrm{H_2O}(l) \rightleftharpoons \mathrm{B}(aq) + \mathrm{H_3O^+}(aq) \tag{1.7}$$

Alternativou je použití $\mathrm{p}K_b$, tedy *záporně vzatý dekadický logaritmus disociační konstanty zásady*. Více o $\mathrm{p}K_b$ je uvedeno v příloze v sekci C.1 na straně 53.

Síla bazí s klesajícím $\mathrm{p}K_a$ konjugované kyseliny klesá. Jako příklad silné báze lze uvést hydroxid sodný s $\mathrm{p}K_a(\mathrm{NaOH_2^+}) = 13{,}8$. Anilin má $\mathrm{p}K_a(\mathrm{C_6H_5NH_3^+}) = 4{,}6$ [4].

Hodnoty $\mathrm{p}K_a$ jsou velice zajímavé pro chemický, biologický, environmentální a farmaceutický výzkum, protože důležité fyzikálně chemické vlastnosti – lipofilicita, rozpustnost a propustnost – jsou závislé na těchto hodnotách.

1.1.1 Význam disociační konstanty pro návrh léků

Obzvláště velkou důležitost mají hodnoty $\mathrm{p}K_a$ při návrhu nových léků. Na základě několika struktur známých léků zaměřených na určité onemocnění lze totiž navrhnout tisíce či desetitisíce podobných molekul, které budou mít potenciálně vyšší účinnost či lepší vlastnosti. Velmi náročným úkolem je pak zjistit, které z těchto molekul jsou opravdu vhodnými léky. Není možno syntetizovat všechny navržené molekuly a testovat jejich vlastnosti. Velice užitečným krokem je proto vyloučit nejdříve molekuly, které vhodnými léky být nemohou. Hodnota $\mathrm{p}K_a$ je jedním z podstatných kritérií, které nám umožní takovéto molekuly najít. Molekuly léků totiž nesmí být ani příliš silnými kyselinami, ani příliš silnými bazemi, protože jinak by poškozovaly organismus. Jejich $\mathrm{p}K_a$ by v případě kyselin nemělo být nižší než 2,5 a v případě bazí by nemělo překročit 11 [5].

Disociační konstanty přinášejí také vhled do interakce nabitých léků s receptory. Navíc je znalost $\mathrm{p}K_a$ nezbytná pro posuzování ADME (absorpce, distribuce, metabolismus, vylučování) profilu [63], jenž je dalším z důležitých vodítek při návrhu léku. Z těchto důvodů je v rámci vědecké komunity kladen velký důraz na vývoj přesnějších a efektivnějších metod pro predikci $\mathrm{p}K_a$.

1.1.2 Výpočet disociační konstanty

Dosud bylo vyvinuto mnoho metod pro predikci pK_a založených na mnoha různých přístupech. Zde je přehled těch nejběžnějších:

LFER Metody založené na lineárním vztahu mezi pK_a a volnou energií byly jedny z prvních [51]. Aplikují Hammettovu a Taftovu rovnici [15, 44] a stále jsou implementovány v softwarech jako ACD/pKa [2], EPIK [54] a SPARC [28]. Přesnost získané hodnoty je závislá především na velikosti databáze s pK_a [6].

DB Databázové metody [64] používají podobnostní metriky a přiřazují tedy vstupní molekule hodnotu pK_a, která přísluší nejvíce podobné molekule v databázi.

Strom Metoda rozhodovacího stromu [65] používá také podobnostní metriku a buduje rozhodovací strom, který poskytuje cestu pro predikci pK_a neznámých sloučenin.

QM Kvantově mechanická simulace [29] je obecně nejpřesnější metodou pro výpočet pK_a. Je implementována např. v rámci modulu pro software Jaguar [53]. Její rozsáhlé využití však znemožňuje časová náročnost.

QSPR Kvantově mechanické výpočty lze využít i takovým způsobem, aby jejich výpočetní náročnost byla tak vysoká. Konkrétně můžeme použít kvantově **kvantově mechanických deskriptorů** [32], které s pK_a silně korelují. Mezi takové deskriptory patří polarizovatelnost, volná energie (např. HOMO energie fenoxidového aniontu [26] nebo relativní transferové energie vodíku [26]), parciální náboje na atomech [16, 19, 24], elektrostatický potenciál molekuly [41] atd. Predikce pK_a na základě deskriptorů probíhá s využitím QSPR (Quantitative Structure-Property Relationship) modelů [24, 32, 35].

I přes úsilí vynaložené na vytvoření optimální metody, je pK_a stále jednou z nejhůře předvídatelných vlastností.

1.2 Atomové náboje

Rozdíl elektronegativit atomů vázaných v molekulách je zdrojem nerovnoměrného rozložení elektronů v chemických vazbách. V důsledku tohoto asymetrického rozložení záporného náboje lokalizujeme na atomech parciální náboje [4]. Pokud je neutrální atom chemicky vázán na další neutrální atom, který má větší elektronegativitu, pak jsou elektrony jádra prvního atomu přitahovány k druhému atomu. První atom takto získává částečný kladný náboj a druhý částečný záporný náboj.

Parciální atomové náboje jsou velmi důležité vlastnosti molekul. Jsou klíčové pro výpočet fyzikálních, chemických a biologických vlastností nebo reaktivity molekul [9, 13, 40, 43, 58, 67]. Navíc, pomocí informací o atomových nábojích je možné

předpovědět stabilitu různých molekul, směr chemických reakcí, interakce s biomolekulami a tak dále.

1.2.1 Výpočet atomových nábojů

Navzdory užitečnosti nábojů neexistuje žádná přímá metoda, která by umožňovala určit náboje experimentálně. Z tohoto důvodu byly vyvinuty různé přístupy ke kalkulaci parciálních nábojů, z nichž každý poskytuje jistým způsobem odlišné výsledky [56].

Nejznámějšími metodami pro výpočet nábojů jsou kvantově mechanické metody [4, 66] následované využitím populační analýzy [19, 66]. Tyto metody využívají kvantovou mechaniku pro výpočet elektonové hustoty v rámci orbitalů a populační analýza pak slouží k rozdělení této elektronové hustoty mezi atomy. Nevýhodou těchto metod je však časová náročnost[1]. Uvedené metody budou využity v rámci této práce, proto je na ně podrobně zaměřen následující text. Dalšími metodami pro výpočet atomových nábojů jsou empirické metody, které rozdělují elektronovou hustotu mezi atomy pomocí různých heuristik. Příkladem takovýchto metod je Gasteiger-Marsiliho metoda [23] nebo EEM (Electronegativity Equalization Method) [45].

1.3 Kvantová mechanika

Kvantová mechanika je vypočetní metoda, která modeluje molekulové systémy užitím **kvantové teorie** [4]. Kvantová teorie byla vyvinuta, protože klasická (Newtonova) mechanika nedokázala popsat některé jevy mikrosvěta, konkrétně kvantování energie mikročástic, Heisenbergův princip neurčitosti, vlnově částicový dualismus atd. Kvantová teorie je založena na následujících principech: Každý fyzikální systém může být reprezentován užitím Hilbertova prostoru[2]. Každý stav systému v Hilbertově prostoru je plně popsán vektorem ψ. Tomuto vektoru se též říká **vlnová funkce**. V teoretické chemii vlnová funkce popisuje rozložení elektronů v molekulových systémech. Vlnová funkce je funkce $\psi : \mathbb{R}^3 \to \mathbb{Q}$, jejíž definičním oborem jsou souřadnice částic v prostoru. Oborem hodnot vlnové funkce jsou komplexní čísla jejichž čtverce popisují pravděpodobnost výskytu částice v nějakém bodě prostoru.

Vlnovou funkci můžeme získat řešením Schrödingerovy rovnice:

$$\hat{H}\psi = E\psi \tag{1.8}$$

[1] Časová složitost je $\theta(B^4)$, kde B je větší nebo rovno počtu elektronů v molekule. [55]

[2] Úplný unitární vektorový prostor.

kde E je energie a $\hat{H}$ Hamiltonův operátor. Schrödingerova rovnice je diferenciální rovnicí druhého řádu a jejím řešení jsou dvojice (ψ, E). Vlnová funkce jednoho elektronu se nazývá **atomový orbital** či **molekulový orbital** a popisuje (pomocí hodnoty čtverce ψ^2) distribuci tohoto elektronu v atomu respektive molekule.

Ukázalo se, že Schrödingerovu rovnici lze řešit *analyticky* pouze pro jednoelektronové systémy (atom vodíku, kationty H_2^+ a He^+)[18]. Pro víceelektronové systémy je potřeba zavést některá zjednodušení – např. Born Oppenheimerovu aproximaci, model nezávislých částic atd. Více informací o uvedených aproximacích uvádím v příloze C.2. Tyto aproximace mohou být velmi užitečné, ale vyžadují od člověka, aby předvídal, kdy jsou aproximace ještě platné a jakou přesnost lze od nich očekávat. K řešení Schrödingerovy rovnice užíváme dvou hlavních přístupů [33, 39]: semiempirické metody a níže diskutované **ab initio metody**.

Ab initio metody jsou výpočetní metody odvozené přímo z teorie a neobsahují žádná experimentální data krom hmotností elementárních částic. Hamiltonián je vyjádřen různými aproximacemi nazývanými **úroveň teorie** [4, 33] a vlnová funkce je nahrazena množinou jednoduchých funkcí nazývaných **bázová sada** [4].

Semiempirické metody zase užívají parametrů a/nebo ignorují některé výrazy v Hamiltonově operátoru. Semiepirické metody nebudou použity v této práci, tudíž není důvod se jím dále věnovat.

1.3.1 Úrovně teorie

Nejvíce využívané úrovně teorie v ab-initio metodách jsou:

Hartree–Fock

Tato metoda [4, 55, 66] využívá Born-Oppenheimerovu aproximaci, model nezávislých částic a konečnou bázi. Jedná se o variační metodu, takže energie takto získaná je vždy vyšší než opravdová hodnota energie. Schrödingerova rovnice je aproximována soustavou Hartree-Fockových rovnic:

$$F_i \psi_i = \varepsilon_i \psi_i \tag{1.9}$$

kde F_i je Fockův operátor i-tého elektronu a reprezentuje aproximaci Hamiltoniánu. ψ_i je vlnová funkce i-tého elektronu a ε_i je Langrangianeův multiplikátor i-tého elektronu. Hartree-Fockova metoda (HF metoda) zahrnuje následující iterativní proces: z množiny libovolných řešení ψ_i se vypočtou Fockovy operátory. S těmi se pak řeší Hartree-Fockovy rovnice, načež získáme druhou množinu řešení ψ_i. Tato řešení jsou použita v další iteraci. Hartree-Fockova metoda tak postupně vylepšuje

jednotlivá řešení, což vede k nižším a nižším celkovým energiím. Tento proces je opakován, až se dosáhne bodu, kdy již energie neklesá.

HF metoda podává výborné výsledky při hledání **optimálních geometrií**. Bohužel naprosto selhává při popisu mezimolekulových interakcí [66].

Teorie funkcionálu hustoty

Density Functional Theory (DFT) metody [7] jsou založeny na dvou Hohenberg-Kohnových teorémech [30]. První H–K teorém říká, že vlastnosti základního stavu mnohaelektronového systému jsou jednoznačně určeny elektronovou hustotou, která je funkcí prostorových souřadnic x, y a z.

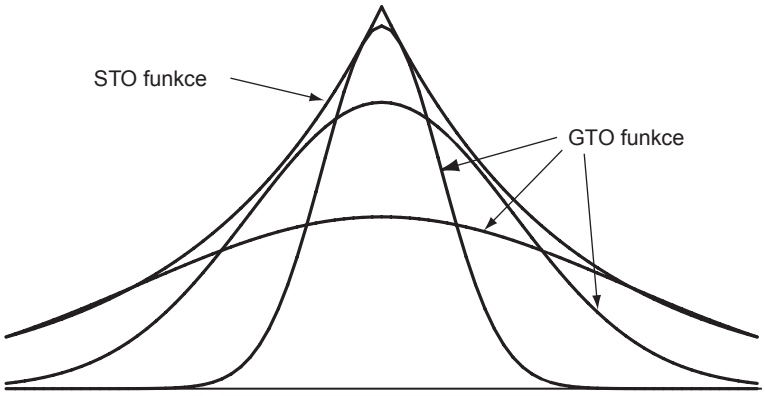
Druhý H–K teorém definuje funkcionál energie $E[\rho]$ a dokazuje, že jeho minimum odpovídá základnímu stavu molekuly, pokud se omezíme na fyzikálně přístupné elektronové hustoty. Tím DFT nahrazuje problém řešení mnohaelektronové Schrödingerovy rovnice problémem nalezení dostatečně přesné apraximace univerzálního funkcionálu elektronové hustoty. Příkladem této skupiny teorií je **BLYP** (pojmenovaná pole svých autorů Becke, Lee, Yank a Parr), její rozšíření **B3LYP** (BLYP kombinovaná s HF) a BP86 (používá starší korelační funkcionál Perdew86).

1.3.2 Bázové sady

Bázová sada [33, 66] je množina funkcí používaných k popisu tvaru atomového orbitalu. Molekulové orbitaly a celé vlnové funkce se vytvářejí **lineárními kombinacemi** (LCAO) bázových a sférických funkcí. K ab initio metodám je nutno specifikovat bázovou sadu. Ačkoliv je možné vytvořit vlastní bázovou sadu, většina kalkulací se uskutečňuje s existující bázovou sadou. Zvolená úroveň teorie a bázová sada jsou dvě hlavní kritéria určující přesnost výsledku.

Nejužívanější bázové sady obsahují funkce GTO (orbitaly Gaussova typu), případně funkce STO (orbitaly Slaterova typu). Narozdíl od STO, lze GTO integrovat analyticky, což je mnohem rychlejší než numerické integrování. Funkce STO byly využívány pro svou přesnost, té se dnes však dosahuje větším počtem GTO.

STO–3G Nejmenší bázové sady se také označují jako minimální bázové sady. Nejpopulárnější minimální bázovou sadou je STO–3G. Tahle notace značí, že tvar STO orbitalu je aproximován třemi GTO orbitaly. Minimální bázové sady se používají pro velmi velké molekuly, kvalitativní výsledky a v určitých případech i kvantitativní výsledky.



Obr. 1.1: Aproximace STO pomocí tří GTO [66].

6–31G Tato notace značí, že každý vnitřní orbital je popsán šesti GTO a každý valenční orbital je popsán množinami bázových funkcí (jedna obsahující tři GTO a druhá jeden). Analogy jsou 3–21G nebo 6–311G.

6–31G* Bázové sady označené hvězdičkou obsahují navíc ještě jednu Gaussovu funkci (polarizační funkci) pro všechny atomy kromě vodíku. Polarizační funkce umožňuje, aby mohla vlnová funkce pružněji měnit tvar.

6–31+G* Znaménko plus značí, že všem nevodíkovým atomům byly přidány difúzní funkce. Jedná se o funkce s malým exponentem a umožňují popis tvaru vlnové funkce daleko od jádra. Používají se především pro anionty, protože jejich elektrony jsou rozprostřeny dál od jader.

1.3.3 Populační analýza

Kvantová mechanika poskytuje informace o molekulových orbitalech (o jejich vlnových funkcích). Čtverec vlnové funkce popisuje elektronovou hustotu v orbitalu, tedy pravděpodobnost výskytu elektronů v definované jednotce objemu. Paricální atomové náboje mohou být vyjádřeny elektronovou hustotou náležící každému atomu v molekule.

Populační analýza (PA) rozděluje elektronovou hustotu lokalizovanou v molekulových orbitalech mezi jednotlivé atomy v molekule. Nejobtížnějším úkolem během populační analýzy je rozdělit elektronovou hustotu molekulových orbitalů patřících dvojici atomů (takzvané vazebné orbitaly) mezi dva atomy.

Pro dělení elektronové hustoty a její distribuci mezi atomy bylo navrženo mnoho schémat [55, 66]:

Mullikenova populační analýza (MPA) Tato populační analýza je nejstarší a nejvíce používaná. MPA rozděluje populace vazebného orbitalu rovnoměrně mezi atomy účastnící se vazby. Tento přístup je velmi zjednodušený a nebere v úvahu, že jeden z vázaných atomů může přitahovat elektrony markatněji než druhý. Na druhou stranu, jednoduchost MPA je občas výhodou, protože může být použita bez potíží.

Přirozená populační analýza NPA (z anglického Natural population analysis) je rozšíření Löwdinovy populační analýzy [55] a využívá explicitně ortogonální (přirozené) atomové orbitaly. Tato metoda zachází s populací vazebného orbitalu z matematického hlediska. Nevýhodou NPA je, že v některých zvláštních případech dává nefyzikálně velké náboje.

Analýza dle elektrostatického potenciálu (ESP) je postavena na zcela odlišném principu – náboje jsou fitovány tak, aby co nejlépe odpovídaly elektrostatickému potenciálu molekuly.

1.4 Zápis molekuly v počítači

Pro účely počítačového zpracování je nutno molekuly vyjadřovat pomocí modelu. Existuje více typů takových modelů, přičemž každý zavádí jistá zjednodušení a popisuje molekulu na určité úrovni abstrakce.

1.4.1 1D struktura molekuly

1D (jednorozměrná) struktura uchovává pouze informace o tom, jaké atomy se v molekule nacházejí a v jakém počtu. Vyjadřuje se molekulovým vzorcem obecného tvaru $E_i n_i$, kde E_i je značka prvku a n_i je počet atomů tohoto prvku v molekule. Například $C_6H_{12}O_6$ značí molekulu, která se skládá z šesti atomů uhlíků, dvanácti atomů vodíku a šesti atomů kyslíku. Tento vzorec však už nic neříká o vazbách. Tato 1D struktura je tak totožná pro všechny aldohexosy, ketohexosy (např. glukosa a fruktosa, dohromady 32 sacharidů) a další molekuly.

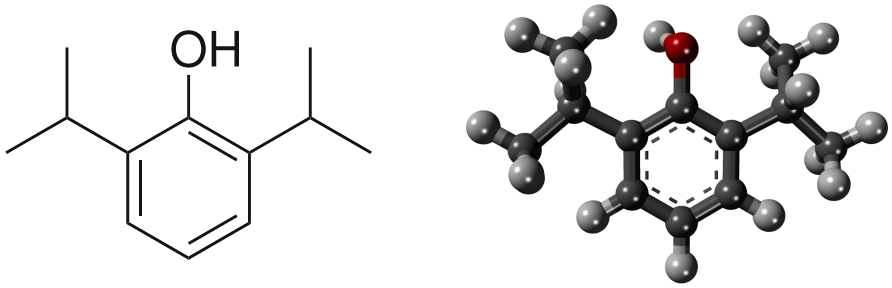
1.4.2 2D struktura molekuly

2D struktura nebo také topologie [49] přidává k jednorozměrné struktuře informaci o vazbách mezi atomy. Je popsána strukturním vzorcem, který je grafickou reprezentací molekuly. Vazby bývají znázorněny násobnými čarami (podle násobnosti vazby). Krom základních se můžeme setkat i se sofistikovanými strukturními vzorci, které

pro zpřehlednění neuvádějí vodíky vázané na uhlících a uhlíky samotné vyobrazují jako zlom (viz obr. 1.2).

1.4.3 3D struktura molekuly

3D struktura, která je také označována jako geometrie molekuly, obsahuje vzhledem ke 2D struktuře navíc informace o umístění každého atomu v prostoru. Tyto informace mohou být zadány pomocí kartézských souřadnic nebo pomocí interních souřadnic (tzv. Z-matice).



Obr. 1.2: 2D a 3D struktura propofolu (anestetikum).

1.5 Predikce 3D struktury z 2D struktury

Automatické metody predikce 3D struktury na základě 2D struktury využívají čtyři základní typy algoritmů [21]: Metody založené na pravidelch a datech, metody pracující s fragmenty, metody využívající konformační analýzu a numerické metody. Vygenerované 3D struktury mohou být dále upřesněny pomocí optimalizací.

1.5.1 Metody založené na pravidelch a datech

Tyto metody jsou založeny na znalostech chemiků, týkajících se geometrických a energetických pravidel a principů konstrukce 3D struktur molekul. Popsané znalosti byly získány na základě experimentálních dat nebo s využitím teoretického výzkumu (např. kvantové mechaniky). Uvedené znalosti jsou zabudovány do softwarových nástrojů v explicitní formě (tedy jako pravidla) nebo v implicitní formě (např. data ohledně povolených konformací cyklu). Metody založené na pravidlech a datech jsou implementovány např. v softwarových nástrojích Wizard [38], CONCORD [59] a CORINA [22].

1.5.2 Metody pracující s fragmenty

Metody pracující s fragmenty využívají informace z experimentálně získaných 3D struktur molekul. Konkrétně, tyto metody konstruují struktury molekul z fragmentů, jejichž 2D struktury jsou pokud možno co nejvíce podobné konstruované molekule. Dané metody se také snaží pracovat s co možná největšími fragmenty. Softwarové nástroje využívající tyto metody obsahují databázi fragmentů a implementují množinu pravidel pro sestavování těchto fragmentů. Uvedené metody využívají např. softwarové nástroje AIMB [27] a X-Chem [17].

1.5.3 Metody využívající konformační analýzu

Nástroje pro generování 3D struktury na základě konformační analýzy se překrývají s programy pro generování konformerů, protože oba tyto typy softwarů využívají podobné algoritmy. Nejběžnějšími metodami využívajícími konformační analýzu jsou systematické metody, náhodnostní metody, genetické algoritmy a simulační přístupy. Všechny tyto metody mohou být využity buď k vyhledání globálního minima v rámci konformačního prostoru molekuly (a tedy ke konstrukci 3D struktury molekuly) nebo k vygenerování všech konformerů s nízkou energií.

1.5.4 Numerické metody

Tyto metody využívají kvantově mechanické výpočty (QM), molekulově mechanické výpočty (MM) a algoritmy pracující s geometrickými vzdálenostmi (distance geometry, DG) a shrnují je do obecné numerické metodiky pro predikci 3D struktury. Pomocí DG algoritmů se nejdříve vygeneruje startovní konformace, ta se poté optimalizuje pomocí MM a dále pak pomocí QM. Tento přístup je použit např. v softwarovém balíku Open Babel [48] a rovněž v softwarovém balíku Balloon [61], který však využívá pouze DG a MM.

1.5.5 Optimalizace 3D struktury

Geometrickému uspořádání reálné chemické molekuly odpovídají nejvíce takové konformery, které mají co možná nejnižší energii. Jsou tedy minimy v rámci konformačního prostoru molekuly. Klasickou metodou optimalizace 3D struktury je její minimalizace [33]. Minimalizací 3D struktury rozumíme sestup do takového lokálního minima konformačního prostoru, které je nejblíže k naší 3D struktuře. Dalším krokem minimalizace je nahrazení vstupní 3D struktury 3D strukturou odpovídající nalezenému lokálnímu minimu. Takovéto optimalizace mohou být realizovány pomocí molekulové mechaniky (MM optimalizace) nebo kvantové mechaniky (QM optimalizace).

1.6 QSPR

QSPR [22, 39] je zkratkou pocházející z anglického výrazu *Quantitative Structure-Property Relationship*, což lze přeložit jako **kvantitativní vztah mezi strukturou a vlastností**. QSPR modely se snaží predikovat fyzikálně-chemické vlastnosti na základě dat (deskriptorů) vypočítaných ze struktury molekuly.

Výstup QSPR modelů se pak často používá k virtuálnímu screeningu [63], při návrhu léků, v rámci počítačově chemických simulací apod.

1.6.1 Deskriptory

Deskriptory [39, 57] jsou reálná čísla, která lze vypočítat na základě struktury molekuly. Rozlišujeme 1D, 2D a 3D deskriptory podle zdrojové struktury.

1D deskriptory vycházejí ze sumárního vzorce. Jedná se například o molární hmotnost, přítomnost vodíku atd.

Ukázkou 2D deskriptorů je počet vazeb, počet benzenových kruhů, počet funkčních skupin, počet postranních řetězců a podobně.

3D deskriptory se počítají na základě prostorové struktury. Mohou to být například vazebné úhly v molekule, informace vypočítané pomocí kvantové mechaniky (náboje, dipólový moment, spinová multiplicita, ...) a informace týkající se povrchu molekuly.

1.6.2 QSPR modely a jejich parametrizace

Klasický QSPR model splňuje tyto dvě podmínky:

- Vlastnost musí být funkcí struktury, tedy i deskriptorů.
- Závislost vlastnosti na deskriptorech musí být lineární.

QSPR model [22, 39] je vícerozměrná funkce $\mathbb{R}^n \to \mathbb{R}$, která vyjadřuje fyzikálně-chemickou vlastnost P jako lineární funkci n různých strukturních deskriptorů D_i s koeficienty c_i odlišujícími jejich relativní význam:

$$P = \sum_{i=1}^{n} c_i D_i + konstanta \tag{1.10}$$

Metoda výpočtu parametrů c_i se nazývá **parametrizace modelu**. Parametrizace se provádí na tréninkové sadě, což je sada molekul, u kterých predikovanou vlastnost P známe.

Vícerozměrná lineární regrese

Nejběžnější metodou parametrizace je vícerozměrná lineární (multilineární) regrese [25]. Je to technika, která slouží k popisu lineární závislosti dvou nebo více proměnných tím, že stanovuje parametry pro regresní model.

Vstupní rovnice pro parametrizaci vypadá obecně takto [36]:

$$M(x^i, p) = \sum_{j=1}^{n} p_j \cdot x_j^i \tag{1.11}$$

kde M je model, n počet souřadnic měření, x^i vektor souřadnic měření a p jsou vektory parametrů. V případě QSPR modelů jsou souřadnicemi měření deskriptory. Pro modely musí platit:

$$M(x^i, p) \sim y^i \tag{1.12}$$

kde y^i jsou hodnoty naměřené pro souřadnice měření x^i. Parametry získáme řešením soustavy rovnic $\mathbf{A} \cdot p = b$, kde pro členy matice $\mathbf{A}$ a vektoru b platí [36]:

$$a_{kl} = \sum_{i=1}^{m} x_k^i \cdot x_l^i \tag{1.13}$$

$$b_k = \sum_{i=1}^{m} y^i \cdot x_k^i \tag{1.14}$$

kde m je počet měření (tedy v našem případě počet molekul tréninkové sady) a $k, j \in \{1, \ldots, n\}$.

1.6.3 Validace QSPR modelu

Klíčová hodnota popisující kvalitu modelu je druhá mocnina Pearsonova korelačního koeficientu R^2 [24].

K výpočtu druhé mocniny Pearsonova korelačního koeficientu slouží vztah:

$$R^2 = \frac{[\sum_{i=1}^{n}(P_i^{calc} - \bar{P}^{calc}) \cdot (P_i^{exp} - \bar{P}^{exp})]^2}{\sum_{i=1}^{n}(P_i^{calc} - \bar{P}^{calc})^2 \cdot \sum_{i=1}^{n}(P_i^{exp} - \bar{P}^{exp})^2} \tag{1.15}$$

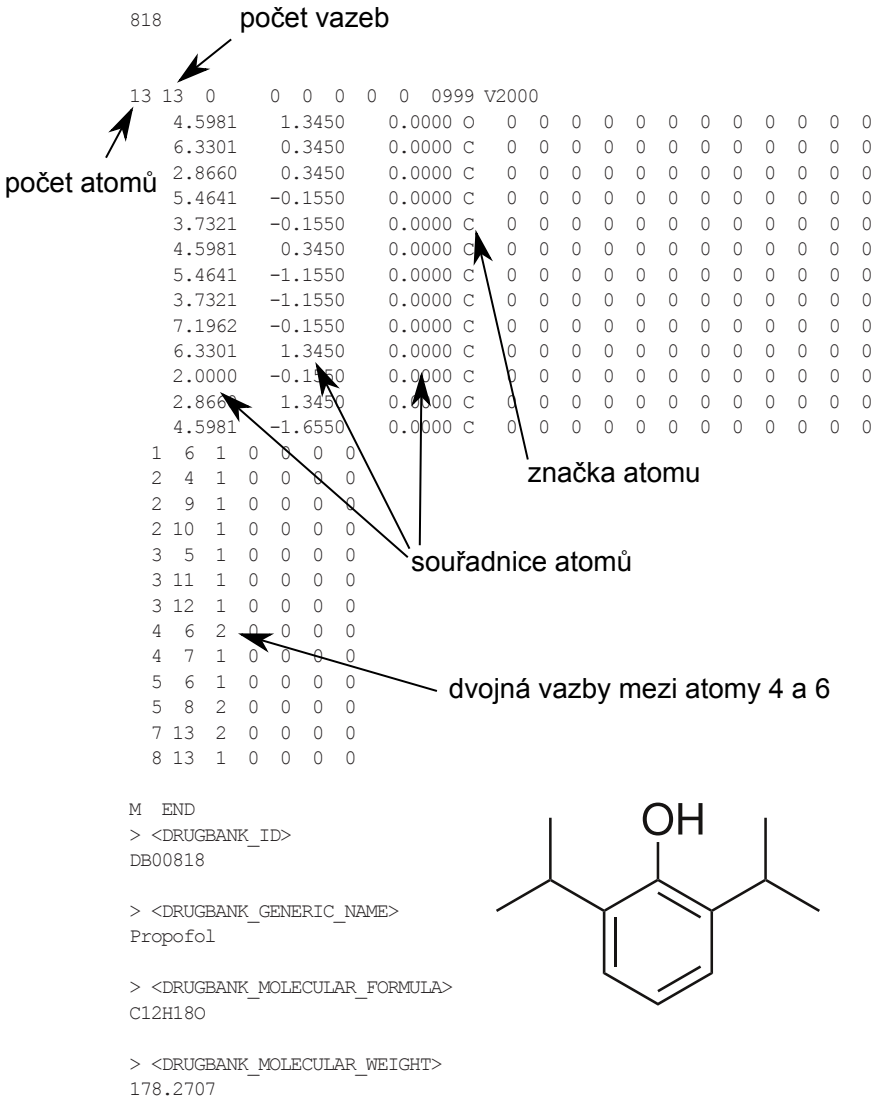
kde $\bar{P}^{calc}$ je průměrná hodnota P_i^{calc} a $\bar{P}^{exp}$ je průměrná hodnota P_i^{exp}.

2 METODY

2.1 Použité datové formáty

2.1.1 Formát SDF

Formát SDF [1] byl navržen společností SYMYX (nyní Accelrys) pro ukládání chemoinformatických dat o molekulách. SDF soubor je uložen v klasickém textovém formátu, což usnadňuje manipulaci s ním. SDF soubory obsahují dvě části: První je část MOL obsahující tzv. spojovací tabulku (CTAB [39]), která obsahuje informace o struktuře a vlastnostech skupiny atomů (např. značka prvku). Druhá část ukládá doplňující informace o molekule. Příklad takového souboru je zobrazen na obrázku 2.1.




```
818
 13 13  0     0  0  0  0  0  0999 V2000
    4.5981    1.3450    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.3301    0.3450    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.8660    0.3450    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.4641   -0.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7321   -0.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.5981    0.3450    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.4641   -1.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7321   -1.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    7.1962   -0.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.3301    1.3450    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.0000   -0.1550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.8660    1.3450    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.5981   -1.6550    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  6  1  0  0  0  0
  2  4  1  0  0  0  0
  2  9  1  0  0  0  0
  2 10  1  0  0  0  0
  3  5  1  0  0  0  0
  3 11  1  0  0  0  0
  3 12  1  0  0  0  0
  4  6  2  0  0  0  0
  4  7  1  0  0  0  0
  5  6  1  0  0  0  0
  5  8  2  0  0  0  0
  7 13  2  0  0  0  0
  8 13  1  0  0  0  0
M  END
> <DRUGBANK_ID>
DB00818

> <DRUGBANK_GENERIC_NAME>
Propofol

> <DRUGBANK_MOLECULAR_FORMULA>
C12H18O

> <DRUGBANK_MOLECULAR_WEIGHT>
178.2707
```

Obr. 2.1: Ukázka souboru ve formátu SDF.

2.1.2 Notace SMILES

Notaci Simplified Molecular-Input Line-Entry Specification (SMILES) vytvořil v roce 1986 americký vědec David Weininger [39]. Jako v ostatních lineárních notacích, je ve SMILES struktura chemickcýh sloučenin reprezentována písmeny a čísly. Na rozdíl od nomenklatury IUPAC je tato notace čitelná i počítači a na rozdíl od notace IUPAC InChI [1] je čitelná i lidmi. Základní pravidla SMILES jsou:

1. Atomy jsou reprezentovány svými chemickými značkami.
2. Atomy vodíku se nezapisují.
3. Značky sousedících atomů jsou vedle sebe.
4. Větvení se značí závorkami.
5. Dvojné a trojné vazby jsou značeny „=“, respektive „#“.
6. Kruhy jsou popsány přiřazením číslic dvěma „spojujícím“ atomům.

2.2 Databáze NCI

Molekuly analyzované v této práci byly získány z NCI Databáze [46]. Databáze byla založena a je udržována Národním ústavem pro rakovinu[2] v USA. Tato databáze shromažďuje molekuly léků, které byly v minulosti alespoň jednou testovány proti rakovině, tedy většinu známých organických molekul.

Databáze NCI je velice populární [62], a proto je často používaná pro analýzy ve výpočetní chemii a chemoinformatice. Pro tuto práci byla použita poslední verze (Release 3 ze září 2003) databáze, která obsahuje 260 071 molekul ve formátu SDF. Pro všechny molekuly jsou k dispozici 2D i 3D strukturní informace. 3D struktury byly vygenerovány z 2D struktur softwarem CORINA verze 2.6.

2.3 Databáze Physprop

Hodnoty pK_a analyzovaných molekul byly získány z Physical Properties Database (Physprop) [31], která byla vytvořena společností SRC sídlící New Yorku. Tato databáze obsahuje chemické struktury, názvy a fyzikálně-chemické vlastnosti přes 41 tisíc chemikálií. Fyzikální vlastnosti jsou shromažďovány z rozličných zdrojů a obsahují experimentální, extrapolované a odhadované hodnoty teploty tání, teploty varu, rozpustnosti ve vodě, rozdělovacího koeficientu oktanol–voda, tlaku nasycených par, pK_a, Henryho konstanty a rychlostní reakce s hydroxylovým aniontem v atmosféře.

[1] International Chemical Identifier

[2] Developmental Therapeutics Program Division of Cancer Treatment, National Cancer Institute, Rockville, USA

2.4 Softwarový balík Gaussian

Atomové náboje a struktury optimalizované kvantovou mechanikou byly získány pomocí softwarového balíku Gaussian 09 [20]. Gaussian je výpočetně chemický software, který byl poprvé vydán v roce 1970 Johnem Poplem a jeho výzkumnou skupinou na Carnegie-Mellon University. Od té doby je pravidelně aktualizován.

Gaussian je používán chemiky, chemickými inženýry, biochemiky a fyziky pro výzkum ve stávajících i nově vznikajících oblastech vědy. S využitím zákonů kvantové mechaniky predikuje Guasian energie, atomové náboje, molekulové struktury, vibrační frekvence a další vlastnosti molekul. Může být použit ke studiu molekul a jejich reakcí za velmi různorodých podmínek.

Na import a export dat tohoto programu byl použit Open Babel.

2.5 Softwarový balík Open Babel

Softwarový balík Open Babel [48] obsahuje několik programů (a knihoven) určených pro práci s molekulovými strukturami. Umožňuje čtení nebo zápis 113 formátů chemických souborů, generování 3D struktur, optimalizaci struktur molekulovou mechanikou, výpočet emprických atomových nábojů, hledat substruktury, počítat fingerprinty apod. Výhodou je otevřený zdrojový kód.

2.6 Program Balloon

Balloon [61] je program pro generování 3D struktur a sekundárně i pro převod mezi formáty (SDF, MOL2, SMILES a VBF). Algoritmus výpočtu je založen na distanční geometrii. Vygenerované struktury umožňuje přímo optimalizovat pomocí molekulové mechaniky.

2.7 Program R

Program R je open-source implementací statistického jazyka S. Tento statistický software umožňuje lineární a nelineární modelování, testování hypotéz, klasifikaci, analýzu časové řady, vykreslování grafů apod.

3 VÝSLEDKY A DISKUZE

3.1 Tréninkové sady molekul

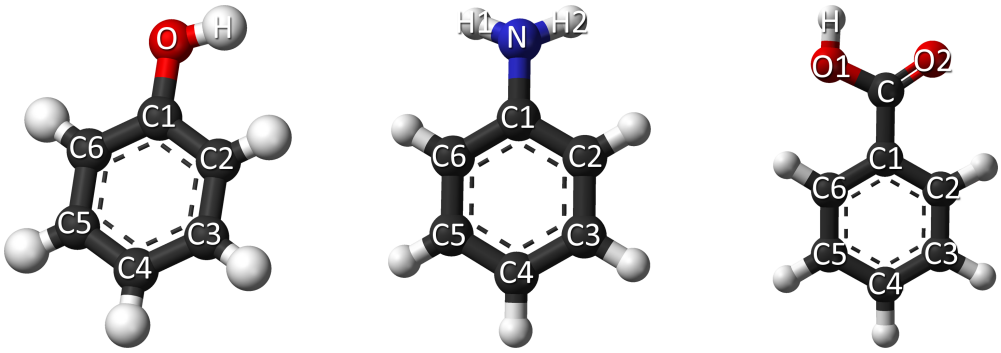
3.1.1 Studované molekuly

Tréninkové sady obsahovaly 124 molekul substituovaných fenolů, 81 molekul substituovaných anilinů a 76 molekul substituovaných benzoových kyselin. Seznam všech molekul včetně jejich struktur a pK_a je k nalezení na přiloženém CD. Vzhledem k odlišnému chování anilinů a benzoových kyselin substituovaných v *ortho* vzhledem ke stejným molekulám substituovaným v polohách *meta* a *para* bylo potřeba rozdělit molekuly do dvou podskupin (meta + para, ortho). Přehled skupin molekul a počtu jejích zástupců je uveden v tabulce 3.1.

typ molekul	meta a para	ortho
Fenoly	124	
Aniliny	41	40
Benzoové kyseliny	37	39

Tab. 3.1: Souhrné informace o tréninkových sadách.

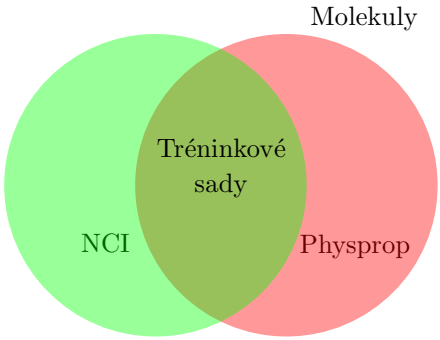
Molekuly ve všech pěti tréninkových sadách dosahovaly značné strukturní diverzity a obsahovaly širokou škálu elektron-akceptorních a elektron-donorních substituentů. Všechny molekuly byly staženy z databáze NCI pod daným identikátorem NSC v podobě 3D struktur generovaných programem CORINA.



Obr. 3.1: Číslování molekul. Zleva: fenol, anilin a kyselina benzoová.

3.1.2 Disociační konstanty

Experimentální hodnoty pK_a byly získány z databáze Physprop. Strukturní informace z databáze NCI a pK_a z databáze Physprop byly propojeny pomocí registračních čísel CAS[1].



Obr. 3.2: Výběr molekul do tréninkových sad: tréninková sada je podmnožinou průniku.

3.1.3 Konstrukce a optimalizace struktur

Struktury molekul získané z databáze NCI (tedy struktury generované programem CORINA) tvořily jednu ze skupin 3D struktur, které se dále analyzovaly. Na základě těchto 3D struktur byla programem Open Babel vygenerována další skupina 3D struktur. Konkrétně byly struktury z NCI převedeny v programu Open Babel do formátu SMILES, který uchovává 2D strukturu (topologii) molekuly. Tyto 2D struktury byly použity pro generování 3D struktur pomocí programu Open Babel.

Třetí skupina 3D struktur byla vytvořena obdobným způsobem, pouze byl pro generování 3D struktur na základě 2D struktur použit software Balloon.

Pro všechny tři takto získané skupiny 3D struktur byly vytvořeny ještě jejich molekulárně mechanicky a kvantově mechanicky optimalizované varianty. Pro molekulárně mechanickou optimalizaci byl použit software Open Babel a silové pole MMFF94. Pro kvantově mechanickou pak software Gaussian 09 s kvantovou metodu HF/6–31+G*. Po provedení výše zmíněných procesů jsme získali pro každou molekulu 9 různých 3D struktur, konkrétně: neoptimalizované struktury generované CORINou, Open Babelem a Balloonem, molekulárně mechanicky optimalizované

[1]Číslo CAS je **unikátní** identifikátor přidělený každé chemické látce Americkou chemickou společností.

struktury generované těmito třemi softwary a kvantově mechanicky optimalizované struktury generované těmito třemi softwary.

3.1.4 Atomové náboje

Pomocí programu Gaussian byly spočítány náboje pomocí dvanácti kvantově mechanických metod pro výpočet náboje. Konkrétně se jednalo o kombinace dvou úrovní QM teorie (HF a B3LYP), dvou bázových sad (STO–3G a 6–31G*) a tří populačních analýz (MPA, NPA a ESP). Pracovali jsme tedy s těmito dvanácti QM metodami: HF/STO–3G/MPA, HF/STO–3G/NPA, HF/STO–3G/ESP, HF/6–31G*/MPA, HF/6-31G*/NPA, HF/6–31G*/ESP, B3LYP/STO–3G/MPA, B3LYP/STO–3G/NPA, B3LYP/STO-3G/ESP, B3LYP/6–31G*/MPA, B3LYP/6–31G*/NPA, B3LYP/6–31G*/ESP.

3.1.5 Souhrné informace o vstupních datech

Z předchozích sekcí vyplývá, že jsme pro každou molekulu vytvořili 9 typů 3D struktur a pro každou 3D strukturu napočítali 12 typů nábojů. Pro každou molekulu jsme tedy měli k dispozici $12 \times 9 = 108$ různých sad nábojů (a tedy i 108 sad nábojových deskriptorů).

Navíc jsme pracovali s pěti typy molekul, které jsme studovali odděleně. Proto bylo nutno vytvořit $5 \times 108 = 540$ QSPR modelů. Více informací o těchto modelech uvádíme v následující sekci.

3.2 Tvorba QSPR modelů

3.2.1 Deskriptory

Tato práce testuje vliv metody generování 3D struktur a metody výpočtu nábojů na kvalitu QSPR modelů (neboli jejich schopnost co nejpřesněji predikovat pK_a). Jako deskriptory byly tedy použity parciální atomové náboje. Výběr deskriptorů byl učiněn na základě publikace [56] (Svobodová a Geidl, JCIM, 2011). Z této práce vyplývá, že s pK_a korelují náboje na atomech vzdálených od kyselého vodíku nejvýše dvě vazby. Proto jsme pro jednotlivé typy molekul využili deskriptory, uvedené v tabulce 3.2.

3.2.2 Parametrizace a validace modelů

Z vybraných deskriptorů jsme sestavili následující rovnice pro výpočet pK_a:

$$\mathrm{p}K_a(\text{fenol}) = par_{\mathrm{H}} \cdot q_{\mathrm{H}} + par_{\mathrm{O}} \cdot q_{\mathrm{O}} + par_{\mathrm{C1}} \cdot q_{C1} + konst. \qquad (3.1)$$

Typ molekul	Náboje
Fenoly	q_{H}, q_{O} a q_{C1}
Aniliny	q_{H1}, q_{H2}, q_{N} a q_{C1}
Benzoové kyseliny	q_{H}, q_{O1}, q_{O2} a q_{C}

Tab. 3.2: Nábojové deskriptory použité pro tvorbu QSPR modelů.

$$\mathrm{p}K_a(\text{anilin}) = par_{\mathrm{H1}} \cdot q_{\mathrm{H1}} + par_{\mathrm{H2}} \cdot q_{\mathrm{H2}} + par_{\mathrm{N}} \cdot q_{\mathrm{N}} + par_{\mathrm{C1}} \cdot q_{C1} + konst. \quad (3.2)$$

$$\mathrm{p}K_a(\text{benz.kys.}) = par_{\mathrm{H}} \cdot q_{\mathrm{H}} + par_{\mathrm{O1}} \cdot q_{\mathrm{O1}} + par_{\mathrm{O2}} \cdot q_{\mathrm{O2}} + par_{\mathrm{C}} \cdot q_{C} + konst. \quad (3.3)$$

kde par_x a *konst.* jsou parametry modelu. Parametrizace QSPR modelů byla provedena pro všechny získané náboje metodou vícerozměrné lineární regrese. Pro parametrizaci byly použity kompletní sady molekul a získaný model byl validován pro všechny molekuly v sadě.

3.2.3 Souhrn kritérií kvality modelů

Část výsledků je uvedena v následujících tabulkách a grafech. Tabulka 3.3 shrnuje vybrané Pearsonovy korelační koeficienty pro korelaci mezi experimentálními hodnotami pK_a a hodnotami pK_a vypočítanými pomocí QSPR modelů. Nejvíce relevantní korelace mezi experimentálními a vypočítanými hodnotami ukazují grafy na obrázku 3.3. Zbylá data byla pro svůj rozsah umístěna na CD a do přílohy B.

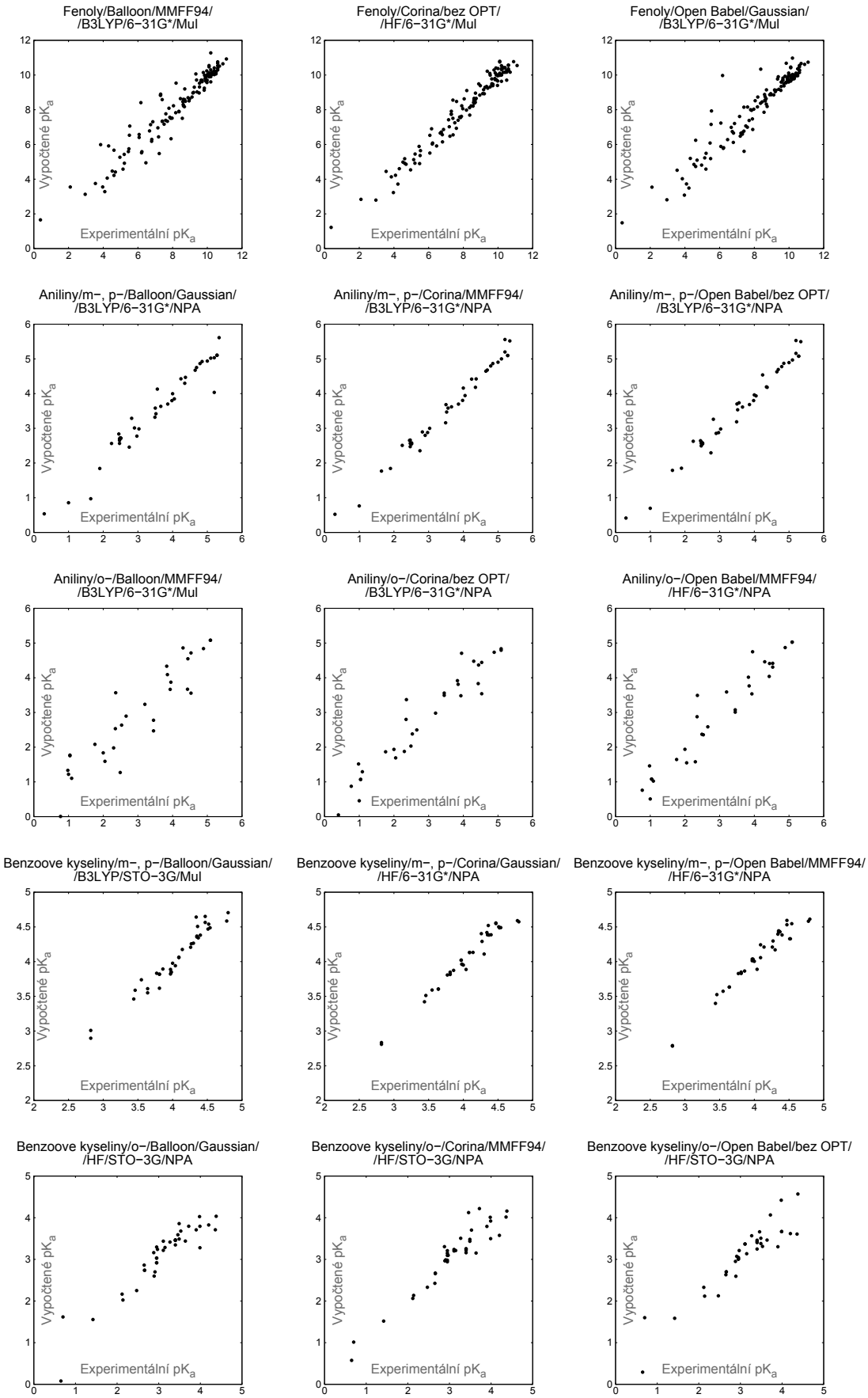
3.3 Diskuze kvality modelů

Z našich výsledků vyplývá, že kvantově mechanické náboje jsou obecně velmi kvalitními deskriptory pro výpočet pK_a. Tento výsledek souhlasí s výsledky publikovanými v [19, 24]. Tato práce také prokazuje, že automaticky generované 3D struktury je možno úspěšně použít pro predikci pK_a. Konkrétně 200 z 540 (tedy 37 %) QSPR modelů dosahovalo $R^2 > 0{,}9$ a pouze méně než 30 % dosahovalo nedostatečné korelace ($R^2 < 0{,}8$) . To jasně dokazuje, že tuto metodu predikce pK_a lze **používat v praxi**.

Otázkou zůstává, která metoda pro generování 3D struktur a která metoda pro výpočet nábojů je k tomuto účelu nejvhodnější.

3.3.1 Vliv softwaru pro generování 3D struktury

Vliv softwaru pro generování 3D struktury je možno analyzovat na základě tabulky 3.3. Z uvedené tabulky vyplývá, že všechny tři programy (Balloon, CORINA a Open



Obr. 3.3: Vybrané grafy.

Aniliny m-, p-		B3LYP/6-31G*		HF/6-31G*	
		MPA	NPA	MPA	NPA
Balloon	bez opt	0,7627	0,7987	0,7445	0,8062
	MM opt	0,8699	0,8698	0,8345	0,8632
	QM opt	0,9372	0,9452	0,9436	0,9427
CORINA	bez opt	0,9698	0,9687	0,9782	0,9800
	MM opt	0,9663	0,9836	0,9486	0,9765
	QM opt	0,9429	0,9518	0,9558	0,9494
Openbabel	bez opt	0,9681	0,9775	0,9416	0,9571
	MM opt	0,9541	0,9653	0,9274	0,9415
	QM opt	0,9452	0,9497	0,9489	0,9466

Aniliny o-		B3LYP/6-31G*		HF/6-31G*	
		MPA	NPA	MPA	NPA
Balloon	bez opt	0,9126	0,8589	0,9089	0,9054
	MM opt	0,9451	0,9240	0,9087	0,9073
	QM opt	0,9446	0,9431	0,9055	0,9208
CORINA	bez opt	0,9176	0,9628	0,9381	0,9728
	MM opt	0,9289	0,9643	0,9223	0,9715
	QM opt	0,9265	0,9224	0,8888	0,9110
Openbabel	bez opt	0,9326	0,9580	0,9136	0,9690
	MM opt	0,9374	0,9651	0,9194	0,9748
	QM opt	0,9340	0,9295	0,8956	0,9166

Benzoové kys. m-, p-		B3LYP/6-31G*		HF/6-31G*	
		MPA	NPA	MPA	NPA
Balloon	bez opt	0,8370	0,8627	0,8895	0,9000
	MM opt	0,8231	0,8740	0,9192	0,9019
	QM opt	0,9016	0,9493	0,9309	0,9443
CORINA	bez opt	0,9450	0,9524	0,9493	0,9486
	MM opt	0,9356	0,9420	0,9494	0,9579
	QM opt	0,9571	0,9535	0,9665	0,9677
Openbabel	bez opt	0,9473	0,9517	0,9506	0,9619
	MM opt	0,9478	0,9526	0,9485	0,9614
	QM opt	0,9598	0,9602	0,9648	0,9696

Benzoové kys. o-		B3LYP/6-31G*		HF/6-31G*	
		MPA	NPA	MPA	NPA
Balloon	bez opt	0,6319	0,6000	0,6976	0,6735
	MM opt	0,6922	0,6574	0,7959	0,7075
	QM opt	0,7644	0,8365	0,8107	0,8520
CORINA	bez opt	0,7547	0,8539	0,7747	0,8484
	MM opt	0,8550	0,8685	0,8798	0,8855
	QM opt	0,9103	0,9057	0,9190	0,9157
Openbabel	bez opt	0,7814	0,8001	0,7673	0,8503
	MM opt	0,7811	0,8043	0,7665	0,8556
	QM opt	0,7587	0,8022	0,7748	0,7558

Fenoly		B3LYP/6-31G*		HF/6-31G*	
		MPA	NPA	MPA	NPA
Balloon	bez opt	0,5717	0,4177	0,6525	0,5464
	MM opt	0,9112	0,8412	0,9095	0,8880
	QM opt	0,9022	0,8867	0,8926	0,8918
CORINA	bez opt	0,9617	0,9582	0,9656	0,9622
	MM opt	0,9354	0,7894	0,9358	0,8530
	QM opt	0,9438	0,9349	0,9439	0,9387
Openbabel	bez opt	0,9019	0,9072	0,8611	0,9054
	MM opt	0,9023	0,9061	0,8598	0,9043
	QM opt	0,9055	0,8943	0,8780	0,8922

Legenda	R^2
excelentní	0,950–0,990
velice dobré	0,920–0,950
dobré	0,900–0,920
akceptovatelné	0,850–0,900
slabé	0,800–0,850

Tab. 3.3: Tabulka R^2 vybraných modelů.

Babel) lze úspěšně použít pro predikci pK_a. Nejlepší výsledky byly dosaženy s programy CORINA a Open Babel. Přes diametrální odlišnost algoritmů programů CORINA a Open Babel poskytují oba velmi podobné výsledky. Struktury navrhované programem Balloon poskytují bez optimalizace velmi slabé modely. K objektivnímu porovnání slouží následující tabulka s průměrnými hodnotami R^2.

3.3.2 Vliv optimalizace

Jak je patrno z tabulky 3.3, vliv optimalizace byl nejmarkantnější v případě struktur generovaných softwarem Balloon. Již optimalizace metodami molekulové mechaniky zlepšila výsledky průměrně o 23 %. Optimalizace kvantovou mechanikou zvýšila

software	Balloon	CORINA	Open Babel
průměrné $\mathbf{R^2}$	0,7556	0,8850	0,8818

Tab. 3.4: Průměrné hodnoty R^2 pro všechny QSPR modely, využívající 3D struktury generované určitým softwarem.

průměrné R^2 o dalších 10 %.

Na struktury pocházející z programu Open Babel měla naopak optimalizace molekulovou mechanikou vliv zcela minimální – průměrné R^2 se snížilo o 0,01 %. Tento výsledek jen potvrzuje, že struktury generované programem Open Babel jsou ihned optimalizovány. Optimalizace kvantovou mechanikou pak přinesla zlepšení o 1,5 %.

V případě 3D struktur generovaných programem CORINA je vliv optimalizace sporný. U molekul fenolů a anilinů měla optimalizace jak molekulovou mechanikou, tak pomocí kvantové mechaniky, minimální či negativní efekt. U benzoových kyselin se však situace změnila – optimalizace MM zvýšila průměrné R^2 o 20 % a pouze náboje vypočítané ze struktur optimalizovaných kvantovou mechanikou poskytovaly modely s $R^2 > 0{,}9$.

Znatelnou nevýhodou optimalizace pomocí kvantové mechaniky je její výpočetní náročnost. Optimalizace jedné molekuly s použitím bázové sady 6–31+G* trvala se čtyřjádrovým procesorem pod taktem 3 GHz průměrně 1 hodinu. Výjimkou nebyl ani pět hodin trvající výpočet.

původ	**bez opt.**	**MM optimalizace**		**QM optimalizace**		celkové
	$\bar{R}^2$	$\bar{R}^2$	% ↑↓	$\bar{R}^2$	% ↑↓	zlepšení
Balloon	0,6377	0,7858	23,21 %	0,8697	10,67 %	36,37 %
CORINA	0,8598	0,8716	1,37 %	0,9062	3,97 %	5,40 %
Open Babel	0,8570	0,8569	−0,01 %	0,8698	1,50 %	1,49 %

Tab. 3.5: Shrnutí vlivu optimalizace.

3.3.3 Vliv kvantově mechanické metody

Vliv použité QM metody vyplývá z tabulky 3.3. Rozdíly jsou sice malé, ale drží se trendu. Náboje počítané pomocí metody Hartree-Fockovy korelovaly s pK_a téměř vždy více než při použití metody B3LYP (průměrné R^2 bylo o 2 % vyšší, viz tabulka 3.6). Kalkulace s HF jsou navíc při použití bázové sady 6–31G* cca o třetinu rychlejší proti B3LYP.

úroveň teorie	HF	B3LYP
průměrné R^2	0,8495	0,8203

Tab. 3.6: Srovnání průměrných R^2 pro obě úrovně teorie.

3.3.4 Vliv bázové sady

Dle očekávání poskytuje vhodnější náboje pro QSPR modelování bázová sada 6–31G*. Proti minimální bázové sadě STO–3G je to však rozdíl velmi malý (v průměru o 5 % vyšší R^2). Výhodou bázové sady STO–3G je však cca 30× nižší výpočetní náročnost než v případě bázové sady 6–31G*.

úroveň teorie	6–31G*	STO–3G
průměrné R^2	0,8414	0,8039

Tab. 3.7: Srovnání průměrných R^2 pro použité bázové sady.

3.3.5 Vliv populační analýzy

Mullikenova PA i NPA poskytují atomové náboje, které jsou vhodné pro predikci pK_a. Rozdíly mezi nimi jsou velmi malé a liší se s každou tréninkovou sadou. Dle průměrných hodnot R^2 je MPA o 2 % přesnější. ESP náboje mají pouze slabou korelaci s pK_a.

populační analýza	MPA	NPA	ESP
průměrné R^2	0,8510	0,8337	0,7546

Tab. 3.8: Srovnání průměrných R^2 pro různé populační analýzy.

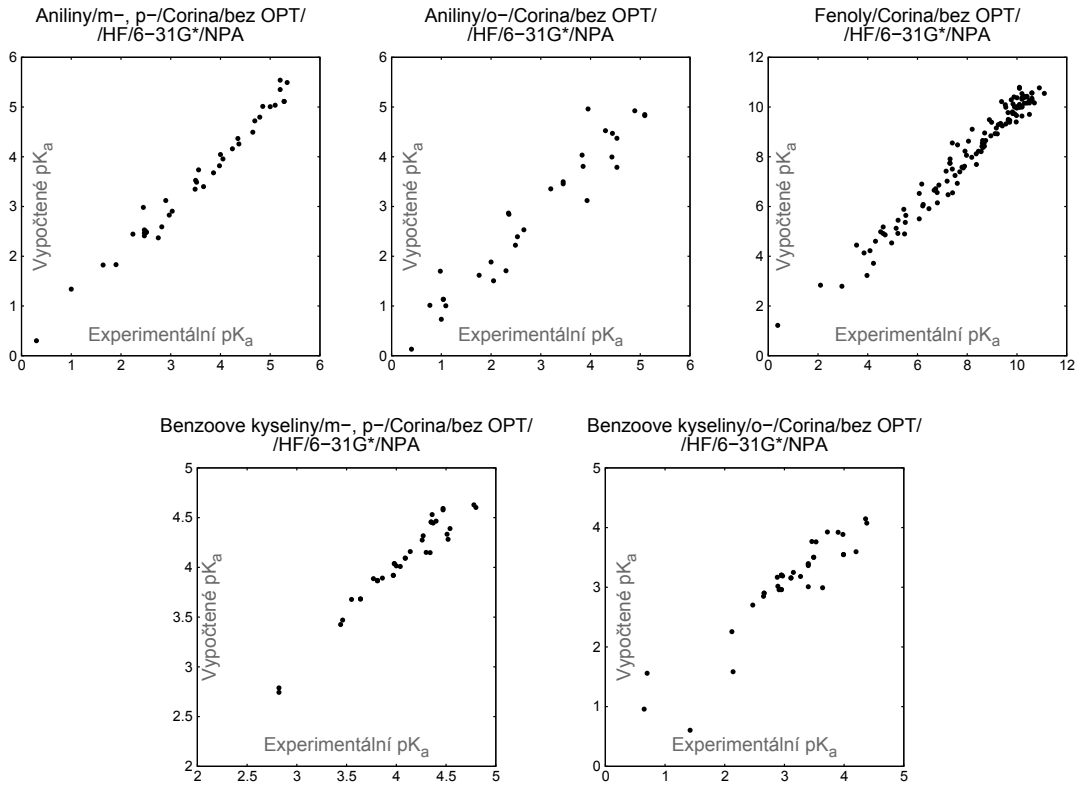
3.3.6 Shrnutí

Provedená studie prokázala, že **nejvhodnější** série metod je použít software CORINA ke generování struktur, vytvořené struktury neoptimalizovat a náboje počítat kombinací metod HF/6–31G*/NPA. MPA totiž sice vychází v průměru lépe, ale v kombinaci s výše uvedeným přístupem (CORINA, bez optimalizace, HF/6–31G*) poskytuje lepší výsledky NPA (R^2 0,9424 vs. 0,9393). Informace o zvolené kombinaci metod jsou uvedeny v tabulce 3.9.

Modely získané touto sérií metod dosahují průměrného R^2 0,9424. Přesnost dobře ilustrují grafy na obrázku 3.4.

proces	metoda
konstrukce	CORINA
optimalizace	—
úroveň teorie	Hartree–Fock
bázová sada	6–31G*
populační analýza	NPA

Tab. 3.9: Popis nejvhodnějších metodik, které je nutno využít při predikci pK_a pomocí QSPR modelů.



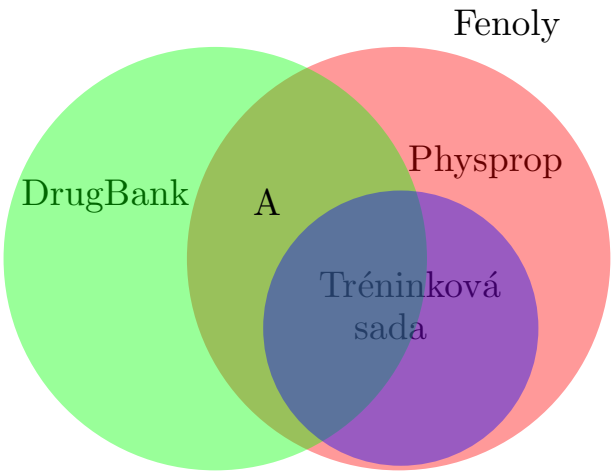
Obr. 3.4: Grafy ukazující korelaci mezi experimentálním a vybranou metodou vypočítaným pK_a.

Místo NPA by se dala rovněž použít Mullikenova populační analýza a úroveň teorie Hartree–Fock lze zaměnit za B3LYP, rozdíl v R^2 je pouze okolo 1 %. Použití optimalizace kvantovou mechanikou také zaručuje vynikající modely, je však velmi časově náročné.

3.4 Testovací datová sada – molekuly léků

Praktickou použitelnost metody predikce pK_a, vytvořené v rámci této práce a popsané v předchozím textu, lze nejlépe ukázat na testech molekul mimo tréninkovou sadu. Vzhledem k důležitosti znalosti pK_a ve farmacii bude analýza provedena na molekulách využívaných jako léky.

Prohledáním databáze DrugBank byly nalezeny 3 molekuly fenolů, pro které je dostupné pK_a v databázi Physprop a zároveň se nenacházejí v tréninkové sadě fenolů. Jmenovitě se jedná o dronabinol, levorfanol a pentazocin.



Obr. 3.5: Výběr molekul do testovací sady: $A = \text{DrugBank} \cap \text{Physprop} \setminus \text{Tréninková sada}$, kde A je množina třech nalezených fenolů. Velikost překryvu je tak opět pouze ilustrační.

3.4.1 Dronabinol

Dronabinol je INN[2] název pro synteticky připravený tetrahydrocannabinol. Prodává se v USA a některých dalších zemích (např. Německo, Anglie, Izrael) [47] pod obchodním názvem Marinol®. V ČR je jeho distribuce zakázana. Využívá se proti a zvracení [52], nechutenství [8] a k tišení bolesti [14, 47].

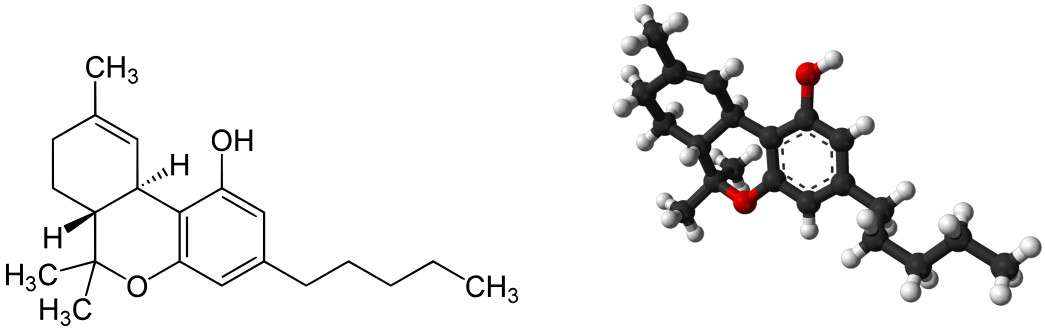
Při pohledu na strukturu molekuly dronabinolu (obr. 3.6) je zřejmé, že se jedná o látku hydrofobního charakteru a překonání hematoencefalické bariéry[3] tak nečiní

[2] International Nonproprietary Name

[3] Odděluje prostředí mozku od cévního systému. Lipofobní látky je nutné přenášet aktivně.

problémy. Přítomnost fenolické hydroxylové skupiny je důležitá kvůli volným elektronovým párům kyslíku, který tak slouží jako donor vodíkové vazby. Studie ukázaly, že donorem vodíkové vazby u těchto receptorů může být i aminoskupina [42, 60].

Experimentální hodnota pK_a dronabinolu je 10,6 (získáno z databáze Physprop), což znamená, že za fyziologického pH není dronabinol disociován ani z jedné tisíciny.



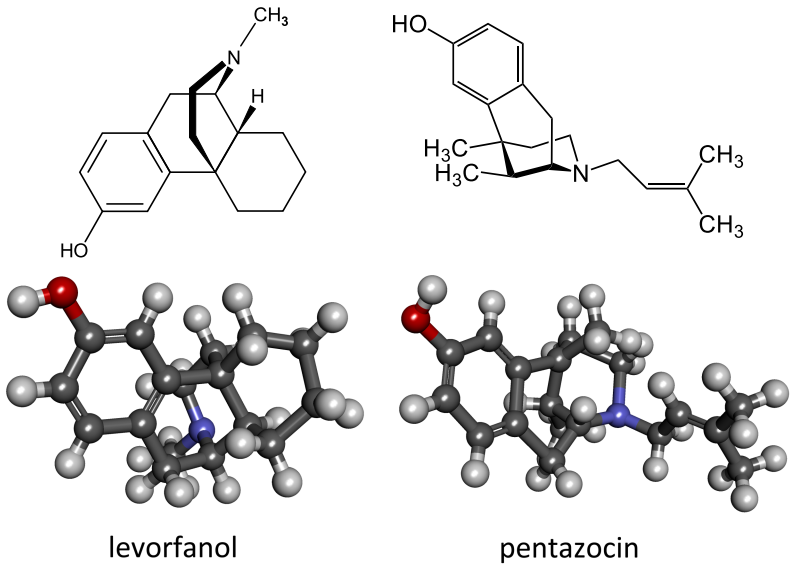
Obr. 3.6: 2D a 3D struktura molekuly dronabinolu.

3.4.2 Levorfanol a pentazocin

Levorfanol i pentazocin jsou opioidní analgetika. Levorfanol je prodávaný v zahraničí pod názvem Levo–Dromoran, v ČR se lék nepoužívá [3]. Pentazocin je v ČR dostupný na předpis pod obchodním názvem Fortral a používá s k tlumení mírných až těžkých bolestí.

Tyto molekuly obsahují několik polárních skupin, a hůře tak zdolávají hematoencefalickou bariéru. Ve fenolické hydroxylové skupině je nejdůležitější vodík, který se na receptor váže vodíkovou vazbou a fenolová skupina tak pro správný analgetický účinek nesmí být disociována. pK_a této skupiny se pohybuje v rozmezí 9 - 11 [34].

Nejdůležitější skupinou je však terciární amin, který se při vazbě na záporně nabitý receptor váže protonovaný. pK_a této skupiny se pohybuje v rozmezí 8 - 10 [50], což znamená, že celková hodnota pK_a je řízena především touto skupinou. Pomocí nábojů na fenolické skupině tak sice nemůžeme spočítat první pK_a, jak bylo původně zamýšleno, ale můžeme spočítat druhé pK_a (pK_{a2}). V databázi Physprop jsou dostupné pouze hodnoty pK_{a1}, proto bylo potřeba hodnoty pK_{a2} hledat v jiných zdrojích. Hodnotu pK_{a2} se podařilo získat pro pentazocin [10], a tato hodnota je 10,35. V případě levorfanolu se hodnotu pK_{a2} získat nepodařilo, protože ji není možno změřit [34]. O experimentální hodnotě pK_{a2} této molekuly je tedy známo pouze, že se nachází v intervalu 9 - 11 [34].



Obr. 3.7: 2D a 3D struktury molekul levorfanolu a pentazocinu.

3.5 Predikce pK_a pro molekuly léků

Pro predikci pK_a těchto molekul byla použita nejúspěšnější z našich pK_a predikčních metodik, popsaných a analyzovaných v předchozím textu (viz sekce 3.3.6). Konkrétně byl použit software CORINA ke generování struktur, které nebyly nijak optimalizovány a na kterých byly náboje spočítány pomocí metody HF/6–31G*/NPA. Predikované hodnoty jsou uvedeny v tabulce 3.10.

testovaná látka	experimentální pK_a	vypočítané pK_a	odchylka
dronabinol	10,6	9,92	0,68
levorfanol	9 - 11	10,39	—
pentazocin	10,35	10,44	0,09

Tab. 3.10: Porovnání experimentálních a predikovaných hodnot pK_a pro molekuly léků z testovací sady.

Predikované pK_a tedy velmi přesně odpovídají experimentálním hodnotám pK_a daných léků. Tyto výsledky potvrzují, že metodika popsaná v tabulce 3.9 je vhodným způsobem predikce pK_a.

3.6 Publikační činnost

Výsledky představené v této práci byly prezentovány na následujících konferencích:

- Geidl, S., Beránek, R., Svodobodová Vařeková, R., Bouchal, T., Brumovský, M., Kudera, M., Skřehota, O., Koča, J.: How the methodology of 3D structure preparation influences the quality of QSPR models? 7th German Conference on Chemoinformatics. 2011. (poster, listopad 2011)
- Beránek, R., Geidl, S., Bouchal, T., Svobodová Vařeková, R.: Výpočet pK_a na základě atomových nábojů a studie vlivu 3D stuktury na přesnost výpočtu. Studentská odborná konference Chemie a společnost. 2011. (poster, prosinec 2011)

ZÁVĚR

Predikce hodnot disociačních konstant pro dosud nesyntetizované molekuly je oblastí, která má velký význam pro farmaceutický průmysl. Na základě znalosti pK_a totiž můžeme z rozsáhlých sad molekul potenciálních léků vyloučit nevhodné molekuly, tedy molekuly příliš kyselé nebo bazické. Velmi slibnou metodikou pro predikci pK_a je aplikace QSPR modelů využívajících jako deskriptory parciální atomové náboje. Hodnoty nábojů je nutno vypočítat na základě 3D struktury molekuly. Tyto struktury nemůžeme získat experimentálně, protože pracujeme s dosud nesyntetizovanými molekulami. Proto je nutno tyto struktury vygenerovat pomocí vhodných softwarových nástrojů a poté získané struktury dále optimalizovat. Kvalita takto vytvořených struktur je klíčovým faktorem ovlivňujícím přesnost predikce pK_a. Proto jsem se v rámci své práce zaměřil právě na tuto tématiku.

Prvním krokem mé práce bylo analyzovat vhodnost využití různých softwarových nástrojů pro generování 3D struktur, metod pro optimalizaci a metod pro výpočet nábojů pro predikci pK_a pomocí QSPR modelů. Konkrétně jsem pracoval se softwarovými nástroji CORINA, Open Babel a Balloon a optimalizaci jsem buď neprováděl žádnou nebo pro ni využíval molekulovou případně pak kvantovou mechaniku. Jako metody pro výpočet nábojů jsem používal kvantově mechanické metody s úrovní teorie HF nebo B3LYP, bázemi STO–3G nebo 6–31G* a populačními analýzami NPA, MPA nebo ESP. Takto jsem vytvořil 540 QSPR modelů, vypočítal jejich kritéria kvality a vzájemně je porovnal. Výsledky těchto analýz potvrdily, že automaticky generované struktury jsou vhodnými vstupy pro predikci pK_a (37 % našich QSPR modelů mělo hodnoty $R^2 > 0,9$). Dále jsem pak na základě uvedených analýz našel nejlepší metodiku pro predikci pK_a. Tato metodika je následující: Vygenerovat 3D struktury molekul pomocí software CORINA, tyto 3D struktury dále neoptimalizovat a vypočítat pro ně náboje pomocí HF/6–31G*/NPA.

Uvedenou metodiku jsem poté otestoval v praxi – konkrétně jsem ji využil k predikci pK_a tří molekul léků (dronabinol, levorfanol a pentazocin), které nebyly součástí naší tréninkové sady. Hodnty pK_a, predikované tímto způsobem, velmi přesně odpovídaly experimentálním hodnotám pK_a daných léků. Tyto výsledky potvrzují, že metodika vyvinutá v rámci mé práce je vhodným a efektivním způsobem predikce pK_a.

Má SOČ byla realizována ve spolupráci s vědeckými pracovníky a studenty z Národního centra pro výzkum biomolekul, které je součástí Masarykovy univerzity a také projektu CEITEC. V této spolupráci plánuji pokračovat a výsledky své práce dále rozšiřovat (např. analyzovat nábojové deskriptory z disociovaných molekul, využít molekuly z databáze Pubchem, prozkoumat využitelnost empirických nábojů apod.).

Výsledky mé práce byly prezentovány na mezinárodní konferenci Chemoinformatics v Goslaru (Německo) a na studentské odborné konferenci Chemie a společnost. Rozšířenou verzi analýz bychom poté rádi publikovali v impaktovaném vědeckém časopisu Journal of chemical information and modelling.

LITERATURA

[1] *MDL CTfile Formats*. Accelrys, San Diego, CA, USA, 2010. Dostupné z: `http://download.accelrys.com/freeware/ctfile-formats/ctfile-formats.zip`.

[2] *ACD/pKa*. Advanced Chemistry Development, Inc., 110 Yonge St., 14th Floor, Toronto, Ontario, Canada M5C 1T4.

[3] ANZENBACHER, P. – JEZDINSKý, J. Léčiva a chiralita. *Klin Farmakol Farm 2003; 17: 148*. 150.

[4] ATKINS, P. – PAULA, J. *Atkins' Physical chemistry*. Oxford : Oxford University Press, 9. vydání, 2010. ISBN 9780199543373.

[5] BAJORATH, J. *Chemoinformatics: concepts, methods, and tools for drug discovery*. Methods in molecular biology. New York, NY, USA : Humana Press, 2004. ISBN 9781588292612.

[6] BARTLETT, P. – ENTZEROTH, M. *Exploiting chemical diversity for drug discovery*. RSC biomolecular sciences. London : Royal Society of Chemistry, 2006. ISBN 9780854048427.

[7] BARTOLOTTI, L. – FLURCHICK, K. An introduction to density functional theory. *Reviews in computational chemistry*. 1996, s. 187–216.

[8] BEAL, J. et al. Dronabinol as a treatment for anorexia associated with weight loss in patients with AIDS. *Journal of pain and symptom management*. 1995, 10, 2, s. 89–97.

[9] BORK, N. et al. Ab initio charge analysis of pure and hydrogenated perovskites. *Journal of Applied Physics*. 2011, 109, 3, s. 033702–033702.

[10] BRITTAIN, H. G. – FLOREY, K. *Analytical Profiles of Drug Substances and Excipients:*. Č. vol. 13 v Analytical Profiles of Drug Substances, Excipients, and Related Methodology. Waltham, Massachusetts, US : Academic Press, 1984. ISBN 9780122608230.

[11] BUNIN, B. *Applications of combinatorial synthesis to drug discovery and catalysis development*. Berkeley : University of California, 1996.

[12] BUNIN, B. *Chemoinformatics: theory, practice, & products*. New York, NY, USA : Springer, 2007. ISBN 9781402050008.

[13] BURDEN, F. – POLLEY, M. – WINKLER, D. Toward novel universal descriptors: Charge fingerprints. *Journal of chemical information and modeling*. 2009, 49, 3, s. 710–715.

[14] BURNS, T. – INECK, J. Cannabinoid analgesia as a potential new therapeutic option in the treatment of chronic pain. *The Annals of pharmacotherapy*. 2006, 40, 2, s. 251–260.

[15] CAREY, F. – SUNDBERG, R. *Advanced organic chemistry: Reactions and synthesis*. Advanced Organic Chemistry. New York, NY, USA : Springer, 2007. ISBN 9780387683546.

[16] CITRA, M. Predicting pK_a. *Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods*. 1999, 38, 1, s. 192–206. PMID: 10903100.

[17] DAVIESR, K. – UPTON, T. Experiences building and searching the Chapman & Hall Dictionary of Drugs. *Tetrahedron Computer Methodology*. 1990, 3, 6, Part C, s. 665–671. ISSN 0898-5529. doi: 10.1016/0898-5529(90)90165-5. Three-dimensional chemical structure handling.

[18] Dirac, P. A. M. Quantum Mechanics of Many-Electron Systems. *Royal Society of London Proceedings Series A*. April 1929, 123, s. 714–733.

[19] DIXON, S. L. – JURS, P. C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.* December 1993, 14, s. 1460–1467. ISSN 0192-8651. doi: 10.1002/jcc.540141208.

[20] FRISCH, M. J. et al. Gaussian 09 Revision A.1, 2009. Gaussian Inc. Wallingford CT.

[21] GASTEIGER, J. *Handbook of chemoinformatics: from data to knowledge in 4 volumes*. Č. sv. 1 v Advances in Electrochemical Sciences and Engineering Series. New York, NY, USA : Wiley-VCH, 2003. ISBN 9783527306800.

[22] GASTEIGER, J. – ENGEL, T. *Chemoinformatics: a textbook*. Bognor Regis : Wiley-VCH, 2003. ISBN 9783527306817.

[23] GASTEIGER, J. – MARSILI, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*. 1980, 36, 22, s. 3219–3228.

[24] GEIDL, S. Výpočty pK_a na základě atomových nabojů [online]. Bakalářská práce, Masarykova univerzita, Přírodovědecká fakulta, Brno, 2011. Dostupné z: `http://is.muni.cz/th/327887/prif_b/`.

[25] GRAHAM, R. *Data analysis for the chemical sciences: a guide to statistical techniques*. New York, NY, USA : Wiley, 1993. ISBN 9781560810483.

[26] GROSS, K. C. – SEYBOLD, P. G. Substituent effects on the physical properties and pK_a of phenol. *International Journal of Quantum Chemistry*. 2001, 85, 4-5, s. 569–579. ISSN 1097-461X. doi: 10.1002/qua.1525.

[27] HAHN, M. A. – WIPKE, W. T. *In Chemical Structures; Warr, W. E.,. Ed.,*. 1988, 1, s. 269–278.

[28] HILAL, S. H. – KARICKHOFF, S. W. – CARREIRA, L. A. A Rigorous Test for SPARC's Chemical Reactivity Models: Estimation of More Than 4300 Ionization pK_as. *Quantitative Structure-Activity Relationships*. 1995, 14, 4, s. 348–355. ISSN 1521-3838. doi: 10.1002/qsar.19950140405.

[29] HO, J. – COOTE, M. A universal approach for continuum solvent pK_a calculations: are we there yet? *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*. 2010, 125, s. 3–21. ISSN 1432-881X. 10.1007/s00214-009-0667-0.

[30] Hohenberg, P. – Kohn, W. Inhomogeneous Electron Gas. *Physical Review*. November 1964, 136, s. 864–871. doi: 10.1103/PhysRev.136.B864.

[31] HOWARD, P. – MEYLAN, W. *Physical/chemical property database (Physprop)*. Syracuse Research Corporation, Environmental Science Cente, 1999.

[32] JELFS, S. – ERTL, P. – SELZER, P. Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *Journal of Chemical Information and Modeling*. 2007, 47, 2, s. 450–459. doi: 10.1021/ci600285n.

[33] JENSEN, F. *Introduction to computational chemistry*. New York, NY, USA : Wiley, 2007. ISBN 9780470011874.

[34] KAUFMAN, J. J. – SEMO, N. M. – KOSKI, W. S. Microelectrometric titration measurement of the pK_a's and partition and drug distribution coefficients of narcotics and narcotic antagonists and their pH and temperature dependence. *Journal of Medicinal Chemistry*. 1975, 18, 7, s. 647–655. doi: 10.1021/jm00241a001.

[35] KLOPMAN, G. – FERCU, D. Application of the multiple computer automated structure evaluation methodology to a quantitative structure-activity relationship study of acidity. *J. Comput. Chem.* September 1994, 15, s. 1041–1050. ISSN 0192-8651. doi: 10.1002/jcc.540150911.

[36] KUDERA, M. Softwarové nástroje pro výpočet disociačních konstant [online]. Bakalářská práce, Masarykova univerzita, Fakulta informatiky, Brno, 2010. Dostupné z: `http://is.muni.cz/th/207767/fi_b/`.

[37] KVASNIČKA, V. – KRATOCHVÍL, M. – KOČA, J. *Matematická chemie a počítačové řešení syntéz.* Pokroky chemie. Praha : Academia, 1987.

[38] LEACH, A. R. – PROUT, K. – DOLATA, D. P. The application of Artificial Intelligence to the conformational analysis of strained molecules. *Journal of Computational Chemistry.* 1990, 11, 6, s. 680–693. ISSN 1096-987X. doi: 10.1002/jcc.540110603. Dostupné z: `http://dx.doi.org/10.1002/jcc.540110603`.

[39] LEACH, A. – GILLET, V. *An Introduction to Chemoinformatics.* Dordrecht : Springer, 2007. ISBN 9781402062902.

[40] LEE, A. C. – CRIPPEN, G. M. Predicting pK_a. *Journal of Chemical Information and Modeling.* 2009, 49, 9, s. 2013–2033. doi: 10.1021/ci900209w.

[41] LIU, S. – PEDERSEN, L. G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J Phys Chem A.* Apr 2009, 113, s. 3648–3655. PMID: 19317439.

[42] MARZO, V. *Cannabinoids.* Neuroscience intelligence unit. Austin, Texas, USA : Landes Bioscience / Eurekah.com, 2004. ISBN 9780306482281.

[43] MCMURRY, J. – VUT. *Organická chemie.* Překlady vysokoškolských učebnic. Brno : VUTIUM, 2007. ISBN 9788070806371.

[44] MCNAUGHT, A. D. – WILKINSON, A. *IUPAC Compendium of Chemical Terminology.* Č. 2nd. New York, NY, USA : IUPAC, 1997. Dostupné z: `http://goldbook.iupac.org`.

[45] MORTIER, W. – GHOSH, S. – SHANKAR, S. Electronegativity-equalization method for the calculation of atomic charges in molecules. *Journal of the American Chemical Society.* 1986, 108, 15, s. 4315–4320.

[46] *NCI Open Database Compounds.* National Cancer Institute, Rockville, USA, September 2003.

[47] NOYES, R. et al. The analgesic properties of delta-9-tetrahydrocannabinol. *Clin Pharmacol Ther*. 1975, 15, s. 139–143.

[48] O'BOYLE, N. et al. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. 2011, 3, s. 33.

[49] ONDRÁK, M. Vyhledávání strukturních motivů významných pro studium disociačních konstant molekul [online]. Bakalářská práce, Masarykova univerzita, Fakulta informatiky, Brno, 2011. Dostupné z: `http://is.muni.cz/th/255906/fi_b/`.

[50] PATRICK, G. – SPENCER, J. *An introduction to medicinal chemistry*. Oxford : Oxford University Press, 2009. ISBN 9780199234479.

[51] PERRIN, D. – DEMPSEY, B. – SERJEANT, E. *pK_a prediction for organic acids and bases*. London : Chapman and Hall, 1981. ISBN 9780412221903.

[52] SALLAN, S. E. – ZINBERG, N. E. – FREI, E. Antiemetic Effect of Delta-9-Tetrahydrocannabinol in Patients Receiving Cancer Chemotherapy. *New England Journal of Medicine*. 1975, 293, 16, s. 795–797. doi: 10.1056/NEJM197510162931603.

[53] *Jaguar*. Schrödinger, Inc., New York, NY, USA, 2010.

[54] SHELLEY, J. C. et al. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* Dec 2007, 21, s. 681–691. PMID: 17899391.

[55] SKŘEHOTA, O. *Quantitative structure-property relationship modeling algorithms, challenges and IT solutions [online]*. Rigorózní práce, Masarykova univerzita, Fakulta informatiky, Brno, 2011. Dostupné z: `http://is.muni.cz/th/60606/fi_r/`.

[56] SVOBODOVÁ VAŘEKOVÁ, R. et al. Predicting pK_a Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes. *Journal of Chemical Information and Modeling*. 2011, 51, 8, s. 1795–1806. doi: 10.1021/ci200133w.

[57] TODESCHINI, R. – CONSONNI, V. – MANNHOLD, R. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*. Methods and Principles in Medicinal Chemistry. New York, NY, USA : Wiley-VCH, 2009. ISBN 9783527318520.

[58] TORRENS, F. – CASTELLANO, G. Topological charge-transfer indices: from small molecules to proteins. *Current Proteomics*. 2009, 6, 4, s. 204–213.

[59] *Concord*. Tripos, Inc., St Louis, MO, USA. Dostupné z: `http://www.tripos.com`.

[60] ULISS, D. B. et al. Hashish. Importance of the phenolic hydroxyl group in tetrahydrocannabinols. *Journal of Medicinal Chemistry*. 1975, 18, 2, s. 213–215. doi: 10.1021/jm00236a025.

[61] VAINIO, M. – JOHNSON, M. Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of chemical information and modeling*. 2007, 47, 6, s. 2462–2474.

[62] VOIGT, J. et al. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci*. 2001, 41, 3, s. 702–712.

[63] WAN, H. – ULANDER, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin Drug Metab Toxicol*. Feb 2006, 2, s. 139–155. PMID: 16863474.

[64] XING, L. – GLEN, R. Method for accurately estimating pK_a of molecules using atom type definitions and partial least squares, February 2006.

[65] XING, L. – GLEN, R. C. – CLARK, R. D. Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. *Journal of Chemical Information and Computer Sciences*. 2003, 43, 3, s. 870–879. doi: 10.1021/ci020386s.

[66] YOUNG, D. *Computational chemistry: a practical guide for applying techniques to real world problems*. New York, NY, USA : Wiley, 2001. ISBN 9780471333685.

[67] ZHANG, J. – KLEINÖDER, T. – GASTEIGER, J. Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *Journal of chemical information and modeling*. 2006, 46, 6, s. 2256–2266.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

BOA Born-Oppenheimerova aproximace

DFT Density Functional Theory

ESP Electrostatic potential analyses

GTO Gaussian Type Orbital

HF Hartree-Fock

HPE hyperplocha potenciální enrgie

LCAO Linear Combination of Atomic Orbitals

LFER Linear Free Energy Relationship

MMFF94 Merck Molecular Force Field 94

MPA Mullikenova populační analýza

NCI National Cancer Institute

NPA Natural population analysis

NSC National Sort Code

PA populační analýza

pK_a záporně vzatý dekadický logaritmus disociační konstanty kyseliny

pK_b záporně vzatý dekadický logaritmus disociační konstanty zásady

QSPR Quantitative Structure-Property Relationship

RMSE root mean square error

SCF self-consistent field

STO Slater Type Orbital

SMILES Simplified Molecular-Input Line-Entry Specification

A OBSAH PŘILOŽENÉHO CD

Součástí této práce je přiložené CD, které obsahuje:

- tréninkové sady molekul ve složce `molekuly/trenink`,
- testovací sadu molekul ve složce `molekuly/test`,
- grafy s výsledky ve složce `vysledky/grafy`,
- tabulku s výsledky ve složce `vysledky` (soubor `tabulka.xlsx`),
- text této práce včetně zdrojového kódu a obrázků ve složce `prace` a
- skripty použité během získávání a zpracování výsledků ve složce `skripty`.

Fenoly		B3LYP/6–31G*			B3LYP/STO–3G			HF/6–31G*			HF/STO–3G		
		MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP
Balloon	bez opt	0,5717	0,4177	0,5534	0,2817	0,1556	0,4332	0,6525	0,5464	0,5889	0,3920	0,2306	0,6751
	MM opt	0,9112	0,8412	0,7383	0,8167	0,8347	0,6991	0,9095	0,8880	0,7383	0,8695	0,8661	0,7536
	QM opt	0,9022	0,8867	0,8157	0,8443	0,8544	0,7787	0,8926	0,8918	0,8084	0,8739	0,8786	0,7834
Corina	bez opt	0,9617	0,9582	0,8061	0,9066	0,8959	0,8174	0,9656	0,9622	0,7828	0,9217	0,9206	0,8671
	MM opt	0,9354	0,7894	0,7668	0,8070	0,8486	0,6657	0,9358	0,8530	0,7688	0,8727	0,8624	0,7941
	QM opt	0,9438	0,9349	0,8464	0,8685	0,8806	0,8202	0,9439	0,9387	0,8282	0,9093	0,9140	0,8434
Openbabel	bez opt	0,9019	0,9072	0,7008	0,8659	0,8601	0,7749	0,8611	0,9054	0,7124	0,8987	0,9040	0,8170
	MM opt	0,9023	0,9061	0,6960	0,8635	0,8584	0,7688	0,8598	0,9043	0,7076	0,8974	0,9030	0,8145
	QM opt	0,9055	0,8943	0,7923	0,8731	0,8809	0,8337	0,8780	0,8922	0,7849	0,8965	0,9032	0,8450

Legenda	R^2
excelentní	0,950–0,990
velice dobré	0,920–0,950
dobré	0,900–0,920
akceptovatelné	0,850–0,900
slabé	0,800–0,850

Tab. B.1: Porovnání R^2 pro tréninkovou sadu molekul fenolů.

Aniliny m-, p-		B3LYP/6-31G*			B3LYP/STO-3G			HF/6-31G*			HF/STO-3G		
		MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP
Balloon	bez opt	0,7627	0,7987	0,4558	0,6900	0,7089	0,2918	0,7445	0,8062	0,4010	0,7186	0,8669	0,3332
	MM opt	0,8699	0,8698	0,7252	0,7971	0,7806	0,6121	0,8345	0,8632	0,7076	0,7384	0,7616	0,6060
	QM opt	0,9372	0,9452	0,8736	0,8963	0,9008	0,8079	0,9436	0,9427	0,7808	0,9141	0,9265	0,8308
Corina	bez opt	0,9698	0,9687	0,9024	0,8599	0,8600	0,8662	0,9782	0,9800	0,8394	0,8993	0,9057	0,8587
	MM opt	0,9663	0,9836	0,8831	0,8737	0,8730	0,7806	0,9486	0,9765	0,8266	0,8315	0,8440	0,7443
	QM opt	0,9429	0,9518	0,8778	0,9078	0,9086	0,8691	0,9558	0,9494	0,8044	0,9270	0,9370	0,8881
Openbabel	bez opt	0,9681	0,9775	0,8608	0,8877	0,8793	0,7965	0,9416	0,9571	0,7637	0,8287	0,8284	0,7422
	MM opt	0,9541	0,9653	0,8330	0,8735	0,8767	0,7657	0,9274	0,9415	0,7510	0,8376	0,8393	0,7507
	QM opt	0,9452	0,9497	0,8870	0,9017	0,9054	0,8385	0,9489	0,9466	0,8145	0,9227	0,9343	0,8524

Aniliny o-		B3LYP/6-31G*			B3LYP/STO-3G			HF/6-31G*			HF/STO-3G		
		MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP
Balloon	bez opt	0,9126	0,8589	0,7336	0,6465	0,7881	0,4724	0,9089	0,9054	0,8125	0,7854	0,8697	0,6678
	MM opt	0,9451	0,9240	0,8616	0,8153	0,8114	0,5854	0,9087	0,9073	0,8616	0,8557	0,8737	0,7438
	QM opt	0,9446	0,9431	0,8514	0,9163	0,9208	0,7695	0,9055	0,9208	0,8488	0,9044	0,9145	0,8539
Corina	bez opt	0,9176	0,9628	0,8609	0,8101	0,7800	0,7711	0,9381	0,9728	0,8644	0,8689	0,8731	0,8719
	MM opt	0,9289	0,9643	0,9183	0,8220	0,7978	0,7974	0,9223	0,9715	0,9149	0,8799	0,8855	0,8728
	QM opt	0,9265	0,9224	0,9156	0,9032	0,8984	0,9130	0,8888	0,9110	0,9085	0,9036	0,9072	0,9327
Openbabel	bez opt	0,9326	0,9580	0,9090	0,8306	0,8088	0,7763	0,9136	0,9690	0,9152	0,8866	0,8909	0,8576
	MM opt	0,9374	0,9651	0,9187	0,8319	0,8096	0,7930	0,9194	0,9748	0,9237	0,8900	0,8943	0,8762
	QM opt	0,9340	0,9295	0,9263	0,9126	0,9090	0,9105	0,8956	0,9166	0,9252	0,9142	0,9165	0,9306

Tab. B.2: Porovnání R^2 pro tréninkovou sadu molekul anilinů.

Ben. kyseliny m-, p-		B3LYP/6-31G*			B3LYP/STO-3G			HF/6-31G*			HF/STO-3G		
		MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP
Balloon	bez opt	0,8370	0,8627	0,5183	0,7170	0,8332	0,5664	0,8895	0,9000	0,5552	0,8862	0,9120	0,6974
	MM opt	0,8231	0,8740	0,7897	0,7443	0,8120	0,5631	0,9192	0,9019	0,8200	0,9043	0,9178	0,6722
	QM opt	0,9016	0,9493	0,8984	0,9422	0,8960	0,7973	0,9309	0,9443	0,9012	0,9463	0,9436	0,8509
Corina	bez opt	0,9450	0,9524	0,9337	0,8689	0,8618	0,7987	0,9493	0,9486	0,9269	0,9196	0,9187	0,8490
	MM opt	0,9356	0,9420	0,9284	0,8572	0,8457	0,8081	0,9494	0,9579	0,9370	0,9289	0,9338	0,8822
	QM opt	0,9571	0,9535	0,9366	0,9333	0,8840	0,7767	0,9665	0,9677	0,9496	0,9359	0,9320	0,8143
Openbabel	bez opt	0,9473	0,9517	0,9479	0,8839	0,8684	0,8534	0,9506	0,9619	0,9476	0,9318	0,9334	0,8820
	MM opt	0,9478	0,9526	0,9465	0,9043	0,8679	0,8561	0,9485	0,9614	0,9458	0,9306	0,9309	0,8861
	QM opt	0,9598	0,9602	0,9524	0,9386	0,8954	0,8751	0,9648	0,9696	0,9508	0,9371	0,9364	0,8886

Benz. kyseliny o-		B3LYP/6-31G*			B3LYP/STO-3G			HF/6-31G*			HF/STO-3G		
		MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP	MPA	NPA	ESP
Balloon	bez opt	0,6319	0,6000	0,4267	0,4578	0,4887	0,3609	0,6976	0,6735	0,4542	0,6559	0,6795	0,5238
	MM opt	0,6922	0,6574	0,7369	0,5662	0,6307	0,4834	0,7959	0,7075	0,7347	0,7990	0,8055	0,6698
	QM opt	0,7644	0,8365	0,8319	0,7730	0,7959	0,6988	0,8107	0,8520	0,8341	0,8547	0,8740	0,7485
Corina	bez opt	0,7547	0,8539	0,5654	0,6255	0,6445	0,6268	0,7747	0,8484	0,6161	0,7490	0,7496	0,7601
	MM opt	0,8550	0,8685	0,8376	0,8178	0,8419	0,7866	0,8798	0,8855	0,8414	0,8940	0,9060	0,8667
	QM opt	0,9103	0,9057	0,8702	0,9004	0,9041	0,9172	0,9190	0,9157	0,8629	0,9173	0,9124	0,9067
Openbabel	bez opt	0,7814	0,8001	0,7191	0,7404	0,7796	0,5930	0,7673	0,8503	0,7147	0,8406	0,8681	0,7150
	MM opt	0,7811	0,8043	0,7210	0,7495	0,7870	0,5863	0,7665	0,8556	0,7204	0,8456	0,8719	0,7139
	QM opt	0,7587	0,8022	0,6982	0,6250	0,6437	0,7357	0,7748	0,7558	0,7271	0,6891	0,7121	0,7879

Tab. B.3: Porovnání R^2 pro tréninkovou sadu molekul benzoových kyselin.

C DOPLŇUJÍCÍ INFORMACE K TEORII

C.1 pK_b

Disociační konstanta K_b vyjadřuje rovnováhu disociace zásady podle rovnice (C.1).

$$\mathrm{B}(aq) + \mathrm{H_2O}(l) \rightleftharpoons \mathrm{BH^+}(aq) + \mathrm{OH^-}(aq) \tag{C.1}$$

$$K_b = \frac{a_{\mathrm{OH^-}} a_{\mathrm{BH^+}}}{a_{\mathrm{B}}} \approx \frac{[\mathrm{OH^-}][\mathrm{BH^+}]}{[\mathrm{B}]} \tag{C.2}$$

Mezi pK_a konjugované kyseliny a pK_b zásady lze snadno převádět. Jejich součet je vždy roven 14, tedy záporně vzatému dekadickému logaritmu iontového součinu vody.

$$\begin{aligned} pK_a + pK_b &= -\log \frac{[\mathrm{H_3O^+}][\mathrm{B}]}{[\mathrm{BH^+}]} - \log \frac{[\mathrm{OH^-}][\mathrm{BH^+}]}{[\mathrm{B}]} \\ &= -\log \frac{[\mathrm{H_3O^+}][\mathrm{B}][\mathrm{OH^-}][\mathrm{BH^+}]}{[\mathrm{BH^+}][\mathrm{B}]} \\ &= -\log [\mathrm{H_3O^+}][\mathrm{OH^-}] \\ &= -\log K_w = 14 \end{aligned} \tag{C.3}$$

C.2 Kvantová mechanika

C.2.1 Born-Oppenheimerova aproximace

Born-Oppenheimerova aproximace (BOA) umožňuje oddělit pohyb elektronů od pohybu jader [66]. Pro izolovanou molekulu je možno rozdělit její hamiltonián na následující složky:

$$\hat{H} = T_e + T_n + V_{ee} + V_{en} + V_{nn} \tag{C.4}$$

kde T_e a T_n jsou operátory kinetické energie elektronů a jader a V_{ee}, V_{en} a V_{nn} operátory potenciální energie zahrnující elektrostratické interakce elektron-elektron, elektron-jádro a jádro-jádro.

Elektrony, jejichž hmotnost se pohybuje o tři a více řádů níž než hmotnost jádra, se zároveň i pohybují řádově rychleji. Lze tedy očekávat, že elektrony reagují na změnu polohy jader téměř okamžitě, zatímco změna polohy elektronu se chování jader nijak nedotkne. Operátor potenciální energie pro interakce jádro-jádro tedy můžeme zanedbat:

$$\hat{H} = T_e + T_n + V_{ee} + V_{en} \tag{C.5}$$

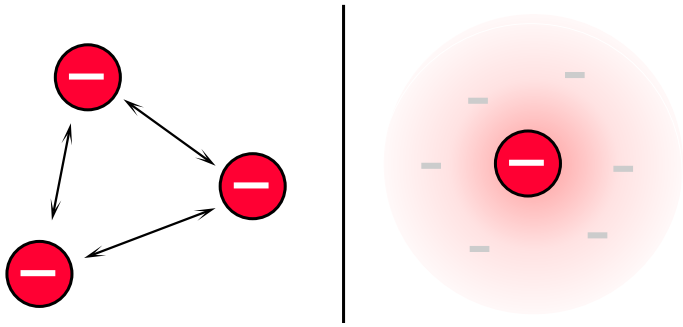
To znamená, že v rámci BOA řešíme pohyb elektronů pro statická jádra a to opakujeme pro různé polohy jader [66]. Výsledkem je pak hyperplocha potenciální enrgie (HPE) – tedy závislost energie elektronů na poloze jader. Následně můžeme studovat pohyb jader po HPE. Tato aproximace se díky své přesnosti používá v drtivé většině metod kvantové chemie [33].

C.2.2 Model nezávislých částic

Model nezávislých částic neboli self-consistent field (SCF) [33, 66] se nezabývá časově závislými jevy a pracuje pouze se spin-orbitálními interakcemi (interakce elektronů) a uspořádání jader považuje za konstantní. Hamiltonián tak redukuje na tuto formu:

$$\hat{H} = T_e + V_{ee} + V_{en} \tag{C.6}$$

Důvodem, proč nemůžeme najít analytické řešení Schrödingerovy rovnice je přítomnost odpudivých sil mezi elektrony. Interakce mezi elektrony je ze své podstaty párová k energii tak odpovídá součtu příspěvků všech možných dvojich elektronů v systému. Nelze ji tedy rozdělit na příspěvky, z nichž by každý odpovídal jednomu elektronu. Model nezávislých částic umožňuje díky zanedbání párové interakce rozdělit problém pohybu N elektronů rozdělit N nezávislých jednoelektronových problémů, které můžeme řešit.

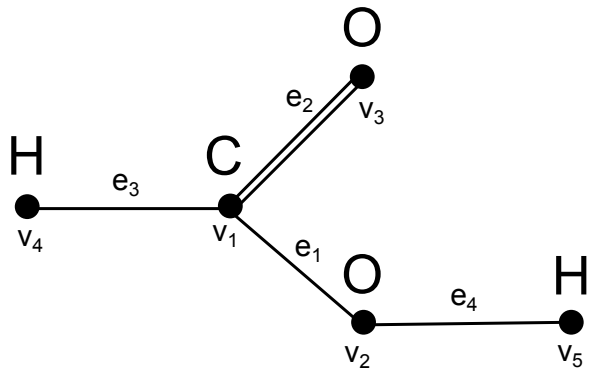


Obr. C.1: Zanedbání interakcí mezi elektrony umístěním elektronu do průměrného (středního) elektrického pole.

Elektron-elektronová repulze je příliš významná na to, aby se dala zanedbat. Vzniklý model by byl od skutečnosti příliš vzdálený. Proto se interakce mezi elektrony nahradí interakcí jednoho elektronu se středním polem ostatních elektronů.

C.3 Molekulové grafy

Topologie jsou v počítači reprezentovány molekulovými grafy. Uzly molekulového grafu reprezentují atomy a hrany reprezentují vazby. Uzly obsahují informaci o typu atomu, hrany obsahují řád vazby. Molekulové grafy mají různé typy, které se liší například v různém pojmenování množit vrcholů a hran.



Obr. C.2: Molekulový graf kyseliny mravenčí.

Jako příklad uvedeme definici molekulového grafu podle Kvasničky et al.[37], která obsahuje atomové souřadnice ale nikoliv volné elektrony (Ty by se daly implementovat jako smyčky na vrcholech.). Tento molekulový graf je pětinásobný (V, E, ϕ, β, γ), kde:

- V je množina vrcholů (atomů).
- E je multimnožina[1] hran (vazeb).
- φ je zobrazení $V \to \beta$, které přiřazuje atomům chemické značky.
- β je množina chemický značek.
- γ je funkce $V \to \mathbb{R}^3$, která přiřazuje atomům souřadnice x, y a z

Formálně může být molekula mravenčí kyseliny (HCOOH) reprezentována molekulovým grafem $G_{\mathrm{HCOOH}} = (V, E, \varphi, \beta, \gamma)$, kde:

- $V = \{v_1, v_2, v_3, v_4, v_5\}$
- $E = \{\{v_1, v_4\}, \{v_1, v_3\}, \{v_1, v_3\}, \{v_1, v_2\}, \{v_2, v_5\}\}$
- $\varphi(v_1)$ =C, $\varphi(v_2)$ =O, $\varphi(v_3)$ =O, $\varphi(v_4)$ =H, $\varphi(v_5)$ =H
- β = {C, O, H}
- $\gamma(v_1) = \{1,8733; -0,6193; 4,5914\}, \gamma(v_2) = \{4,0306; -1,9142; 2,7446\}$, $\gamma(v_3) = \{2,9708; -1,4085; 2,0522\}, \gamma(v_4) = \{1,9042; -0,7896; 2,7040\}$, $\gamma(v_5) = \{2,9683; -1,5204; 0,6618\}$

Molekulový graf G_{HCOOH} je vyobrazen na obrázku C.2.

[1] Množina, která může obsahovat více stejných prvků.