Diplomová práce

České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra počítačů

Vývoj molekulárních reprezentací pro aplikaci strojového učení v kvantové chemii a spektroskopii

Ing. Štěpán Sršeň

Vedoucí: prof. Ing. Filip Železný, Ph.D.
Květen 2019

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Sršeň** Jméno: **Štěpán** Osobní číslo: **420839**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Otevřená informatika**

Studijní obor: **Umělá inteligence**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Vývoj molekulárních reprezentací pro aplikaci strojového učení v kvantové chemii a spektroskopii

Název diplomové práce anglicky:

Development of molecular representations suitable for machine learning application in quantum chemistry and spectroscopy

Pokyny pro vypracování:

1. Rešerše metod strojového učení používaných v oblasti teoretické výpočetní chemie.
2. Rešerše molekulárních reprezentací doposud používaných v kvantové chemii a spektroskopii.
3. Testování vybraných reprezentací v rámci strojového učení základních molekulárních vlastností.
4. Testování vybraných reprezentací v rámci strojového učení spektroskopicky významných vlastností.
5. Analýza konstrukce molekulárních deskriptorů a její vliv na popis extenzivních a intenzivních vlastností molekul.
6. Vývoj nových reprezentací vhodných pro strojové učení základních i spektroskopicky významných molekulárních vlastností.
7. Testování nově vyvinutých molekulárních reprezentací.
8. Aplikace strojového učení na modelování elektronových spekter s použitím nově vyvinutých molekulárních reprezentací a porovnání s ab initio výpočty a experimenty.

Seznam doporučené literatury:

(1) Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning; Springer Series in Statistics; Springer New York: New York, NY, 2009.
(2) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. J. Chem. Phys. 2018, 148 (24), 241401.
(3) Huang, B.; von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. J. Chem. Phys. 2016, 145 (16), 161102.
(4) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. J. Chem. Phys. 2018, 148 (24), 241717.
(5) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Von Lilienfeld, O. A. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. J. Chem. Phys. 2015, 143 (8), 084111.

Jméno a pracoviště vedoucí(ho) diplomové práce:

prof. Ing. Filip Železný, Ph.D., Intelligent Data Analysis FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **28.01.2019** Termín odevzdání diplomové práce: **24.05.2019**

Platnost zadání diplomové práce: **20.09.2020**

prof. Ing. Filip Železný, Ph.D. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Ing. Pavel Ripka, CSc. podpis děkana(ky)

 © ČVUT v Praze, Design: ČVUT v Praze, VIC

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací.
Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

.	
Datum převzetí zadání	Podpis studenta

© ČVUT v Praze, Design: ČVUT v Praze, VIC

Poděkování

Tímto bych chtěl poděkovat vedoucímu této práce prof. Filipovi Železnému. Dále bych chtěl poděkovat babičce Marušce za neutuchající entusiasmus ohledně mých studií.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 20. května 2019 .....

Ing. Štěpán Sršeň

Abstrakt

Tato práce se zabývá aplikací strojového učení do oblasti kvantové chemie a spektroskopie, kde je hlavním problémem kvantitativních výpočtů jejich časová náročnost. Hlavním cílem je vývoj a testování vhodných molekulárních reprezentací neboli forem vstupního vektoru učení, které představují v této oblasti nejdůležitější součást regrese. Navrhuji zde různé přístupy založené mimo jiné na maticích atomových interakcí, na distribucích interakcí či na vypočítaných molekulových vlastnostech.

Navrhované reprezentace poskytují na testovaných datových sadách aktuálně nejlepší publikované výsledky. Použití těchto reprezentací pro modelování spekter umožňuje snížení výpočetní náročnosti od jeden až dva řády.

Klíčová slova: strojové učení, hřebenová regrese, jádrová transformace, molekulární reprezentace, spektroskopie, absorpční spektrum

Vedoucí: prof. Ing. Filip Železný, Ph.D.

Abstract

This work is focused on the application of machine learning into quantum chemistry and spectroscopy. The main problem of quantitative quantum chemical calculations are their high computational demands which we want to overcome by using machine learning techniques. The main goal of this work is the development of new molecular representations as it is the most important part of regression in this field. I propose various molecular representations based, for example, on matrices of atomic interactions, distributions of atomic interactions or precomputed molecular properties.

Proposed representations provide better results in comparison with available literature. By using these representations, it possible to accelerate spectra modelling by up to two orders of magnitude.

Keywords: machine learning, kernel ridge regression, molecular representation, spectroscopy, absorption spectrum

Title translation: Development of molecular representations suitable for machine learning application in quantum chemistry and spectroscopy

Obsah

1 Úvod **1**

2 Teoretický základ **3**

2.1 Regresní metody 3

2.1.1 Lineární regrese 3

2.1.2 Nelineární regrese 4

2.2 Úvod do kvantové mechaniky molekul 6

2.3 Absorpční spektra 8

2.3.1 Veličiny popisující absorpci . . . 9

2.3.2 Modelování empirickým rozšířením 10

2.3.3 Reflexní princip 11

2.4 Statistické zpracování spektrálních dat 12

2.4.1 Odhad hustoty pravděpodobnosti 12

2.4.2 Statistická chyba modelování 13

2.4.3 Porovnání distribucí 14

3 Analýza současného stavu **17**

3.1 Molekulární reprezentace 17

3.2 Datové sady 19

3.3 Strojové učení v kvantové chemii a spektroskopii 21

4 Návrh a implementace **23**

4.1 Molekulární deskriptory 24

4.1.1 Matice interakcí 24

4.1.2 Množiny interakcí 25

4.1.3 Distribuce interakcí 26

4.1.4 Výčet orbitálních energií 26

4.2 Strojové učení rozdílu 28

4.3 Vstupní formát 28

5 Výsledky a testování **31**

5.1 Datová sada QM7b 31

5.2 Datová sada QM8 36

5.3 Molekula CH_2OO 38

5.4 Molekula azobenzenu 41

6 Závěr **47**

Literatura **49**

Seznam použitých zkratek **55**

Obrázky

2.1 Schematické znázornění reflexního principu. Jaderná hustota základního stavu je přes hyperplochu excitovaného stavu reflektována až na energetickou osu spektra. 12

4.1 Znázornění distribuce založené na energiích molekulových orbitalů. . . 28

5.1 Porovnání střední absolutní chyby Laplaceova a Gaussova jádra pro různé reprezentace na datové sadě QM7b. 32

5.2 Porovnání učících časů pro různé reprezentace na datové sadě QM7b. 32

5.3 Porovnání střední absolutní chyby pro různé verze CM reprezentace na datové sadě QM7b. 33

5.4 Porovnání střední absolutní chyby pro různé reprezentace na datové sadě QM7b. 35

5.5 Porovnání střední absolutní chyby 1. excitační energie (vlevo) a 2. excitační energie (vpravo) pro datovou sadu QM8. 36

5.6 Porovnání střední absolutní chyby oscilátorové síly 1. (vlevo) a 2. (vpravo) excitovaného stavu pro datovou sadu QM8. 37

5.7 Porovnání střední absolutní chyby Laplaceova a Gaussova jádra pro excitační energie (vlevo) a tranzitní dipólové momenty (vpravo) pro molekulu CH_2OO. 38

5.8 Porovnání průměrné střední absolutní chyby nejlepších reprezentací s a bez použití ΔML pro molekulu CH_2OO. 40

5.9 Srovnání spektra s a bez použití ML pro molekulu CH_2OO. 41

5.10 Porovnání průměrné střední absolutní chyby excitačních energií (vlevo) a střední absolutní chyby energií 2. excitovaného stavu (vpravo) pro molekulu azobenzenu. 43

5.11 Porovnání průměrné střední absolutní chyby oscilátorových sil pro molekulu azobenzenu. 44

5.12 Porovnání průměrné střední absolutní chyby nejlepších reprezentací s a bez použití ΔML pro molekulu azobenzenu. 45

5.13 Srovnání spektra s a bez použití ML pro molekulu (E)-azobenzenu. 45

Tabulky

5.1 Porovnání středních absolutních chyb pro různé mocniny vzdáleností v BoB reprezentaci na datové sadě QM7b. 33

5.2 Porovnání středních absolutních chyb pro vyvíjené reprezentace založené na distribucích na datové sadě QM7b. 34

5.3 Porovnání středních absolutních chyb excitačních energií s a bez zahrnutí tříatomových interakcí na datové sadě QM7b. 36

5.4 Porovnání středních absolutních chyb pro distribuci orbitálních energií s a bez použití ΔML na datové sadě QM8. 37

5.5 Porovnání středních absolutních chyb excitačních energií pro různá nastavení reprezentací pro molekulu CH_2OO. 39

5.6 Porovnání středních absolutních chyb tranzitních dipólových momentů pro různá nastavení reprezentací pro molekulu CH_2OO. 39

5.7 Porovnání středních absolutních chyb excitačních energií pro různá nastavení reprezentací pro molekulu azobenzenu. 42

5.8 Porovnání středních absolutních chyb průměru tranzitních dipólových momentů pro různá nastavení reprezentací pro molekulu azobenzenu. 42

Kapitola 1

Úvod

Cílem této práce je prostudování možností nasazení strojového učení (ML, z anglického *Machine Learning*) v kvantové chemii a hlavně spektroskopii. Důvodem k aplikaci přístupů ML je časová náročnost exaktních kvantově-chemických výpočtů, kdy při požadované přesnosti nejsme často schopni k výsledkům v rozumném časovém měřítku vůbec dojít. Hlavním předmětem této práce je tedy vývoj a testování tzv. molekulárních deskriptorů, jelikož právě ty hrají při použití ML v kvantové chemii největší roli. Jedná se o způsob zakódování molekuly do vstupního vektoru učení. Speciální pozornost je pak věnována aplikaci v podoblasti výpočetní elektronové spektroskopie, kde je v tuto chvíli použití ML velmi omezené.

Absorpční spektra jsou grafickým znázorněním funkční závislosti intenzity absorpce světla na energii resp. vlnové délce záření. Spektra obecně jsou jednou ze základních charakteristik molekul. Jejich znalost nám podává informace o vlastnostech těchto molekul a umožňuje jejich spektroskopickou identifikaci. Pokud disponujeme absolutními intenzitami, jsme schopni látku ve vzorku nejen rozpoznat, ale i kvantitativně určit její koncentraci. Chemické přeměny jsou velmi často podmíněny dodáním energie pro překonání aktivační bariéry, ta může být dodána v podobě absorpce záření. Absorpční spektra tak představují i zásadní zdroj informací pro výpočet rychlostních konstant fotochemických reakcí. Například v kontextu chemie atmosféry je rychlostní konstanta fotochemické reakce dána jako:

$$j = \int_{\lambda_i}^{\lambda_j} \sigma(\lambda)\Phi(\lambda)I(\lambda)\mathrm{d}\lambda,$$

kde λ je vlnová délka, $\sigma(\lambda)$ absorpční účinný průřez vyjadřující intenzitu absorpce, $\Phi(\lambda)$ kvantový výtěžek reakce a $I(\lambda)$ je aktinický tok (neboli intenzita záření) v daném prostředí. I zde přitom musíme disponovat absolutními hodnotami absorpčních účinných průřezů.[20]

Některé významné atmosférické molekuly mají však tak krátkou dobu života, že je experimentálně nelze zachytit nebo je takové měření velmi nákladné. K takovým příkladům patří například tzv. Criegeeho meziprodukty, důležité v chemii troposféry.[59] V takovýchto situacích je užitečné modelovat absorpční

spektra na základě teoretických výpočtů. I při současné úrovni poznání však stále nejsme schopni kvantitativně a efektivně modelovat absorpční spektra větších molekul či molekul s komplikovanou elektronovou strukturou. Jsme tak často stále odkázáni na nedostatečně přesné kvantově-chemické metody. Základním problémem je časová náročnost výpočtů, které by byly potřeba pro kvantitativní modelování, a omezená přesnost technik používaných pro modelování spekter. Řešení tohoto problému za pomocí přístupů strojového učení je předmětem této práce. Další aplikací zde vyvíjených přístupů je rychlé skenování chemického prostoru za účelem nalezení molekul s požadovanými fotochemickými vlastnostmi.

Kapitola 2

Teoretický základ

2.1 Regresní metody

Za účelem regrese byla vyvinuta celá řada ML algoritmů. Velmi jednoduchým příkladem regresní úlohy může být lineární regrese, kdy konečnou sadu bodů x, y prokládáme lineární funkcí, která nám umožní předpovědět hodnotu y pro nové vybrané hodnoty x. V našem případě však hledáme komplexní funkci a musíme tak použít metody nelineární regrese, nicméně princip zůstává stejný, akorát nastavujeme parametry složitějšího modelu.

2.1.1 Lineární regrese

Vraťme se pro názornost k lineární regresi. Při lineární regresi v d dimenzích hledáme predikovanou veličinu jako lineární kombinaci příslušných d deskriptorů:

$$f(\mathbf{x}) = \sum_{i=1}^{d} \alpha_i x_i.$$

Regresní koeficienty α_i zde představují váhy jednotlivých dimenzí/deskriptorů x_i. Z předpisu je zřejmé, že taková funkce musí procházet počátkem. Předpokládáme zde totiž, že data jsou vystředěná, jinak bychom museli přidat ještě posun. Otázkou zůstává, jakým způsobem získat regresní koeficienty. Chceme nějakým způsobem minimalizovat odchylku predikovaných hodnot od opravdových hodnot pro trénovací data. Nejčastěji používaná je metoda nejmenších čtverců neboli minimalizace sumy čtverců odchylek:

$$\min_{\alpha} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2,$$

kde i tentokrát indexuje trénovací data. Výhodou této funkce je snadná optimalizace, ale například oproti sumě absolutních odchylek je více citlivá na odlehlé hodnoty v trénovacích datech.[50]

Další problém může představovat přítomnost tzv. multikolinearity neboli stav, kdy jsou jednotlivé deskriptory korelované. Metoda se pak stává velmi

citlivou, kvůli čemuž máme poté velkou varianci predikovaných koeficientů. Reálně se jen málokdy setkáme s problémy, kdy by dimenze byly úplně ortogonální a multikolinearita tedy představuje běžný problém.[9] Propojeným problémem je tzv. přeučení, což je efekt, kdy náš model naučíme příliš dobře predikovat data z trénovací množiny, avšak na úkor predikce dat nových, která je naším cílem. K přeučení dochází, pokud dáváme modelu k dispozici příliš mnoho stupňů volnosti (v tomto případě deskriptorů), což vede k natrénování trénovacích dat včetně statistického šumu. Takovým velmi jednoduchým příkladem přeučení by bylo, pokud bychom zašuměná data pocházející z přímky prokládali nějakým polynomem vysokého stupně. Při dosti vysokém stupni polynomu bychom data sice proložili dokonale, avšak predikce nových hodnot by byla značně nepřesná. Řešením výše uvedených problémů může být přidání regularizačního členu, který bude zároveň se sumou odchylek minimalizovat sumu regresních koeficientů. Jedná se o uplatnění Occamovy břitvy v praxi, kdy preferujeme jednodušší řešení. Opět je možné volit, jakým způsobem budeme koeficienty sčítat. V případě sumy absolutních hodnot se jedná o tzv. LASSO regresi, v případě sumy čtverců koeficientů jde o hřebenovou regresi:

$$\min_{\alpha} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda \sum_{i=1}^{n} \alpha_i^2,$$

kde λ představuje hyperparametr určující zapojení regularizačního členu. Tento parametr získáme další minimalizací. Ta probíhá buďto přes validační data, což je množina trénovacích dat předem oddělená pro tento účel, nebo pomocí křížové validace. Při křížové validaci rozdělíme trénovací data do sady n podmnožin a na každé z nich vyhodnotíme parametry nastavené na zbývajících $n-1$ podmnožinách, přičemž minimalizujeme průměrnou chybu přes všechna tato vyhodnocení.[50]

2.1.2 Nelineární regrese

V případě nelineární regrese lze rozlišit dvě základní kategorie algoritmů. První z nich jsou metody založené na tzv. jádrové transformaci. Jádrová transformace umožňuje aplikovat metody lineární regrese na nelineární problémy. V principu se jedná o to, že v těchto metodách transformujeme vstupní data $\mathbf{x}$ do prostoru s vyšší dimenzionalitou, kde je problém již řešitelný pomocí lineárního algoritmu. Jakým způsobem však najít správnou transformaci $\phi(\mathbf{x})$? Ukazuje se, že často nejvhodnější mapovací funkce jsou velmi složité a vedou na prostor s nekonečnou dimenzionalitou. Řešení takových problémů je zjevně výpočetně nedostupné. Metody založené na jádrové transformaci obchází tento problém využitím dvou pozorování. Za prvé většina lineárních metod lze přepsat tak, že používají vstupní data pouze ve formě skalárního součinu mezi dvěma vstupy $\mathbf{x}_i\mathbf{x}_j$ neboli v případě použití transformace $\phi(\mathbf{x}_i)\phi(\mathbf{x}_j)$. Za druhé můžeme místo mapování dat a následného výpočtu skalárních součinů v transformovaném prostoru spojit tyto operace do jedné tzv. jádrové funkce $k(\mathbf{x}_i, \mathbf{x}_j)$, která operuje v původním prostoru vstupních dat, ale poskytne

nám stejný výsledek jako skalární součin v transformovaném prostoru. Jinak řečeno, vhodnou definicí jádrové funkce můžeme data transformovat do prostoru o vyšší (a klidně i nekonečné) dimenzionalitě zcela implicitně bez použití transformace samotné.[55]

Pokud uplatníme jádrovou transformaci na hřebenovou regresi, predikce veličiny pro nové vstupy v rovnici 2.1.1 přejde na lineární kombinaci přes jádrové transformace nového vstupu s trénovacími daty:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}).$$

Regresní koeficienty získáme modifikací rovnice 2.1.1 opět jako minimalizaci přes trénovací data:[26]

$$\min_{\alpha} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda \sum_{i=1,j=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) \alpha_j.$$

Otázkou zůstává, jakým způsobem zvolit jádrovou funkci. Dle Mercerova teorému může jádro $\mathbf{K}$, neboli matici s prvky $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, tvořit jakákoliv symetrická pozitivně semidefinitní matice.[37] V praxi se ukazuje, že v chemii (a v mnoha dalších aplikacích) bývá nejlepší použít jako jádrovou funkci Gaussovu funkci:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{x}_i - \mathbf{x}_j||_2^2\right)$$

či Laplaceovu funkci:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{\sigma} ||\mathbf{x}_i - \mathbf{x}_j||_1\right).$$

Obě transformují implicitně vstupní data do nekonečně dimenzionálního prostoru. Rozdíl mezi nimi tkví v použité normě v exponentu. Zatímco Gaussovo jádro používá eukleidovskou normu, Laplaceovo jádro používá manhattanskou/oktaedrickou normu. Parametr σ představuje rozšíření Gaussovy, resp. Laplaceovy funkce a je vhodné ho optimalizovat jako hyperparametr. Jádrovou funkci si lze představit jako metriku, která udává, jako moc odlišné dva vstupní vektory jsou.[26]

Výše odvozená metoda se nazývá hřebenová regrese s jádrovou transformací (KRR, z angl. *Kernel Ridge Regression*). Do kategorie metod založených na jádrové transformaci dále spadá například metoda podpůrných vektorů (SVM, z angl. *Support Vector Machines*) či regrese gaussovskými procesy (GPR, z angl. *Gaussian Process Regression*).[55] Druhou hlavní kategorií regresních metod jsou neuronové sítě. Jedná se o kombinaci mnoha jednoduchých funkcí (neuronů), ze kterých můžeme složit libovolně složitý model. Jejich základní nasazení není již v dnešní době s pomocí knihoven složité, avšak jejich správné navržení a použití je stále otázkou empirických zkušeností a dlouhého vývoje. Při použití neuronových sítí nemáme žádnou záruku funkčnosti a je tak třeba

je vždy validovat oproti nějakému základnímu modelu, kterým může být právě některá z výše zmíněných metod používajících jádrovou transformaci. Při dlouhodobém vývoji většinou neuronové sítě překonávají ostatní metody, avšak potřebují velké množství učících dat, kterými v kvantové chemii většinou nedisponujeme. Ukazuje se tak, že v této oblasti zatím nepřináší oproti metodám s jádrovou transformací prokazatelně lepší výsledky.[19] Metody s jádrovou transformací a neuronové sítě jsou pouze dva způsoby, jak nafitovat komplexní flexibilní funkci. Další možností by mohly být přístupy založené na stromech jako třeba regresní stromy či náhodné lesy.[50]

Jako regresní metodu pro testování molekulových reprezentací v této práci jsem si vybral KRR. Jedná se totiž o relativně jednoduchý algoritmus oblíbený právě v kvantové chemii,[52] který mi umožní zaměřit se na práci se vstupními daty. Je totiž zřejmé, že při množství dat, které míváme k dispozici v kvantové chemii, je forma vstupního vektoru $\mathbf{x}$ tou nejdůležitější součástí učení. Typicky chceme do tohoto vektoru nějakým způsobem zakódovat strukturu molekuly, mluvíme potom o tzv. molekulární reprezentaci. Ne nutně však musíme pro učení používat pouze strukturu, někdy může být žádoucí použít například nějakou předpočítanou vlastnost molekuly, potom hovoříme obecněji o molekulárních deskriptorech.[19]

2.2 Úvod do kvantové mechaniky molekul

Pro pochopení konstrukce molekulárních reprezentací a opodstatnění nasazení ML v kvantové chemii je nutné nejprve objasnit některé koncepty z této oblasti. Kvantová mechanika je založena na vlnově-částicovém dualismu, kdy částice se chovají částečně jako vlnění a elektromagnetické vlnění se naopak někdy chová jako částice. Stav kvantově-mechanické soustavy definujeme pomocí vlnové funkce $\Psi(t)$, a nikoliv pomocí trajektorie, jak jsme zvyklí z klasické mechaniky. Stav systému, a tedy jeho vlastnosti, jsou plně určeny touto vlnovou funkcí. Její druhá mocnina nám pak udává pravděpodobnost výskytu částice v daném intervalu vybrané souřadnicové reprezentace. Vlnovou funkci libovolně složitého systému lze získat vyřešením časově závislé Schrödingerovy rovnice:

$$i\hbar\frac{\partial}{\partial t}\Psi(t) = \hat{H}\Psi(t), \tag{2.1}$$

kde i je imaginární jednotka, $\hbar$ je redukovaná Planckova konstanta, t je čas a $\hat{H}$ představuje Hamiltonův operátor celkové energie. Pro řešení stacionárních problémů lze předchozí rovnici zjednodušit do časově nezávislé formy:

$$\hat{H}\Psi = E\Psi, \tag{2.2}$$

kde E představuje vlastní číslo Hamiltonova operátoru a udává možné hodnoty celkové energie systému. Zde narážíme na další charakteristiku kvantové mechaniky, a to je kvantování fyzikálních veličin, které mohou často nabývat pouze diskrétních hodnot.[2]

Hamiltonův operátor pro částici o hmotnosti m má následující tvar:

$$\hat{H} = -\frac{\hbar^2}{2m}\Delta + \hat{V}, \tag{2.3}$$

kde první člen představuje kinetickou energii systému a druhý člen $\hat{V}$ energii potenciální. Δ je Laplaceův operátor, který zastupuje sumu druhých derivací dle polohových souřadnic:

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.4}$$

Pro molekulu lze za určitých předpokladů (absence vnějšího elektromagnetického pole, zanedbání tzv. spin-orbitální interakce a relativistických efektů) vyjádřit Hamiltonův operátor jako sumu kinetických příspěvků jednotlivých částic a jejich vzájemného elektrostatického působení:

$$\hat{H} = -\sum_i \frac{\hbar^2\Delta}{2m_e} - \sum_k \frac{\hbar^2\Delta}{2m_k} + \frac{e^2}{4\pi\epsilon_0}\left(\sum_{i<j}\frac{1}{r_{ij}} + \sum_{k<l}\frac{Z_k Z_l}{r_{kl}} - \sum_i\sum_k \frac{Z_k}{r_{ik}}\right), \tag{2.5}$$

kde proměnné i a j indexují elektrony, k a l indexují atomová jádra, m_e je hmotnost elektronu, m_k hmotnost protonu, ϵ_0 permitivita vakua, Z značí náboj příslušného jádra a r je vzdálenost dvou částic.[15]

Hamiltonův operátor umíme tedy sestavit pro libovolně velikou molekulu, problém je však v následném řešení Schrödingerovy rovnice. Analyticky ji nelze vyřešit ani pro druhý nejjednodušší atom, tedy helium, natož pak pro molekuly. U molekul narážíme velmi rychle i s numerickým řešením, zavadí se proto různé aproximace. Takovou základní aproximací je Bornova–Oppenheimerova aproximace, která je založena na výrazně rozdílné hmotnosti, a tedy i rychlosti elektronů a atomových jader. Oba pohyby proto považujeme za nezávislé a vlnovou funkci můžeme napsat jako součin vlnové funkce jader a vlnové funkce elektronů:

$$\Psi = \Psi_{\text{jad}}\Psi_{\text{el}}. \tag{2.6}$$

V Bornově–Oppenheimerově aproximaci tedy předpokládáme, že na časové škále pohybu elektronů jsou jádra stacionární, a řešíme tzv. elektronovou Schrödingerovu rovnici:

$$\hat{H}_{\text{el}}\Psi_{\text{el}} = E_{\text{el}}\Psi_{\text{el}}. \tag{2.7}$$

kde E_{el} je elektronová energie a elektronový Hamiltonův operátor $\hat{H}_{\text{el}}$ neobsahuje oproti původnímu Hamiltonovu operátoru z rovnice 2.5 člen kinetické energie jader a člen zastupující elektrostatickou repulzi jader se stává konstantním. [27]

Hamiltonův operátor je v tuto chvíli závislý na pozicích jader pouze parametricky a ty se stávají vstupním parametrem výpočtu. Díky tomuto konceptu lze zavést pojem hyperplochy potenciální energie, což je funkční závislost elektronové energie na vnitřních souřadnicích molekuly neboli geometrických

parametrech jako jsou délky vazeb a úhly. V dalším textu bude Bornova–Oppenheimerova aproximace uvažována implicitně a Schrödingerovou rovnicí bude vždy myšlena její elektronová forma.

Náročnost řešení Schrödingerovy rovnice roste exponenciálně s počtem částic v systému a rolí kvantové chemie je tak zavádět další aproximace, aby bylo její řešení časově i paměťově únosné i pro běžné systémy. Při současném stavu poznání a dostupném výpočetním výkonu však stále často nejsme schopni získat kvantitativní výsledky v rozumném čase. Alternativu nám přináší metody strojového učení. Z kvantové mechaniky již víme, že vlnová funkce nám plně definuje daný systém a pro řešení Schrödingerovy rovnice potřebujeme znát jen Hamiltonův operátor, jehož parametry jsou pouze polohy atomových jader a jejich náboje. Existuje zde tedy fyzikálně podložená funkční závislost, která mapuje strukturu molekuly na její vlastnosti. Strojové učení, konkrétně regresní metody, nám umožňuje takovou závislost najít na základě známých učících dat. Ta lze teoreticky získat experimentálně, ale vzhledem k jejich potřebnému množství se budeme pro generování dat obracet opět na kvantovou mechaniku. Následné predikce molekulárních vlastností mohou být pouze tak kvalitní jako použitá učící data. Je důležité poznamenat, že se tím nezbavujeme kvantové mechaniky, fyzika problému je zakódována právě v učících datech.

2.3 Absorpční spektra

Jak bylo uvedeno v kapitole 2.2, fyzikální veličiny bývají v případě kvantových systémů často kvantovány. Stejně tomu je i v případě absorpce. Molekula může nabývat pouze diskrétních hodnot energie a tudíž nemůže absorbovat fotony (kvanta elektromagnetického záření) o libovolné energii, ale pouze fotony, jejichž energie splňuje tzv. Bohrovu rezonanční podmínku. Ta udává, že může být absorbováno/emitováno pouze záření, které odpovídá přesnému rozdílu dvou energetických hladin molekuly:

$$E_2 - E_1 = E_f = h\nu,$$

kde E_1 a E_2 jsou energetické hladiny molekuly a E_f je energie fotonu daná součinem jeho frekvence záření ν a Planckovy konstanty h.[32] Podle energie absorbovaného/emitovaného fotonu rozlišujeme přechody elektronové, vibrační a rotační. Rotační přechody odpovídají nejméně energetickému záření, typicky v mikrovlnné oblasti. Přechody mezi vibračními hladinami pak odpovídají standardně infračervenému záření. Nejenergetičtější záření přísluší elektronovým přechodům, kde se pohybujeme typicky v oblasti viditelného a ultrafialového světla. V případě absorpce fotonu s ještě větší energií, např. v rentgenové oblasti, již dochází k ionizaci molekuly, kdy je elektron zcela "vyražen"z molekuly.[2] V této práci zabývám elektronovými spektry neboli přechody mezi elektronovými hladinami, přičemž v případě absorpce hovoříme o tzv. excitaci elektronu do vyšší hladiny.

2.3.1 Veličiny popisující absorpci

Intenzitu absorpce lze vyjádřit pomocí celé řady veličin. Experimentálně přímo měřitelnou veličinou je transmitance, která je daná jako poměr záření prošlého vzorkem I k původnímu záření zdroje I_0:

$$T = \frac{I}{I_0}. \tag{2.8}$$

Množství pohlceného záření neboli absorbance je pak získáno jako záporně vzatý logaritmus transmitance:

$$A = -\log \frac{I}{I_0}.$$

Obě tyto veličiny jsou však závislé na uspořádání experimentu, neboli na délce měřící nádoby (kyvety) a koncentraci vzorku. Pro získání porovnatelné veličiny se používá Lambertův–Beerův zákon:

$$A = \epsilon c l,$$

kde c je koncentrace vzorku, l je délka kyvety a ϵ je molární absorpční koeficient udávající intenzitu absorpce nezávisle na daném uspořádání. Mikroskopickou veličinou je pak již zmíněný absorpční účinný průřez, který je propojen s molárním absorpčním koeficientem následujícím vztahem:

$$\sigma = \ln(10)\frac{\epsilon}{N_\mathrm{A}},$$

kde N_A je Avogadrova konstanta.[64]

Doposud byly diskutovány experimentální veličiny, teorie však pracuje ještě s jinými veličinami, zmíním alespoň dvě nejpoužívanější. Snadno uchopitelnou veličinou je tzv. oscilátorová síla f. Jedná se o bezrozměrnou veličinu, která udává pravděpodobnost absorpce fotonu o dané vlnové délce, přičemž referencí je klasický harmonický oscilátor o stejné frekvenci. Ta tak prakticky nabývá pro nedegenerované stavy (stavy bez energetických hladin se stejnou energií) hodnot od nuly do jedné.[2] Z této definice je zřejmé, že absorpční spektrum není z pohledu statistiky nic jiného než hustota pravděpodobnosti a pro práci se spektry tak lze s výhodami využít nástrojů strojového učení a aplikované statistiky vyvinutými pro práci s hustotami pravděpodobnosti. Je však nutné poznamenat, že v případě absorpčních spekter se jedná o nenormalizované funkce a je tak s nimi potřeba pracovat.

Druhou používanou veličinou z pohledu teorie je čtverec tranzitního dipólového momentu μ^2. Jedná se o dipólový moment vyvolaný přechodem elektronu do vyšší energetické hladiny. Obě dvě teoretické veličiny jsou propojeny následujícím vztahem:

$$f = \frac{4 m_e \pi \nu}{3 \hbar e^2} \mu^2,$$

kde ν je frekvence záření, m_e je hmotnost elektronu, $\hbar$ je redukovaná Planckova konstanta a e je elementární náboj.[28]

Zatímco kvantově-chemické výpočty poskytují výsledky pomocí oscilátorových sil či tranzitních dipólových momentů, a to vždy pro konkretní geometrii molekuly, my chceme získat experimentálně porovnatelné spektrum reprezentující celkově danou molekulu při dané teplotě. Způsobům, jak toho docílit, se věnuji v následujících sekcích.

2.3.2 Modelování empirickým rozšířením

Z informací v předchozích sekcích vyplývá, že energie může být absorbována pouze po kvantech a spektrum by tedy mělo být tvořeno sérií čar. Tak spektra skutečně vypadají pro atomy a teoreticky i pro izolované molekuly. Nicméně reálná spektra molekul jsou typicky spojitá, to je způsobeno několika příspěvky. Za prvé je každá elektronová hladina rozmělněna na řadu vibračních hladin a každá z nich má dále velké množství rotačních hladin. Běžné měřící přístroje nejsou tak přesné, aby tyto téměř stejné hladiny odlišily. Každá z těchto rotačních hladin je navíc velmi drobně rozšířena díky Dopplerově jevu neboli díky tomu, že molekuly se pohybují v různých směrech a různými rychlostmi. Důležitým příspěvkem je také vzájemná interakce molekul, hlavně pak v kapalné fázi. Energetické hladiny jednotlivých molekul spolu interagují a navzájem se posouvají, čímž vzniká prakticky nekonečné množství hladin. Dalším zdrojem rozšíření jsou tzv. disociativní stavy, kdy absorpce fotonu vede k rozpadu molekuly. Vznikající fragmenty mohou mít libovolnou kinetickou energii a tedy i absorbovaná energie nemusí být kvantována.[2]

Modelovat absorpční spektra lze celou řadou přístupů. Nejjednodušším z nich je modelování pomocí empirického rozšíření. V rámci toho přístupu reprezentujeme molekulu pouze jednou geometrií, tzn. jedním konkretním uspořádáním atomů v prostoru. Standardně se vybírá tzv. minimální neboli rovnovážná geometrie. Tu získáme minimalizací energie systému neboli nalezením energeticky nejvýhodnějšího uspořádání atomů. Pro tuto jednu geometrii získáme pomocí kvantově-chemického výpočtu sadu diskrétních hodnot excitačních energií a příslušných pravděpodobností/intenzit přechodů do jednotlivých excitovaných stavů. Toto čárové spektrum je následně převedeno na spojité spektrum, které by mělo alespoň vzdáleně odpovídat reálné distribuci geometrií při dané teplotě. Toho je docíleno empirickým rozšířením absorpčních čar pomocí Gaussových funkcí:

$$\sigma(E) = \frac{1}{8065{,}54 \cdot 4{,}319 \cdot 10^{-9} \cdot \sqrt{2\pi} H} \sum_{i=1}^{n} f_i \exp\left(\frac{1}{2}\left(\frac{E - E_i}{H}\right)^2\right),$$

kde $\sigma(E)$ je absorpční účinný průřez pro záření o energii E, i indexuje jednotlivé excitované stavy, do kterých se můžeme dostat absorpcí záření, přičemž E_i je excitační energie do daného stavu (neboli rozdíl energetických hladin mezi excitovaným a původním stavem) a f_i je oscilátorová síla udávající intenzitu tohoto přechodu. H zde trochu nestandardně (kvůli kolizi symbolů)

značí empirický parametr rozšíření Gaussovy funkce, to bývá voleno kolem 0,3 eV. Tento přístup je vhodný pro rychlé kvalitativní sestrojení spektra nicméně je nedostatečný pro kvantitativní modelování.[41, 23]

2.3.3 Reflexní princip

Pokud chceme modelovat spektra na kvantitativní úrovni, nevystačíme si bohužel již jen s jedním relativně jednoduchým kvantově-chemickým výpočtem pro jednu geometrii. Jako zástupce pokročilejších přístupů jsem zvolil tzv. reflexní princip, jelikož se jedná stále o snadno uchopitelný přístup, který v řadě případů poskytuje velmi dobré výsledky. Jedná se o tzv. semiklasický přístup, který může být odvozen jako aproximace k plně kvantové simulaci, která je však pro běžné systémy mimo naše výpočetní možnosti.[62, 33] Reflexní princip efektivně zahrnuje tzv. necondonovské efekty (závislost velikosti tranzitního dipólového momentu na geometrii), avšak nezachycuje vibrační rozlišení (přechody do různých vibračních hladin excitovaného elektronového stavu). Je tak vhodný zejména pro modelování disociativních stavů a molekul v kapalné fázi.[41]

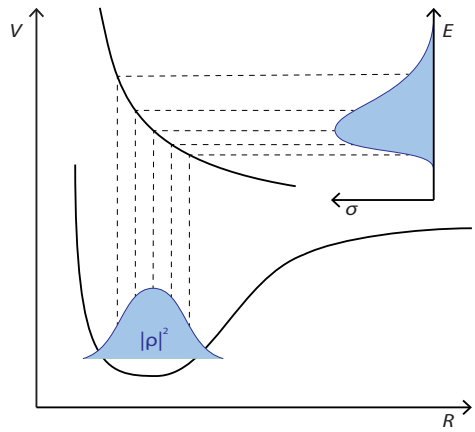
Absorpční účinný průřez v rámci reflexního principu je definován následovně:

$$\sigma(E) = \frac{\pi E}{3\hbar\epsilon_0 c} \sum_{i=1}^{n} \int \rho \mu_i^2 \delta(E - E_i) \mathrm{d}\mathbf{R},$$

kde E_i a μ_i značí excitační energii a tranzitní dipólový moment i-tého excitovaného stavu pro geometrii $\mathbf{R}$, ρ je jaderná hustota základního stavu neboli distribuce geometrií při dané teplotě, ϵ_0 je permitivita vakua a c je rychlost světla. V praxi se integrál přes spojitou distribuci geometrií nahrazuje sumou přes množinu geometrií, která ji reprezentuje. Jedná se tak vlastně o přirozené rozšíření předchozího přístupu, akorát reálnou distribuci geometrií nereprezentujeme jednou jedinou geometrií, ale sadou geometrií čítající typicky stovky až tisíce vzorků. Pro tyto geometrie spočítáme opět excitační energie a intenzity přechodů a z těch modelujeme spektrum některou z metod odhadu hustoty pravděpodobnosti. Schéma reflexního principu je zachyceno na obrázku 2.1.[41]

Množinu geometrií reprezentující hustotu základního stavu při dané teplotě získáme vzorkováním distribuce geometrií pomocí molekulové dynamiky. Buďto je možné použít klasickou dynamiku, kde integrujeme Newtonovy pohybové rovnice, nebo některý pokročilejší přístup zahrnující kvantové efekty jader jako je metoda dráhových integrálů[61] či kvantový termostat[7]. Síly působící na atomy v rámci dynamiky jsou získány opět pomocí kvantově-chemických výpočtů. Data použitá v této práci byla získána za pomocí kombinace metody dráhových integrálů a kvantového termostatu[8] při teplotě simulace 300 K s časovým krokem integrace 20 a.u. (zhruba 0,5 fs).

Dva uvedené přístupy k modelování spekter zároveň zastupují dva různé přístupy k použití strojového učení. Zatímco u modelování empirickým rozšířením trénujeme model na sadě učících molekul a poté predikujeme vlastnosti



Obrázek 2.1: Schematické znázornění reflexního principu. Jaderná hustota základního stavu je přes hyperplochu excitovaného stavu reflektována až na energetickou osu spektra.

molekul mimo učící sadu, reflexní princip zastupuje přístup modelování s použitím většího množství geometrií, model tedy trénujeme na sadě geometrií vybrané molekuly a pak predikujeme vlastnosti pro další geometrie té samé molekuly. Ačkoli by se mohlo zdát, že pro učení samotné to není podstatné, ve výběrů či návrhu molekulárních reprezentací to hraje zásadní rozdíl, jak bude ukázáno dále.

2.4 Statistické zpracování spektrálních dat

Jak již bylo naznačeno, spektrum budeme rekonstruovat jako hustotu pravděpodobnosti. A jako s hustotou pravděpodobnosti s ním dále můžeme i pracovat, díky tomu můžeme použít zavedené statistické nástroje pro odhad statistické chyby či porovnaní shody dvou spekter.

2.4.1 Odhad hustoty pravděpodobnosti

Excitační energie a intenzity/pravděpodobnosti přechodu pro sadu geometrií je třeba převést na spojité spektrum metodou odhadu hustoty pravděpodobnosti. Nejjednodušším způsobem je metoda histogramů, akorát zde pro každý interval/sloupec sčítáme tranzitní dipólové momenty jako váhy místo jednotkových příspěvků v klasickém histogramu:

$$\sigma_i = \frac{\pi E_i I_i}{3\hbar\epsilon_0 c N \Delta E}, \tag{2.9}$$

kde I_i je suma tranzitních dipólových momentů spadajících do daného intervalu energií se střední energií E_i a ΔE je šířka intervalů, N je počet geometrií a konstanty jsou stejné jako v rovnici 2.3.3.[41] Tento přístup má však relativně velký poměr šumu ku signálu, je citlivý na zvolení počátku a

šířku intervalů. Malou změnou některého z parametrů může dojít ke skokové změně ve výsledném histogramu. Část informace navíc ztrácíme nahrazením energie dané geometrie střední energií intervalu.

Proto zde používám pokročilejší přístup jádrového odhadu hustoty pravděpodobnosti.[43] V principu je to také relativně jednoduchý přístup, kdy každý bod nahrazujeme jádrovou funkcí (shoda s názvem u regresních metod založených na jádrové transformaci není náhodná) a výslednou hustotu pravděpodobnosti získáme jako sumu přes tyto funkce. Obecný vzorec pro jádrový odhad hustoty pravděpodobnosti má následující tvar:

$$f_H(x) = \frac{1}{NH} \sum_{j=1}^{N} K\left(\frac{x - x_j}{H}\right), \tag{2.10}$$

kde N je počet bodů, K představuje právě onu jádrovou funkci a H je hyperparametr udávající typicky její šířku. Na jádrovou funkci máme nějaké elementární požadavky: musí být nezáporná, mít průměr v nule a její integrál musí být roven jedné. Jelikož zdroje rozšíření spekter popsané v sekci 2.3.2 mají převážně gaussovský charakter, používám zde normalizovanou Gaussovu funkci. Po dosazení vztahu 2.3.3 dostáváme následující vzorec:

$$\sigma_H(E) = \frac{1}{NH\sqrt{2\pi}} \sum_{j=1}^{N} \sum_{i=1}^{n} \frac{\pi E_{i,j} \mu_{i,j}^2}{3\hbar\epsilon_0 c} \exp\left(\frac{1}{2}\left(\frac{E - E_{i,j}}{H}\right)^2\right), \tag{2.11}$$

kde vnější suma jde přes jednotlivé geometrie, zatímco vnitřní suma jde přes jednotlivé excitované stavy.

Stejně jako u regresních metod tu musíme u jádrové funkce nastavit parametr rozšíření H. To lze udělat několika způsoby. Pro unimodální (s jedním maximem) spektra tvořená jedním excitovaným stavem lze použít jednoduché Silvermanovo empirické pravidlo:

$$H = \left(\frac{4s^5}{3N}\right)^{\frac{1}{5}}, \tag{2.12}$$

kde s je směrodatná odchylka.[56] Alternativně existuje řada časově náročnějších způsobů, jak tento parametr nastavit exaktněji. Pro vyhodnocení spekter v této práci používám křížovou validaci v případě, že Silvermanovo pravidlo nepostačuje.

2.4.2 Statistická chyba modelování

Aby byl zřejmý přínos použití strojového učení, je důležité si vůbec ukázat, jak veliké chyby se dopouštíme modelováním spektra z konečného počtu geometrií. Vhledem k tomu, že spektrum zde představuje neparametrickou distribuci pravděpodobnosti, je vhodné použít některou z metod převzorkování pro odhad intervalů spolehlivosti. Já zde používám *bootstrap* metodu, která spočívá v tom, že původní spojitou distribuci nahradíme empirickou distribucí

neboli prohlásíme navzorkované excitační energie a pravděpodobnosti absorpce za distribuci, kterou používáme zdroj dalších dat. Z této distribuce potom vybíráme s opakováním nové vzorky o původní velikosti. Pokud mám tedy excitační energie a tranzitní dipólové momenty například pro tisíc geometrií, tak z nich vybírám vždy opět tisíc vzorků, avšak s tím, že se mohou opakovat. Některé tedy vybereme vícekrát a některé vůbec. Tento proces opakujeme a pro každou takto převzorkovanou sadu modelujeme spektrum. V této práci používám konkretně cirkulární blokový *bootstrap*, kde nejsou vzorky vybírány z empirické distribuce po jednom, nýbrž po blocích. Tímto způsobem vezmeme v potaz drobnou časovou korelaci, kterou v datech máme z molekulové dynamiky.[40]

Tímto způsobem získáme velké množství spekter a pro každý bod tak můžeme jednoduše spočítat výběrovou směrodatnou odchylku:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Ze Studentova t rozdělení poté získáme koeficient t, kterým musíme vynásobit směrodatnou odchylku pro získání intervalu spolehlivosti na zvolené hladině významnosti:[40]

$$(\bar{x} - ts, \bar{x} + ts).$$

2.4.3 Porovnání distribucí

Abychom byli schopní kvantifikovat přínos strojového učení, je nutné nějakým způsobem porovnat, jak se od sebe dvě spektra liší. Jak již bylo několikrát zmíněno, spektrum je distribuce pravděpodobnosti a pro ty máme k dispozici řadu porovnávacích kritérií jako je například Kolmogorovův–Smirnovův test, Kuiperův test nebo Jensenova–Shannonova divergence.[5] Volbou testu si zároveň vybíráme, co chceme vlastně hodnotit. Zda nám jde o relativní či absolutní hodnoty, zda nás zajímají maximální rozdíly či integrální odchylky atd.

V našem případě máme jedno spektrum/distribuci získanou modelováním z velkého množství geometrií, kterou bereme jako pravdivou hustotu pravděpodobnosti, které se chceme co nejvíce přiblížit. Vybral jsem proto Kullbackovu–Leiblerovu (KL) divergenci, která je asymetrická a odchylky váží pravděpodobností prvního ze spekter:

$$D_{\mathrm{KL}}(P||Q) = \int_{-\infty}^{\infty} p(x)\frac{p(x)}{q(x)}\mathrm{d}x.$$

Pro distribuce vzorkované na stejné mřížce můžeme použít místo integrálu sumu:[16]

$$D_{\mathrm{KL}}(P||Q) = \sum_{x} P(x)\frac{P(x)}{Q(x)}.$$

Problém je v tom, že tato divergence počítá s normalizovanými hustotami pravděpodobnosti. My však nemůžeme spektra normalizovat, jelikož kdyby bylo jedno spektrem násobkem druhého, tak to pro nás nemůže znamenat to samé. Nicméně z obecnější α divergence lze odvodit nenormalizovaná verze KL divergence:[16]

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x)\frac{P(x)}{Q(x)} - P(x) + Q(x).$$

Pro lepší porovnatelnosti divergencí pro úplně odlišné dvojice spekter ještě obě spektra vždy normalizuji sumou prvního z nich (je důležitý rozdíl oproti normalizaci každého zvlášť).

Kapitola 3

Analýza současného stavu

V této kapitole budou probrány aktuálně publikované molekulární reprezentace, dostupná testovací data a také nasazení strojového učení učení v kvantové chemii včetně přehledu predikcí spektroskopických vlastností různými výzkumnými skupinami.

3.1 Molekulární reprezentace

Forma molekulární reprezentace či molekulárního deskriptoru $\mathbf{x}$ představuje v oblasti kvantové chemie prakticky nejdůležitější součást regrese.[19] Používané reprezentace můžeme rozdělit dle dvou základních kritérií. Za prvé je důležité, zda reprezentace kóduje kompletní trojrozměrnou geometrii molekuly nebo je založena pouze na molekulárních grafech, tzn. bere v potaz pouze konektivitu.[14] V rámci této práce zkoumám spektroskopické vlastnosti i pro různé geometrie jedné molekuly, je tedy důležité aby reprezentace umožňovala takové učení. Za druhé může reprezentace kódovat jednotlivé atomy v jejich chemickém prostředí (funkce symetrie[Behler2011], hladký překryv atomovách pozic [Bartók2013, Grisafi2018]) nebo molekulu jako celek. První možnost je vhodná pro predikci lokálních vlastností, jako jsou síly působící na daný atom či chemický posun. Druhá možnost je naopak vhodnější pro predikci globálních vlastností molekuly jako celku, kam spadají i zkoumané spektroskopické vlastnosti jako excitační energie či intenzita absorpce.[31] Při použití lokálních prostředí je potřeba dát si obzvláště pozor na predikci intenzivních globálních vlastností, které jsou projevem molekuly jako celku, ale nelze je jednoduše poskládat z atomových příspěvků, což je právě náš případ. Dále jsou tedy probírány pouze reprezentace kódující trojrozměrnou strukturu pro molekulu jako celek.

Strukturu molekuly potřebujeme tedy nějakým způsobem zakódovat do vektoru, a to takovým způsobem, aby daná reprezentace splňovala určité požadavky. Hlavním z nich je invariance vůči translaci, rotaci a záměně dvou stejných atomů. Například soubor kartézských souřadnic atomů tak není vhodným kandidátem, neboť pro stejnou molekulu, jen jinak orientovanou v prostoru, bychom dostali odlišný vektor, což neodpovídá fyzikální realitě. Na

druhou stranu chceme, aby se změny ve struktuře molekuly, které mají vliv na zkoumanou vlastnost, projevily i ve vstupním vektoru neboli vyžadujeme unikátnost. Dále je žádoucí, aby reprezentace byla spojitá a diferencovatelná, jelikož model by měl být hladký, aby mohla regrese správně fungovat. Reprezentace by také měla být schopna zakódovat libovolnou molekulu a její výpočet by měl být rychlý, jelikož rychlost je prvotní důvod k použití ML. Posledním kritériem je efektivita neboli kolik učících dat potřebujeme, abychom dosáhli požadované přesnosti. Zde záleží na preferencích, respektive množství dostupných učících dat. Některým reprezentacím stačí málo učících dat pro kvalitní odhady, ale s dalším množstvím dat se už moc nezlepšují. Některé naopak potřebují veliké množství dat pro jakýkoliv solidní odhad, ale lze je dále systematicky zlepšovat pomocí větší trénovací množiny. Většinou však budeme chtít pracovat s co nejmenším množstvím dat, protože jejich získání je typicky výpočetně velmi drahé. Další výhodou je, pokud má reprezentace konstantní velikost, jelikož to usnadní práci s daty a učení napříč různými velikostmi molekul.[31] Některé reprezentace umožňují kódování periodických systémů jako jsou krystaly, takové systémy však nejsou cílem této práce.

V současné době existuje řada používaných molekulárních reprezentací spadajících do námi zvolené kategorie. Takovou základní reprezentací je coulombovská matice[51] (CM), která kóduje strukturu molekuly pomocí elektrostatistických interakcí:

$$M_{i,j} = f(\mathbf{x}) = \begin{cases} 0{,}5Z_i^{2,4} & i = j \\ \frac{Z_i Z_j}{r_{i,j}} & i \neq j \end{cases},$$

kde Z_i je náboj jádra a $r_{i,j}$ je vzdálenost dvou jader. CM však nesplňuje podmínku invariance, jelikož je závislá na uvedeném pořadí atomů v molekule. Toto lze řešit různými způsoby, například použitím vlastních hodnot matice místo jejích prvků, taková reprezentace však porušuje podmínku unikátnosti. Další možností je seřazení řádků matice podle jejich normy, taková reprezentace ale zase porušuje požadavek spojitosti. Je zřejmé, že CM sice není optimální, nicméně je velmi jednoduchá a je z ní odvozena řada dalších reprezentací. Takovou reprezentací je i relativně úspěšná BoB[25] reprezentace (z angl. *Bag of Bonds*), kde jsou mimodiagonální prvky CM nejprve roztříděny do jednotlivých skupin podle zúčastněných atomů (CH, CC atd.) a v těchto skupinách jsou seřazeny podle velikosti. Velikost každé skupiny je určena jako maximální počet dané interakce v molekulách datasetu a pro molekuly s menším počtem interakcí je skupina doplněna nulami. Rozšířením BoB o interakce vyššího řádu je BAML[29] reprezentace (z anglického *Bonds and Angles Machine Learning*). Tříatomové interakce představují typicky úhly, čtyř-atomové pak torzní úhly. Reprezentace $F_{2\mathrm{B}}$ a $F_{3\mathrm{B}}$ jsou založeny na podobném principu, akorát obsahují pro každou dvojici či trojici atomů ne jednu, ale vždy celou sadu interakcí, které se liší pouze různým umocněním meziatomových vzdáleností.[46]

Všechny doposud uvedené reprezentace škálují s počtem interakcí v molekule, což odpovídá druhé mocnině počtu atomů pro CM, BoB a $F_{2\mathrm{B}}$ a dokonce

třetí mocnině pro F_{3B} a BAML (u BAML dokonce čtvrté mocnině při zahrnutí torzních úhlů). Tento nešvar se snaží řešit další reprezentace, které místo výpisu všech interakcí v molekule zrekonstruují distribuci těchto interakcí a jako deskriptory používají vzorkování této distribuce. Tyto reprezentace tak škálují lineárně nikoliv s počtem interakcí, ale s počtem typů interakcí (neboli nezáleží na tom, kolikrát se daná n-tice atomů v molekule vyskytuje), a jejich velikost je tak prakticky konstantní vzhledem k počtu atomů. Reprezentace kódovaných vazeb[14] (z anglického *Encoded Bonds*) takto upravuje BoB pomocí odhadu hustoty pravděpodobnosti, kdy pro každý typ interakce (CH, CC atd.) získáme její distribuci v molekule a přímo z ní nebo z její kumulativní distribuční funkce získáme vzorkováním výslednou reprezentaci. Pro každý pár prvků (nikoliv však každý pár atomů) pak dostaneme sadu deskriptorů, které mají následující tvar:

$$v_i = \sum_{j<k} f(r_{j,k}, d_i, \beta),$$

kde index i označuje pořadí bodu na vzorkovací mřížce, d_i je hodnota tohoto bodu a $r_{j,k}$ je vzdálenost atomů j a k příslušících k danému typu interakce. Funkce f vrací hodnoty v jednotkovém intervalu a její hladkost závisí na parametru β, v případě odhadu hustoty pravděpodobnosti se může jednat například o Gaussovu funkci a parametr β představuje její rozšíření. Reprezentaci lze doplnit o maximální geodetickou vzdálenost, kdy do distribuce započteme pouze páry atomů, jejichž vzdálenost v počtu vazeb je menší než dané číslo.

Podobný přístup používá i SLATM[30] reprezentace (z angl. *Spectrum of London and Axilrod-Teller-Muto potential*), která místo distribuce vzdáleností v molekule používá metriku založenou na radiálních distribučních funkcích potenciálů a přidává i tříatomové interakce. Pro dvouatomové interakce je použit Londonův potenciál, pro tříatomové pak ATM[3, 39] potenciál. FCHL[18] reprezentace (*Faber-Christensen-Huang-Lilienfeld*) dále rozšiřuje tento přístup použitím multidimenzionálních distribucí a definováním chemických prvků pomocí umístnění v periodické soustavě prvků namísto náboje jádra, což umožnuje predikovat vlastnosti molekul obsahující atomy, které nejsou zahrnuty v učících datech. Další moderní reprezentací založenou na distribucích je MBTR[31] (z angl. *Many-Body Tensor Representation*), která je pro dvouatomové interakce velmi podobná kódovaným vazbám, akorát používá inverzní hodnotu vzdáleností a přidává funkci vážící interakce, která například snižuje příspěvek k distribuci velmi vzdáleným atomům. Tato reprezentace zároveň volitelně přidává tříatomové interakce formou distribucí úhlů.

3.2 Datové sady

V rámci této práce aplikuji ML v oblasti elektronové spektroskopie. Molekulárními vlastnostmi mého zájmu jsou tedy excitační energie a příslušné intenzity přechodu. Pro testování těchto veličin existují v tuto chvíli dvě

používané datové sady, přičemž obě obsahují pro každou molekulu pouze jednu geometrii, a to geometrii optimalizovanou (s minimální energií). Zatím není k dispozici žádná datová sada obsahující uvedené vlastnosti pro velké množství geometrií jedné molekuly.

Datová sada QM7b[4, 38] obsahuje data pro 7211 organických molekul s až 23 atomy (maximálně však 7 těžkých atomů typu C, N, O nebo S). Součástí této sady je 14 různých veličin včetně excitační energií pro 1. excitovaný stav, excitačních energií v maximu absorpce molekuly a intenzit přechodu pro maximum absorpce. Ty byly pro tuto sadu získány semiempirickou kvantově-chemickou metodou ZINDO/S (z angl. *Zerner's Intermediate Neglect of Differential Overlap for Spectroscopy*). Jedná se o velmi rychlou aproximativní metodu, nicméně pro potřeby testování je dostačující.

Datová sada QM8[49, 48] obsahuje data pro 21786 organických molekul s až 8 těžkými atomy (C, O, N, F). Sada obsahuje excitační energie a oscilátorové síly pro první dva excitované stavy získané metodami časově závislé teorie funkcionálu hustoty (TDDFT, z angl. *Time-Dependent Density Functional Theory*) a aproximativní metodou spřažených klastrů druhého řádu (CC2, z angl. *Second-Order Approximate Coupled-Cluster*). Metoda CC2 je považována za přesnější a v této datové sadě je vedena jako referenční. Metoda CC2 škáluje s velikostí systému jako $O(n^5)$,[12] zatímco DFT metody škálují typicky s $O(n^3)$.[36]

Pro testování predikcí pro modelování za pomocí reflexního principu jsem si zvolil dvě molekuly, u kterých mám k dispozici vstupní data neboli výpočty z kvantově-chemických programů.[58, 57] První molekulou je nejjednodušší Criegeeho intermediát CH_2OO. Jedná se o významnou atmosférickou molekulu, která se podílí na odbourávání těkavých organických látek z troposféry, ale její experimentální i teoretická charakterizace je velmi složitá.[60, 54] U této molekuly přispívá k absorpci (má nezanedbatelnou intenzitu absorpce v dané spektrální oblasti) pouze jeden (a to druhý) excitovaný stav, nicméně vzhledem ke komplikované elektronové struktuře se ukazují veškeré běžně dostupné kvantově-chemické metody jako nedostatečné.[58] Je proto potřeba použít velmi časově náročné výpočetní metody a chtěli bychom tak zredukovat počet geometrií, pro které je nutné počítat excitační vlastnosti. Zde použitá data pro excitované stavy byla získána již zmíněnou metodou CC2 a také aproximativní metodou spřažených klastrů třetího řádu (CC3, z angl. *Third-Order Approximate Coupled-Cluster*). Přestože tato metoda má v názvu slovo "aproximativní", jedná se o velmi pokročilou metodu, jejíž výpočetní náročnost škáluje s velikostí systému jako $O(n^7)$.[12] Z hlediska strojového učení stačí v tomto případě vytvořit jeden model pro excitační energii a jeden model pro predikci intenzit absorpce.

Druhou molekulou je azobenzen (resp. její trans izomer), který patří do skupiny fotochromních molekulárních přepínačů. Ty při ozáření světlem o vhodné vlnové délce přechází mezi jednotlivými izomery s různými vlastnostmi. Azobenzen tak nachází uplatnění v materiálovém inženýrství,[53, 6] přičemž za jeho použití v molekulárních strojích byla v roce 2016 dokonce udělena

Nobelova cena za chemii. Azobenzen bude představovat jakýsi zatěžkávací test, jelikož se z hlediska strojového učení jedná o velmi složitý problém. Za prvé se jedná již o větší molekulu (24 atomů), což znamená větší či složitější molekulární deskriptory. Za druhé k absorpčnímu spektru přispívá hned několik excitovaných stavů a za třetí pro různé geometrie jsou tyto stavy jinak seřazeny z hlediska intenzit absorpce. Kvantově-chemické programy totiž řadí elektronové stavy pouze podle jejich excitační energie, což není v tomto případě žádoucí z hlediska učení. Když například program pro jednu geometrii zařadí na druhé místo elektronový stav, který byl pro jinou geometrií na prvním místě, tak se s tím prohodí i příslušné oscilátorové síly a my tak učíme model pro oscilátorové síly na nekonzistentních datech. Když bychom geometrii měnily spojitě, tak by najednou došlo ke skokové změně (prohození) oscilátorových sil. Menší problém toto řazení způsobuje i z hlediska predikce excitačních energií, kde sice nezpůsobuje přímo nespojitosti se změnou geometrie, ale způsobuje nespojitosti první derivace, přičemž takové funkce se hůře učí. Pro ideální učení by bylo nejprve nutné stavy seřadit dle jejich fyzikální podstaty, pak by se excitační energie i oscilátorové síly měnily se změnou geometrie hladce. Tomuto řazení se říká diabatizace. Jedná se však o relativně náročný kvantově-chemický problém, který nezapadá do zaměření této práce. Molekula azobenzenu tak bude představovat nejhorší scénář a budeme chtít zjistit, zda i tak bude možné získat použitelná data. Excitační energie a tranzitní dipólové momenty byly pro tuto molekulu získány semiempirickou metodou ZINDO/S a také metodou TDDFT s funkcionálem CAM-B3LYP.[57] Jedná se o spíše méně výpočetně náročnou metodu, azobenzen je ale z hlediska kvantové chemie už relativně veliká molekula. Jeden takový výpočet vyjde na jednom výpočetním jádru na několik desítek minut v závislosti na sestavě a dalších parametrech výpočtu. Je důležité si ale uvědomit, že běžně takových výpočtů potřebujeme tisíce.

3.3 Strojové učení v kvantové chemii a spektroskopii

Strojové učení zažívá enormní rozmach a v kvantové chemii tomu není jinak.[52] ML přístupy jsou běžně používány pro predikci molekulárních vlastností jako jsou atomizační entalpie[17], slučovací entalpie[65], energie nulového bodu, tepelné kapacity[18] či polarizovatelnosti[24], a to s chemickou přesností. Tyto techniky také nalezly uplatnění ve výzkumu materiálů, například při vývoji organických solárních článků[47] či pro predikci indexů lomu v optice[1]. Strojové učení se také ukázalo býti užitečné v oblastech chemie, kde *ab initio* přístupy (přístupy založené pouze na fyzikálních zákonech) selhávají vzhledem ke komplikovanosti problému. K takovým oblastem patří například predikce rozpustností[35] nebo krystalových mřížek[34, 22]. Důležitou oblastí, kde také nachází uplatnění regresní modely je molekulová dynamika. ML přístupy umožňují konstrukci flexibilních silových polí a hyperploch potenciální energie bez fyzikálních aproximací.[10] V neposlední řadě může strojové učení

nalézt uplatnění také v oblasti spektroskopie. Například Gestegger a spol. byli schopní úspěšně modelovat infračervená spektra.[21], nicméně v elektronové spektroskopii zatím strojové učení používáno není.

Při predikci excitačních energií a oscilátorových sil (či jiné ekvivalentní veličiny udávající intenzitu/pravděpodobnost absorpce) pomocí strojového učení narážíme na několik problémů. Například se jedná o globální intenzivní veličiny, čili není vhodné používat reprezentace tvořené lokálními chemickými okolími. Ramakrishnan a kol.[48] se snažili docílit co nejlepšího výsledku pomocí tzv. Δ ML přístupu, kdy se hledaná veličina spočítá pomocí rychlé kvantové metody a pomocí strojové učení se doplňuje pouze korekce mezi touto a vybranou přesnější metodou. Podstatou tohoto přístupu je fakt, že většina fyzikální podstaty je zahrnuta i ve velmi aproximativních metodách a za posledních pár procent přesnosti získaných kvalitnější metodou zaplatíme neúměrně větší cenu, přičemž rozdíl mezi dvěma kvantovými metodami se lze snadněji naučit pomocí strojového učení. Ramakrishnan a spol. byli takto schopni dosáhnout na datové sadě QM8 střední absolutní chyby 0,1 eV (vzhledem k přesnější metodě) pro excitační energie při použití 5 tisíc molekul pro učení, avšak narazili u oscilátorových sil, jelikož mezi oběma metodami docházelo k prohození pořadí elektronových stavů. Tentokrát jde o jiné pořadí stavů mezi dvěma metodami použitými v rámci Δ ML, důsledek je ale stejný jako u výše popsaného prohození stavů pro různé geometrie. Ramakrishnan a spol. se navíc zabývali predikcí pouze prvního excitovaného stavu.

Pronobis a kol.[45] zkoušeli přímé strojové učení za použití různých molekulárních reprezentací, byli však schopni dostat pouze na úroveň střední absolutní chyby 0,5 eV pro 1. excitovaný stav a 0,4 eV pro 2. excitovaný stav, přičemž oscilátorovými silami se nezaobírali vůbec. Tyto výsledky se vztahují k datové sadě QM8 při učení na 10 tisících molekulách. Na první pohled lepší výsledky poskytují predikce pro datovou sadu QM7b. Collins a kol.[14] uvádějí střední absolutní chybu 0,23 eV pro 1. excitační energii při učení na necelých 6 tisících molekulách. Pronobis a kol.[46] pak uvádějí střední absolutní chybu 0,15 eV při učení na 5 tisících molekulách. Huang a kol.[29] uvádějí dokonce 0,13 eV taktéž při použití 5 tisíc molekul. Stejnou hodnotu uvádí i Montavon a kol.[38] pro opět stejné množství učících dat. I tato nejlepší hodnota je však stále nedostatečná, 0,13 eV odpovídá 3 kcal/mol, přičemž cílem je tzv. chemická přesnost (1 kcal/mol). Chyba 0,13 eV může znamenat ve viditelné oblasti posun spektra o desítky nanometrů. Rovněž je zřejmé, že zatím nebyl vůbec vyřešen problém predikce oscilátorových sil. Výše uvedené práce se také zabývají pouze výpočty pro optimalizované geometrie sady molekul, ve své práci se však chci do budoucna zabývat i výpočty pro sadu různých geometrií jedné molekuly, které jsou potřebné pro přesné modelování spekter.[58] Dalším požadavkem je predikce většího množství excitovaných stavů.

Kapitola 4

Návrh a implementace

Za účelem vývoje molekulárních deskriptorů a použití strojového učení pro predikci zde zkoumaných vlastností jsem napsal program v jazyce Python 3. Využil jsem přitom ve velkém rozsahu objektově orientovaného programování a scikit-learn knihovny.[44] Všechny klíčové části programu jsou paralelizované, takže je možné učení provádět velmi efektivně na výpočetních klastrech. Program implementuje nativně KRR metodu, ale ta lze jednoduše zaměnit za SVR, GPR či některý z lineárních algoritmů. Jednotlivé reprezentace jsou implementovány pomocí hierarchie tříd, které lze přímo použít pro učení i predikci. Třídy reprezentací zároveň implementují scikit-learn rozhraní, takže je možné je použít přímo jako vstup funkcí z této knihovny, toho lze využít například pro rozličné způsoby optimalizace hyperparametrů či různé metody validace modelu. Program je navržen pro co nejefektivnější vývoj nových reprezentací. Třídy jsou od sebe odvozovány tak, že při vývoji nové reprezentace stačí nastavit co nejpodobnější rodičovskou třídu a přepsat či doplnit pouze rozdílné metody.

Program umožňuje počítání reprezentací za chodu neboli při učení modelu, což je vhodné při optimalizaci jejich parametrů. Zároveň však umožňuje předpočítat si reprezentace, jejich normy nebo přímo jádrovou matici. Rozhodnutí, co konkretně si předpočítat, záleží hlavně na tom, jaké všechny parametry chceme optimalizovat. Při optimalizaci pouze obou hyperparametrů učení (váha regularizačního členu a šířka jádrové funkce) si můžeme předpočítat normy. Pokud chceme optimalizovat i použitou jádrovou funkci, tak si můžeme předpočítat alespoň reprezentace. Druhým hlediskem je paměťová složitost v případě velkých datových sad. Zatímco velikost matice norem či jádra je daná druhou mocninou počtu molekul, reprezentace škálují s počtem molekul lineárně, ale zase mají obecně nespecifikovaný druhý rozměr. Z praktického hlediska je nejzdlouhavější výpočet reprezentací za chodu, jelikož se výpočet opakuje pro každou kombinaci parametrů.

Vyjma zde vyvíjených níže popsaných reprezentací jsem také pro testování implementoval rozhraní pro dostupné reprezentace probrané v sekci 3.1. Kódované vazby byly zahrnuty z knihovny MolML[13], SLATM reprezentace byla použita z programu QMLcode[11] a MBPT reprezentace byla použita z programu qmmlpack[31], který je zatím neveřejně vyvíjen Dr. Mathiasem

Ruppem (*Fritz Haber Institute of the Max Planck Society, Berlin, Germany*) a z toho důvodu není součástí přiloženého programu.

Hyperparametry KRR metody byly v této práci vždy optimalizovány pomocí 5-násobné křížové validace na logaritmické mřížce o velikosti 10x10 s rozlišením jednoho řádu. arametry reprezentací byly buďto použity výchozí d(případě stabilních parametrů) nebo byly optimalizovány také pomocí křížové validace, avšak kvůli velké výpočetní složitosti odděleně od KRR hyperparametrů (pro dva prametry reprezentace a 10 možností na parametr by to znamenalo 10^4 kombinací parametrů).

4.1 Molekulární deskriptory

Výběr či návrh molekulárních deskriptorů se velmi liší pro oba případy modelování spekter. Pro modelování empirickým rozšířením budeme chtít deskriptory, které vhodným způsobem zajistí, mimo jiné vlastnosti popsané v sekci 3.1, i invarianci vůči permutaci atomů stejného typu, neboť vlastnosti molekul by neměly záviset na tom, zda prohodíme dva atomy stejného typu. Dále bude žádoucí aby tyto deskriptory měly konstantní velikost pro všechny molekuly, to umožní snadnější učení mezi molekulami o různých velikostech.

Naproti tomu při modelování reflexním principem či jiným přístupem, který potřebuje velké množství geometrií, učíme model na geometriích jedné molekuly a tudíž máme konstantní velikost i pořadí atomů v molekule předem zaručeno. Použití reprezentací založených na distribucích interakcí tedy nejspíše povede v tomto případě pouze ke ztrátě části informace způsobené rekonstrukcí distribuce a jejím převzorkováním. Nicméně převzorkování pomocí distribuce by pro učení na menším množství geometrií mohlo teoreticky vést k lepším výsledkům v případě jinak enormně velikých deskriptorů, například pro velmi rozsáhlou molekulu. V takovém případě by totiž převzorkovaní mohlo díky menší velikosti poskytnout jednodušší model. Stejně tak řazení atomů v deskriptoru může vést ke ztrátě informace. Například v BoB reprezentaci rozřazujeme coulombovské interakce dle typu zúčastněných atomů, což by samo o sobě nevadilo, problém však může být řazení podle velikosti uvnitř každé takové skupiny, jelikož pořadí velikostí interakcí se můžeme pro jednotlivé geometrie lišit.

4.1.1 Matice interakcí

Maticí interakcí $\mathbf{M}$ zde pro n atomů nazývám matici o velikosti $n \times n$, kde prvek $\mathbf{M}_{i,j}$ kóduje interakci mezi atomy i a j. Pakliže jsou prvky matice $\mathbf{M}_{i,j}$ závislé na mezijaderné vzdálenosti zúčastněných atomů, jedná se o jeden z nejjednodušších způsobů kódování molekulové struktury. Příkladem je již zmiňovaná coulombovská matice. Z důvodů popsaných výše lze očekávat od molekulárních reprezentací s touto strukturou dobré výsledky pro učení na sadě geometrií. Mimodiagonální členy mají pro coulombovskou matici

tvar $Z_i Z_j r_{i,j}^{-1}$. Exponent v mocnině u mezijaderných vzdáleností lze zjevně zvolit i jinak a jeho optimum se může teoreticky lišit pro různé vlastnosti. Záporné mocniny odpovídají potenciálům čili lze očekávat jejich dominanci pro predikci energií.

Jak bylo řečeno v sekci 2.3.1, jednou z veličin popisujících intenzitu/pravděpodobnost absorpce je tranzitní dipólový moment. Koukněme se nejprve na elektrický dipólový moment jako takový. Z fyziky ze střední školy víme, že velikost dipólového momentu pro dvě částice ve vzdálenosti r s opačnými náboji q je dána jako $p = rq$. Spojitě pro prostorovou distribuci náboje ρ dostaneme vzorec pro vektor dipólového momentu:

$$\mathbf{p}(\mathbf{r}_\mathrm{o}) = \int_V \rho(\mathbf{r})(\mathbf{r} - \mathbf{r}_\mathrm{o})\mathrm{d}V,$$

kde r_o značí bod pozorování. V případě celkově neutrálního systému je dipólový moment nezávislý na bodu pozorování, jelikož ten je zintegrován na nulu. Pro bodové náboje pak přejde integrál na sumu. Tranzitní dipólový moment je v kvantové chemii dán pro soubor bodových nábojů a pro přechod z elektronového stavu a do stavu b následovně:

$$\mu_\mathrm{ab} = \int \psi_\mathrm{b}^*(\mathbf{r_1}, \mathbf{r_2}, ...)(\mathbf{r_1} q_1 + \mathbf{r_2} q_2 + ...)\psi_\mathrm{a}(\mathbf{r_1}, \mathbf{r_2}, ...)\mathrm{d}^3\mathbf{r_1}\mathrm{d}^3\mathbf{r_2}...,$$

kde ψ_a představuje vlnovou funkci základního stavu a ψ_b vlnovou funkci excitovaného stavu. Vidíme, že tranzitní dipólový moment je sice daný výchozí a cílovou vlnovou funkcí, nicméně v tomto i v předchozích vzorcích vystupují vzdálenosti přímo a ne jako záporná mocnina. Proto se zaměřím na modifikaci coulombovské matice, kde násobím náboje jader přímo vzdálenostmi.

Náboje jader používané v coulombovské matici však dobře nereprezentují nábojové rozdělení v molekule, koneckonců jsou všechny kladné. Rozdělení nábojů lépe definují elektronegativity jednotlivých atomů. Jedná se o veličinu popisující schopnost atomu přitahovat elektrony resp. elektronovou hustotu. Jedná se o tabelovanou veličinu pro jednotlivé atomy čili nemusíme nic počítat. Zavadím takto matici elektronegativit, kde nahrazuji oproti coulombovské matici náboje jader elektronegativitami a opět testuji i kladný exponent u vzdáleností.

Pro případ učení napříč molekulami, kde je vhodné zajistit invarianci vůči permutaci atomů, je možné seřadit matice interakcí podle norem jednotlivých řádků či sloupců. Další možností je použít k učení pouze vlastní čísla matice.[26] Obě dvě varianty jsem implementoval, nicméně v takovém případě lze očekávat dominanci pokročilejšího řazení, kterým je BoB reprezentace.

4.1.2 Množiny interakcí

Stejným způsobem, jako je BoB reprezentace rozšířením coulombovské matice, definuji obecněji reprezentace založené na množinách interakcí jako rozšíření obecné matice interakcí. Jinak řečeno, elementy z matice interakcí rozdělím

do skupin podle atomů zúčastněných v interakci (CH, CC atd.) a v těchto skupinách je seřadím dle velikosti. Tímto způsobem lze jednoduše odvodit z reprezentací uvedených v předchozí sekcí nová sada reprezentací, která bude vhodnější pro učení napříč molekulami, jelikož má stabilnější řazení interakcí. Implementuji tak rozšíření couombovské matice i matice elektronegativit

4.1.3 Distribuce interakcí

Dalším systematickým rozšířením je transformace množin interakcí na sadu distribucí interakcí. Pro každou množinu interakcí odpovídající jednomu typu interakce (CH, CC, ...) získám odpovídající distribuci. Použiji k tomu opět metody odhadu hustoty pravděpodobnosti popsané v sekci 2.4.1. Bylo by možné použít histogramy, nicméně pro lepší učení je vhodné použít spojitou a hladkou křivku, použiji proto znovu jádrový odhad hustoty pravděpodobnosti. Jako jádrovou funkci použiji opět Gaussovu funkci. Distribuci pak získám jednoduše jako sumu přes tyto Gaussovy funkce umístěné v bodech odpovídajících hodnotám interakcí:

$$\sigma_H(x) = \sum_{i=1}^{n} \exp\left(\frac{1}{2}\left(\frac{x - x_i}{H}\right)^2\right),$$

kde H udává opět šířku Gaussovy funkce. Jelikož nám jde pouze o relativní hodnoty, není potřeba distribuci normalizovat. Normalizace by byla navíc nežádoucí, protože by odstranila rozdíly v intenzitách mezi distribucemi pro hojně a řídce zastoupené typy interakcí. Získané distribuce lze pak libovolně navzorkovat. Hustotu vzorkování a hyperparametr H je možné nastavit dle daného problému. Já ve své implementaci nenastavuji přímo tyto dva parametry, místo hyperparametru H nastavuji jeho poměr k délce kroku vzorkování. Nemá totiž například význam nastavovat malou šířku Gaussových funkcí, pokud k tomu není zároveň nastavena velká hustota vzorkování, která by je byla schopna zachytit. Tímto způsobem je model relativně málo citlivý na volbu poměru a lze tak používat přednastavený poměr. Jak bylo řečeno dříve, výhodou distribucí je konstantní velikost výsledné reprezentace, která je daná pouze množstvím navzorkovaných bodů a množstvím typů interakcí, ale nikoliv počtem interakcí. Díky tomu lze při vhodném nastavení očekávat lepší výsledky těchto reprezentací při učení napříč molekulami.

4.1.4 Výčet orbitálních energií

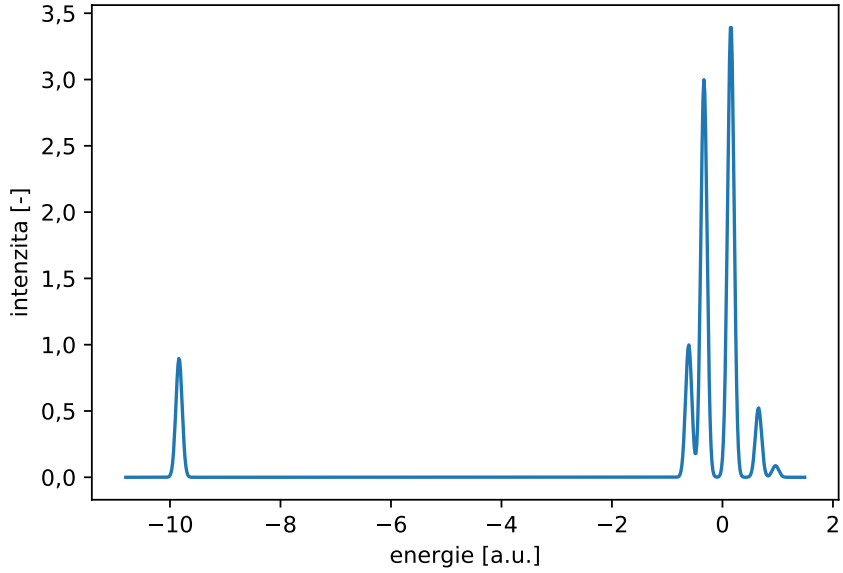
Doposud jsem se zabýval pouze molekulárními reprezentacemi, které kódují strukturu molekuly. Otázkou je však, zda není možné předpočítat si levně nějaké vlastnosti, které by vylepšili proces učení. Navrhuji proto molekulární deskriptory založené na energiích molekulových orbitalů pro učení excitačních energií. Molekulový orbital je jednoduše řečeno vlnová funkce pro jeden elektron. Vlnová funkce molekuly je pak sestrojena jako kombinace těchto molekulových orbitalů. Energie molekulových orbitalů jsou vlastně energetické

hladiny, kterých mohou elektrony v molekule nabývat. Říkají nám tedy, přechody o jakých energiích mohou v molekule nastat. Teoreticky se tak pro excitační energie jedná o dobrý deskriptor. Nicméně vše záleží na způsobu jejich výpočtu. My chceme použít nějakou rychlou kvantovou metodu, tak aby se nám strojové učení stále vyplatilo. Tyto deskriptory lze však systematicky vylepšovat zlepšováním kvantové metody, což může být výhoda. Pokud chceme model učit na datech, jejichž získání trvá v řádu týdnů, tak si můžeme dovolit strávit několik hodin výpočtem kvalitních deskriptorů.

Deskriptory by bylo teoreticky možné předpočítat externě v některém z dostupných kvantově-chemických programů a poté načíst. Já jsem nicméně využil program Psi4[42], který má rozhraní pro programovací jazyk Python a zabudoval jsem tyto výpočty nativně do třídy implementující tyto deskriptory. Tento program bohužel neobsahuje žádnou semiempirickou (neboli extrémně rychlou) metodu, pro výpočet orbitální energií proto používám DFT metodu s funkcionálem PBE, což je samo o sobě také relativně rychlá metoda.

Při kvantových výpočtech je hlavním vstupním parametrem výpočtu tzv. bázový set nebo také báze. Při řešení Schrödingerovy rovnice (viz sekce 2.2) se totiž z praktického hlediska hledá vlnová funkce jako kombinace sady jednodušších funkcí (typicky se jedná o Gaussovy funkce) a sadu těchto funkcí definuje právě bázový set. Ten nám tedy udává efektivní velikost systému neboli kolik parametrů optimalizujeme. Pro výpočet orbitálních energií jsem schválně použil velmi malý bázový set 3-21g, který je pro běžné výpočty zdaleka nedostačující, deskriptory chceme totiž získat v rozumném čase. S tímto nastavením trvá výpočet deskriptorů pro tisíc geometrií CH_2OO v řádu minut. Pro ostatní použité datové sady s tisíci až desetitisíci molekulami/geometriemi větších rozměrů trvá výpočet v řádu jednotek hodin (na výpočetním uzlu s 24 jádry). To je sice již relativně dlouhá doba, nicméně z pohledu kvantových výpočtů a množství přepočítávaných molekul se jedná o relativně bezvýznamnou hodnotu. Pokud chceme použít strojové učení, tak z toho důvodu, abychom se vyhnuli výpočtům trvajícím v řádu týdnů až měsíců. Z výpočtů získáme i velké množství orbitalů, jejichž energie je příliš vysoká a již se tak nemohou účastnit hledaných přechodů, porovnávám proto i deskriptory, kde tyto orbitaly nepoužiji. Výhodou těchto deskriptorů je také lineární velikost s počtem atomů v molekule oproti kvadratické závislosti pro matice a množiny interakcí.

Pro učení napříč molekulami opět zavádím deskriptory tvořené převzorkováním distribuce. Výchozí vzorec 4.1.3 je stejný, akorát v tomto případě získám pouze jednu distribuci. Jak bylo řečeno výše, řada nejvyšších orbitalů už nám nepřináší žádnou hodnotnou informaci, stejně tak nejníže položené orbitaly přináší jen málo informace. Zavádím proto vážící funkci , která upřednostňuje orbitaly s energií kolem nuly. Jedná se o Gaussovu funkci, u které volím zvlášť parametr rozšíření pro levou stranu a pravou stranu funkce. Zatímco na levé straně klesají váhy pomalu, na pravé straně pro vysoko umístěné orbitaly klesá funkce velmi rychle. Ukázka výsledné distribuce je pro molekulu methanu CH_4 zachycena na obrázku 4.1.



Obrázek 4.1: Znázornění distribuce založené na energiích molekulových orbitalů.

4.2 Strojové učení rozdílu

Δ ML popsaný v sekci 3.3 budu testovat i v rámci této práce. Místo přímého učení zkoumané veličiny se učíme rozdíl její hodnoty mezi výpočetně nenáročnou a přesnější kvantově-chemickou metodou. Fitovaná funkce by se díky tomu měla měnit pomaleji, být hladší a tedy i snadněji naučitelná. Z hlediska programu však nedochází k žádné změně, akorát se učíme jinou veličinu. Při testování budu sledovat do jaké míry použití Δ ML pomáhá k vylepšení jednotlivých deskriptorů v závislosti na jejich kvalitě. Lze totiž očekávat, že s rostoucí přesností deskriptorů pro odhad dané veličiny přímo bude zároveň klesat přínos Δ ML přístupu, jelikož se začnou projevovat chyby způsobené nedostatečnými výsledky levné kvantové metody.

4.3 Vstupní formát

V rámci této práce je potřeba ukládat, načítat a obecně pracovat s geometriemi molekul a k nim přiřazenými vlastnostmi. Dodržuji zde zavedený konsensus z oboru a data ukládám v podobě rozšířeného `xyz` formátu. Základní `xyz` formát má obecně pro jednu molekulu na první řádce počet atomů v molekule n, na druhé řádce libovolný komentář a dalších n řádků slouží k zápisů kartézských souřadnic jednotlivých atomů. Molekuly či geometrie následují v základním `xyz` formátu rovnou jedna za druhou. Rozšířený `xyz` formát umožňuje přidávat další informace jak k jednotlivým atomům, tak pod výčet kartézských souřadnic, a jednotlivé molekuly/geometrie jsou v něm oddělené prázdným řádkem. Vyvinutý program umí načítat obě dvě verze

formátu, přičemž vlastnosti jsou čteny z řádku pro komentář, což je ve shodě s dostupnými datovými sadami. Následuje ukázka pro geometrie molekuly CH_2OO s excitační energií a tranzitním dipólovým momentem pro jeden excitovaný stav.

```
5
3.38477 1.15766
C     0.53831745E+00    0.11185424E+00    0.69760773E+00
O    -0.59651587E+00    0.64543105E+00    0.10701378E+01
O    -0.17297993E+01   -0.73653329E-01    0.65666884E+00
H     0.62430793E+00   -0.81515201E+00    0.11294458E-01
H     0.13258040E+01    0.91523741E+00    0.79823071E+00

5
4.04830 1.26916
C     0.59641884E+00    0.98347894E-01    0.69246112E+00
O    -0.55873726E+00    0.62174034E+00    0.90125395E+00
O    -0.17236143E+01   -0.54921874E-01    0.72041017E+00
H     0.57428955E+00   -0.10769552E+01    0.34813458E+00
H     0.13687666E+01    0.68436139E+00    0.12091788E+01

...
```

Kapitola 5

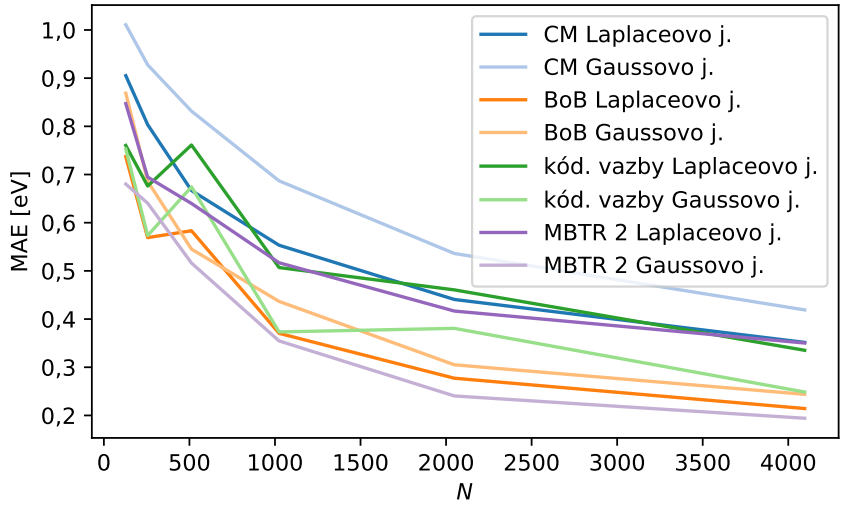
Výsledky a testování

V této sekci nejprve probírám výsledky pro učení na sadě molekul. Jelikož spektra modelovaná empirickým rozšířením jsou daná pouze jedním výpočtem excitačních energií a intenzit absorpce, není nutné vyhodnocovat přesnost výsledných spekter zvlášť. V druhé části pak ukazuji výsledky z učení na sadě geometrií pro dva popsané modelové systémy včetně porovnání výsledných spekter modelovaných reflexním principem.

5.1 Datová sada QM7b

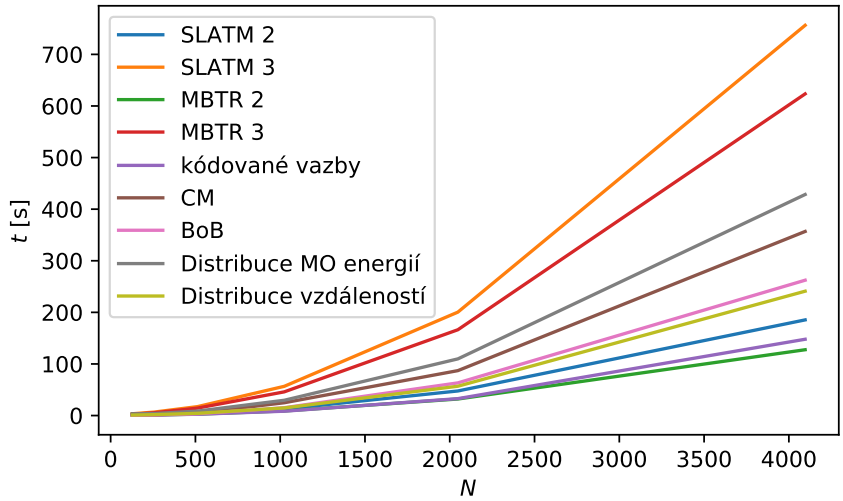
Datovou sadu QM7b jsem použil jako první pro testování predikce excitačních energií prvního excitovaného stavu. Nejprve jsem testoval vliv použité jádrové funkce v rámci KRR na střední absolutní chybu predikcí. Ukázalo se, že v případě reprezentací založených na distribucích je striktně výhodnější Gaussova funkce. Naproti tomu reprezentace ve formě matic nebo množin interakcí vykazují vždy lepší výsledky s Laplaceovou funkcí. Ukázka rozdílu pro CM, BoB a dvě reprezentace založené na distribucích je zachycena na obrázku 5.1. Tento výsledek je obecného charakteru a platí i pro sadu QM8 a obě dvě zkoumané veličiny.

Dále jsem testoval, jakým způsobem škáluje výpočetní náročnost učení modelu s velikostí trénovací množiny. Trénování bylo prováděno za použití výpočetního uzlu s 24 jádry s použitím předpočítávání reprezentací. Učení každé reprezentace bylo prováděno s použitím optimální jádrové funkce. Výsledky jsou shrnuty v obrázku 5.2. Trend je pro všechny reprezentace stejný, a sice exponenciální růst s počtem vzorků. Z grafu vyplývá, že dle očekávání jsou nejnáročnější reprezentace zahrnující i tříatomové interakce (SLATM 3, MBTR 3), takových interakcí je totiž v molekulách velké množství a výrazně zvětšují velikost reprezentace. Pohybujeme se však stále pouze kolem deseti minut pro více než čtyři tisíce trénovacích dat. Na grafu je zachyceno trénování modelu pouze s optimalizací hyperparametrů učení (na mřížce 10x10). Řada reprezentací má ještě k tomu svoje vnitřní parametry, jejichž optimalizace už je o poznání složitější, jelikož si nelze předem reprezentaci pouze jednou



Obrázek 5.1: Porovnání střední absolutní chyby Laplaceova a Gaussova jádra pro různé reprezentace na datové sadě QM7b.

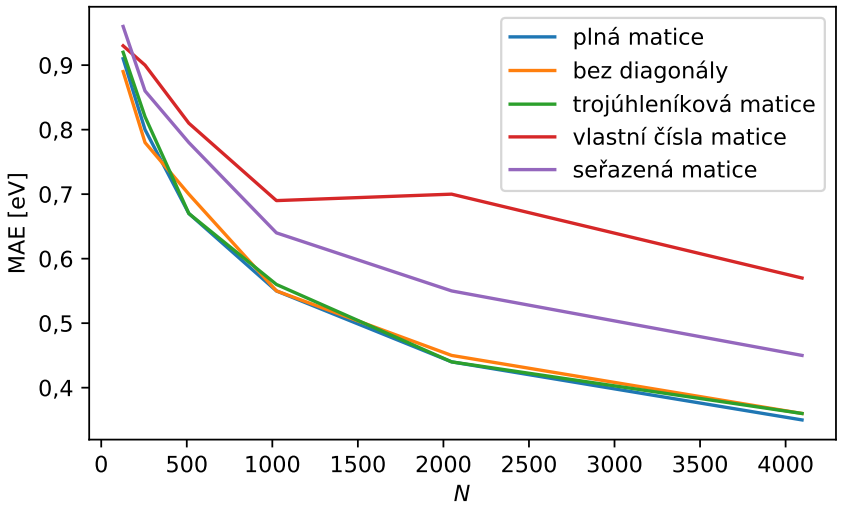
předpočítat a také exponenciálně roste počet kombinací parametrů, které je třeba testovat.



Obrázek 5.2: Porovnání učících časů pro různé reprezentace na datové sadě QM7b.

Dále jsem zkoumal, jaký vliv mají na výsledky různé modifikace CM reprezentace navržené pro její vylepšení v některých situacích (viz sekce 3.1) a jaký vliv má použití diagonálních prvků neboli jednoatomové příspěvky. CM bývá také někdy implementována jako čtvercová matice, zatímco někdy jako horní trojúhelníková matice. To může mít částečně vliv na výsledky, i když je matice symetrická, jelikož je tím změněn poměr příspěvků diagonálních prvků. Všechny tyto modifikace jsou porovnány na obrázku 5.3. Je zřejmé,

že diagonální prvky mají zanedbatelný vliv a tedy i rozdíl mezi plnou a trojúhelníkovou maticí je mizivý. Naproti tomu je vidět veliké zhoršení při použití řazení matice podle norem řádků i při použití vlastních čísel místo samotné matice. Dále používám plnou matici, jelikož má nepatrně lepší výsledky, nicméně na zanedbatelné úrovni.



Obrázek 5.3: Porovnání střední absolutní chyby pro různé verze CM reprezentace na datové sadě QM7b.

V sekci 4.1.1 jsem nastínil problematiku mocniny vzdáleností v CM reprezentaci. Tu jsem také zkoumal podrobněji, akorát na reprezentaci BoB, která z CM reprezentace vychází a má předpoklady pro lepší výsledky při učení na množině molekul. Pro BoB reprezentaci jsem provedl plnou optimalizaci hyperparametrů učení a mocniny zároveň na mřížce 10x10x8. Ukázalo se, že optimální jsou v tomto případě vyšší záporné mocniny (-4 až -6), mezi kterými už je poté malý rozdíl. Porovnání výsledků pro běžnou zápornou první mocninu, zápornou pátou mocninu a optimální mocniny je v tabulce 5.1.

množství trénovacích dat	128	256	512	1024	2048	4096
BoB (-1)	0,74	0,57	0,58	0,37	0,28	0,21
BoB opt. (-6,-5,-4,-5,-4,-5)	0,51	0,40	0,38	0,22	0,19	0,13
BoB (-5)	0,50	0,40	0,43	0,22	0,20	0,13

Tabulka 5.1: Porovnání středních absolutních chyb pro různé mocniny vzdáleností v BoB reprezentaci na datové sadě QM7b.

Výsledek je velmi překvapivý, pro pátou negativní mocninu dostáváme střední absolutní chybu 0,13 eV pro 4096 trénovacích molekul oproti 0,21 eV u běžné verze BoB. Výsledek s první zápornou mocninou je ve shodě s literaturou, která uvádí hodnoty velmi těsně kolem 0,2 eV pro 5000 učících

molekul (na stejné učící sadě).[29, 14] I v kontextu doposud nejlepších výsledků uváděných v literatuře (viz sekce 3.3) je toto extrémně dobrá hodnota posouvající BoB reprezentaci do čela pelotonu (i přes menší množství učících dat), přitom se jedná o jednu ze základních reprezentací. Mocniny vzdáleností jsou tedy nadále u ostatních odvozovaných reprezentací také optimalizovány, i když z výpočetních důvodů již ne zároveň s parametry učení. BoB reprezentaci s touto mocninou jsem vyzkoušel i pro predikci atomizační energie, která je většinou základní vlastností, na kterou jsou reprezentace testovány. Obdržel jsem výsledek 1,63 kcal/mol pro 4096 geometrií oproti hodnotě z literatury 1,84 kcal/mol pro 5000 geometrií.[29] Je tedy zřejmé, že použitá mocnina nemá vliv pouze na predikci excitačních vlastností a měla by jí být v celé komunitě věnována větší pozornost.

Do testování jsem samozřejmě zařadil i reprezentace založené na distribucích v rámci navrženého schématu obecného odvození reprezentací z matice interakcí přes množiny interakcí až po distribuce interakcí. Při konstrukci distribuce vlastně přicházíme o samotné hodnoty interakcí, jelikož ty poté představují hodnoty na ose x, kterou převzorkujeme. To, že například v CM resp. BoB reprezentaci používáme náboje jader jako součást deskriptorů, tak ve výsledků znamená pouhé přeškálování osy x distribuce, jelikož pro danou distribuci je tento člen konstantní (máme jednu distribuci pro každý typ interakce v molekule). Nelze tedy nejspíše ani očekává přílišné rozdíly ve výsledcích. Testoval jsem distribuci vycházející z Coulombických interakcí (tedy z CM, resp. BoB reprezentace), dále distribuci vycházející pouze z mocnin vzdáleností a distribuci vycházející z matice elektronegativit (modifikace CM, kde jsou elektronegativity použity místo nábojů jader). Výsledky jsou sepsány v tabulce 5.2. Dle předpokladu jsou výsledky navzájem opravdu velmi podobné

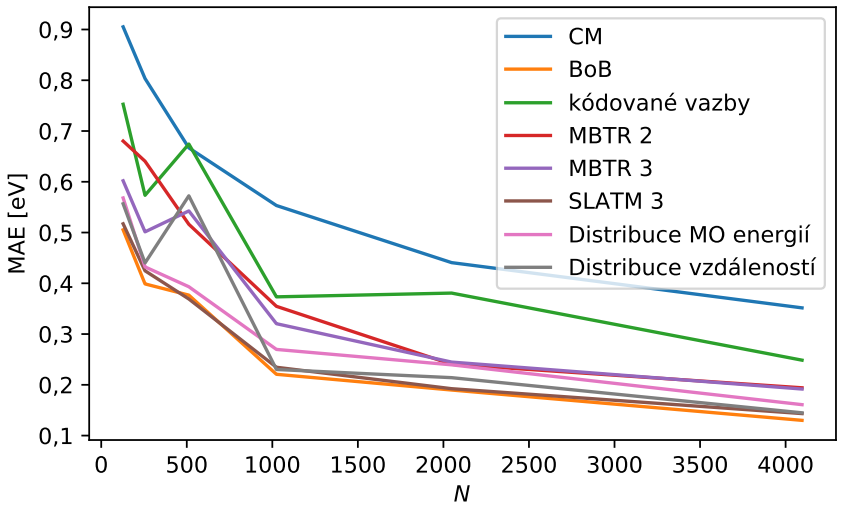
množství trénovacích dat	128	256	512	1024	2048	4096
Distribuce vzdáleností	0,56	0,44	0,57	0,23	0,21	0,14
Distribuce Coulomb. interakcí	0,58	0,47	0,66	0,29	0,26	0,16
Distribuce elektronegativit	0,56	0,43	0,59	0,25	0,21	0,14

Tabulka 5.2: Porovnání středních absolutních chyb pro vyvíjené reprezentace založené na distribucích na datové sadě QM7b.

Finální výsledky jsem sestrojil z vybraných verzí výše uvedených reprezentací, reprezentace založené na orbitálních energiích a referenčních reprezentací z literatury. Tyto výsledky jsou shrnuty na obrázku 5.4. Prvenství zjevně obhájila BoB reprezentace (s pátou zápornou mocninou). V těsném závěsu se nachází všechny mnou dovozené reprezentace založené na distribucích a také SLATM reprezentace. Mírně pozadu zaostávají obě verze MBTR reprezentace (s a bez zahrnutí tříatomových interakcí). Dle očekávání dopadla nejhůře obyčejná CM matice, která je zde víceméně pro srovnání.

Nečekaně špatně dopadly ale také kódované vazby. To je zajímavé, jelikož

tato reprezentace má velmi podobnou konstrukci jako mnou odvozené distribuce vzdáleností a její parametry byly optimalizovány autory právě na této učící sadě (i když pro predikci celé sady vlastností).[14] Její nejdůležitější parametr jsem navíc ještě optimalizoval pro excitační energie. Rozdíl může být v tom, že já jsem jako vhodnější zvolil záporné mocniny vzdáleností, jelikož v tu chvíli mi veškeré interakce spadnou do intervalu (0, 1) a nemusím tedy hledat odkud a kam budu distribuci vzorkovat. Druhým rozdílem může být definice parametrů, kde jsem zvolil možná vhodnější a stabilnější parametry (viz sekce 4.1.3).



Obrázek 5.4: Porovnání střední absolutní chyby pro různé reprezentace na datové sadě QM7b.

Tyto výsledky jsou celkově dobré, jelikož se dostávají na úroveň doposud nejlepších publikovaných hodnot a to s použitím menšího množství dat. Je také zajímavé si všimnout, že z referenčních reprezentací se umístila na předních příčkách pouze SLATM reprezentace. To je nejspíše způsobeno tím, že používá Londonův potenciál pro dvouatomové interakce, přičemž ten je daný šestou negativní mocninou vzdáleností, to nápadně připomíná výsledky optimalizace BoB reprezentace. A skutečně i u mnou odvozených reprezentací založených na distribucích jsem jako nejlepší vyhodnotil vyšší záporné mocniny (-4 až -5).

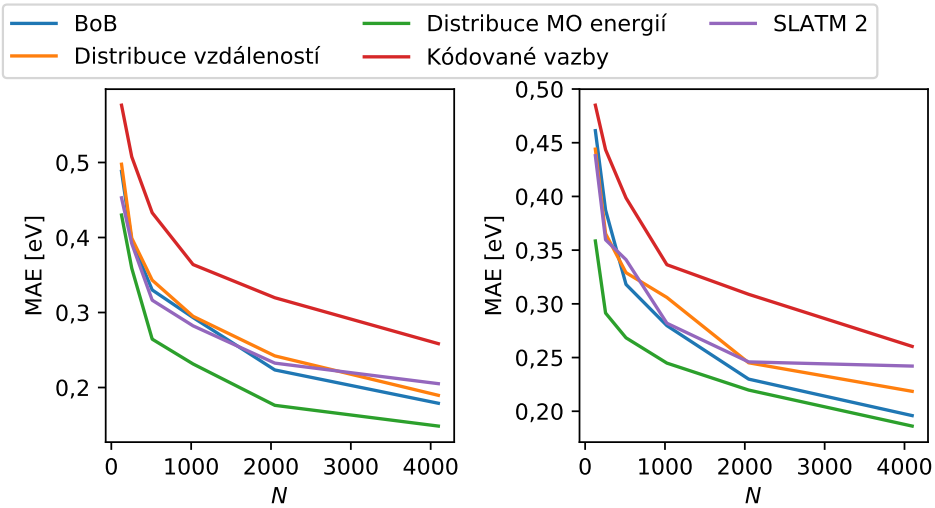
Další věc, která stojí za poznamenání je prakticky nulový rozdíl mezi MBTR 2 a MBTR 3 reprezentacemi neboli vliv zahrnutí tříatomových interakcí. Z těchto výsledků to vypadá, že veškeré informace jsou zahrnuty již ve dvouatomových příspěvcích. Pro ověření jsem modifikoval SLATM reprezentaci, která nativně zahrnuje tříatomové interakce. Porovnání verzí s a bez tříatomových interakcí je pro MBTR i SLATM reprezentaci v tabulce 5.3. Tyto výsledky skutečně potvrzují malý vliv zahrnutí tříatomových příspěvků. Je ale možné, že tento rozdíl by se lehce zvýšil dokonalejší optimalizací parametrů těchto reprezentací, které jich mají relativně velké množství.

množství trénovacích dat	128	256	512	1024	2048	4096
MBTR 2	0,68	0,64	0,52	0,35	0,24	0,19
MBTR 3	0,60	0,50	0,54	0,32	0,24	0,19
SLATM 2	0,53	0,43	0,36	0,24	0,19	0,15
SLATM 3	0,52	0,42	0,37	0,23	0,19	0,14

Tabulka 5.3: Porovnání středních absolutních chyb excitačních energií s a bez zahrnutí tříatomových interakcí na datové sadě QM7b.

5.2 Datová sada QM8

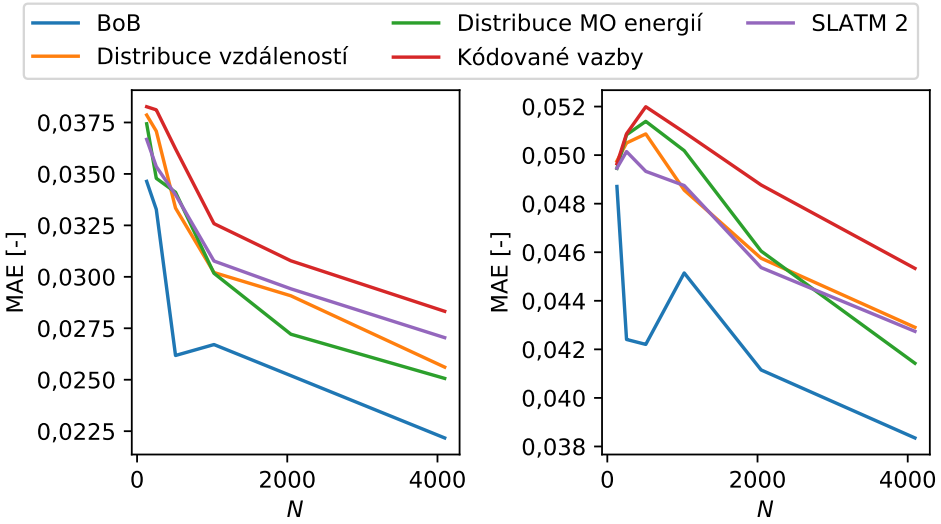
Datová sada QM8 obsahuje excitační energie a oscilátorové síly pro první dva excitované stavy více než 20 000 molekul. Protože kompletní analýza predikcí byla provedena v předchozí sekci pro sadu QM7b, uvádím zde již pouze nejzajímavější výsledky. Porovnání střední absolutní chyby pro první dvě excitační energie je zachyceno na obrázku 5.5. Ukazuje se, že distribuce orbitálních energií překonává ostatní reprezentace pro obě excitační energie. Naproti tomu kódované vazby jsou mezi distribucemi opět nejhorší. Pro BoB reprezentaci byla nalezena optimální mocnina vzdáleností -6, díky které opět drží krok s reprezentacemi založenými na distribucích.



Obrázek 5.5: Porovnání střední absolutní chyby 1. excitační energie (vlevo) a 2. excitační energie (vpravo) pro datovou sadu QM8.

Porovnání reprezentací pro první dvě oscilátorové síly je znázorněno na obrázku 5.6. BoB reprezentace s mocninou -6 systematicky překonává všechny ostatní reprezentace pro obě dvě oscilátorové síly. Nabízí se již pravidlo vyšších záporných mocnin u BoB generalizovat. Výsledky naznačují, že BoB podává při učení na sadě molekul výrazně lepší výsledky pro excitační vlastnosti při

použití záporných mocnin vzdáleností -4 až -6. Ostatní reprezentace jsou na tom navzájem podobně, akorát kódované vazby jsou opět nejhorší.



Obrázek 5.6: Porovnání střední absolutní chyby oscilátorové síly 1. (vlevo) a 2. (vpravo) excitovaného stavu pro datovou sadu QM8.

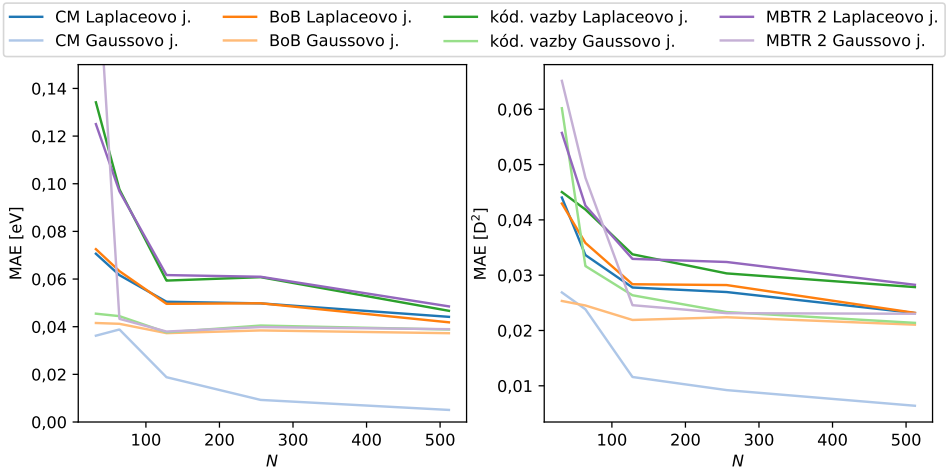
Dále jsem v tabulce 5.4 shrnul výsledky pro distribuci orbitálních energií s a bez použití ΔML. Ukázalo se totiž, že při použití tohoto přístupu předskočí i BoB reprezentaci. Z této tabulky vyplývá několik velmi pozitivních faktů. Zaprvé ΔML přístup zde systematicky výrazně vylepšuje predikci všech excitačních vlastností. A zadruhé jsou tyto výsledky výrazně lepší než výsledky publikované v literatuře pro tuto sadu, navíc s použitím menšího množství učících dat.[48]

množství trénovacích dat	128	256	512	1024	2048	4096
1. exc. energie [eV]	0,430	0,359	0,264	0,231	0,176	0,149
1. exc. energie ΔML [eV]	0,147	0,128	0,109	0,098	0,083	0,073
2. exc. energie [eV]	0,359	0,291	0,268	0,245	0,220	0,186
2. exc. energie ΔML [eV]	0,208	0,168	0,159	0,147	0,135	0,121
1. osc. síla [-]	0,037	0,035	0,034	0,030	0,027	0,025
1. osc. síla ΔML [-]	0,011	0,012	0,012	0,012	0,012	0,012
2. osc. síla [-]	0,049	0,051	0,051	0,050	0,046	0,041
2. osc. síla ΔML [-]	0,024	0,025	0,025	0,030	0,027	0,026

Tabulka 5.4: Porovnání středních absolutních chyb pro distribuci orbitálních energií s a bez použití ΔML na datové sadě QM8.

5.3 Molekula CH_2OO

Pro učení napříč geometriemi jedné molekuly jsem opět na začátku porovnával vliv použité jádrové funkce. Výsledek pro vybrané molekulové reprezentace je zachycen na obrázku 5.7 zvlášť pro excitační energie a tranzitní dipólové momenty. Ukázalo se, že v tomto případě je Gaussova funkce striktně lepší pro obě zkoumané veličiny. Všímavý čtenář si z obrázku také všimne, že CM reprezentace řádově překonává reprezentace založené na distribucích a množinách interakcí, jak bylo předpovězeno v sekci 4.1.1. Vzhledem k obrovskému rozdílu nejsou tyto reprezentace v této aktuální již dále uvažovány.



Obrázek 5.7: Porovnání střední absolutní chyby Laplaceova a Gaussova jádra pro excitační energie (vlevo) a tranzitní dipólové momenty (vpravo) pro molekulu CH_2OO.

Stejně jako v případě učení napříč molekulami zkoumám vliv mocniny vzdáleností v reprezentacích na výsledky. Tentokrát však na maticích interakcí, jelikož u těch lze předpokládat lepší výsledky pro tento typ problému, jak bylo popsáno v sekci 4.1.1. Porovnání středních absolutních chyb excitačních energií pro CM a matici elektronegativit se nachází v tabulce 5.5. Celkem zajímavě se ukazuje, že lépe než klasická záporná mocnina -1 vychází kladná mocnina +1. Tento výsledek byl uvažován pro tranzitní dipólové momenty, u energií je však relativně překvapivý, jelikož potenciální energie běžně klesá se vzdáleností. Nicméně zde se jedná o excitační energii, na kterou mohou mít vliv odlišné skutečnosti. U matice elektronegativit není rozdíl tak velký, pro CM je však již významný. Obě matice přitom při použití kladné mocniny vzdáleností vykazují velmi podobné výsledky. V téže tabulce je také porovnání pro reprezentaci využívající orbitální energie, kde srovnávám vliv redukce horních orbitalů. V tomto případě se zlepší učení pro malé množství dat, ale lehce zhorší výsledky pro větší data. To koresponduje s tím, že se sice snadněji učíme jednodušší model, ale zahazujeme část informace.

Z tabulky 5.5 také vidíme, že přesnost všech tří reprezentaci je velmi dobrá,

s několika sty vzorky se dostáváme na úroveň jednotek meV. Již pro 64 trénovacích molekul pak operujeme s přesností zhruba 10x lepší než u učení napříč molekulami se 4096 vzorky. Taková přesnost je více než dostačující pro cílovou aplikaci, tedy modelování spekter.

množství trénovacích dat	32	64	128	256	512
CM (-1)	0,0362	0,0388	0,0188	0,0093	0,0051
CM (+1)	0,0410	0,0238	0,0153	0,0066	0,0038
Matice elektroneg. (-1)	0,0394	0,0326	0,0166	0,0070	0,0037
Matice elektroneg. (+1)	0,0516	0,0265	0,0157	0,0051	0,0034
Orbitální energie	0,0512	0,0160	0,0097	0,0051	0,0032
Orbitální energie (red.)	0,0227	0,0155	0,0098	0,0059	0,0040

Tabulka 5.5: Porovnání středních absolutních chyb excitačních energií pro různá nastavení reprezentací pro molekulu CH_2OO.

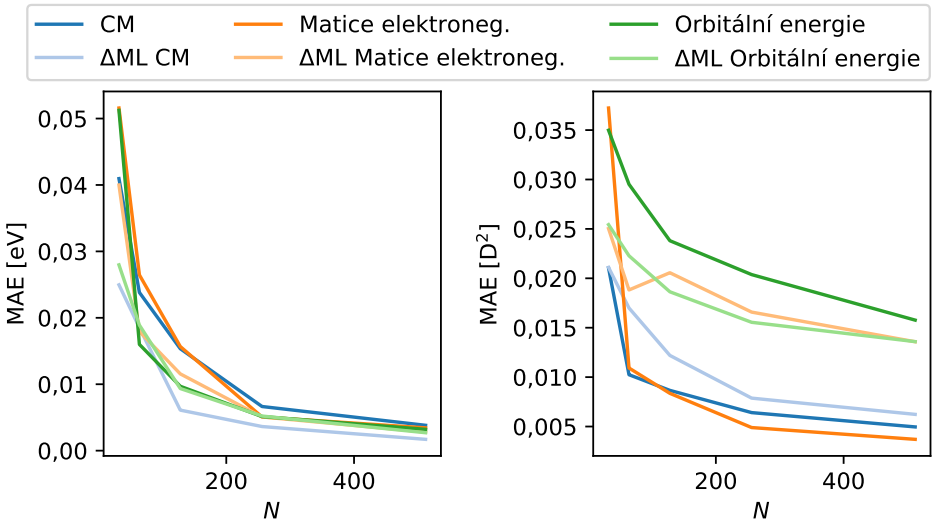
Stejné porovnání jako pro excitační energie jsem provedl i pro tranzitní dipólové momenty. Tyto výsledky jsou shrnuty v tabulce 5.6. Zde je zaprvé ještě více vidět pozitivní vliv kladné mocniny vzdáleností u CM a za druhé si můžeme všimnout, že pro dostatečné množství dat dává lepší výsledky matice elektronegativit. Vliv redukce orbitalů na reprezentaci orbitálních energií je víceméně stejný. Pro malá data dochází ke zlepšení, pro velká data k mírnému zhoršení. I pro tranzitní dipólové momenty vidíme extrémně přesné predikce. Pro srovnání, střední hodnota excitační energie je pro tato data 4,00 eV a střední hodnota čtverce tranzitního dipólového momentu 1,22 D^2.

množství trénovacích dat	32	64	128	256	512
CM (-1)	0,0269	0,0238	0,0116	0,0092	0,0064
CM (+1)	0,0211	0,0102	0,0087	0,0064	0,0050
Matice elektroneg. (-1)	0,0295	0,0216	0,0097	0,0053	0,0039
Matice elektroneg. (+1)	0,0372	0,0109	0,0084	0,0049	0,0037
Orbitální energie	0,0263	0,0132	0,0087	0,0072	0,0040
Orbitální energie (red.)	0,0156	0,0123	0,0095	0,0060	0,0051

Tabulka 5.6: Porovnání středních absolutních chyb tranzitních dipólových momentů pro různá nastavení reprezentací pro molekulu CH_2OO.

Dále jsem vyhodnotil přínos ΔML přístupu jak pro excitační energie, tak pro tranzitní dipólové momenty. Výsledky zachycuje obrázek 5.8. U excitačních energií vidíme systematické zlepšení pro CM a do určitého počtu vzorků i u matice elektronegativit. U orbitálních energií vidíme zlepšení pouze pro 32 vzorků. U tranzitního dipólového momentu naopak vidíme zhoršení u obou matic a jen lehké zlepšení u orbitálních energií. Výsledky celkově naznačují, že ΔML je pro excitační energie přínosný u každé reprezentace jen do určité přesnosti. Výrazně horší výsledky v případě tranzitních dipólových momentů

jsou nejspíše způsobeny porušením některého z předpokladů. Buďto je korelace mezi metodami nedostatečná a nebo není rozdíl tranzitních dipólových momentů ze dvou kvantových metod dostatečně hladká funkce.

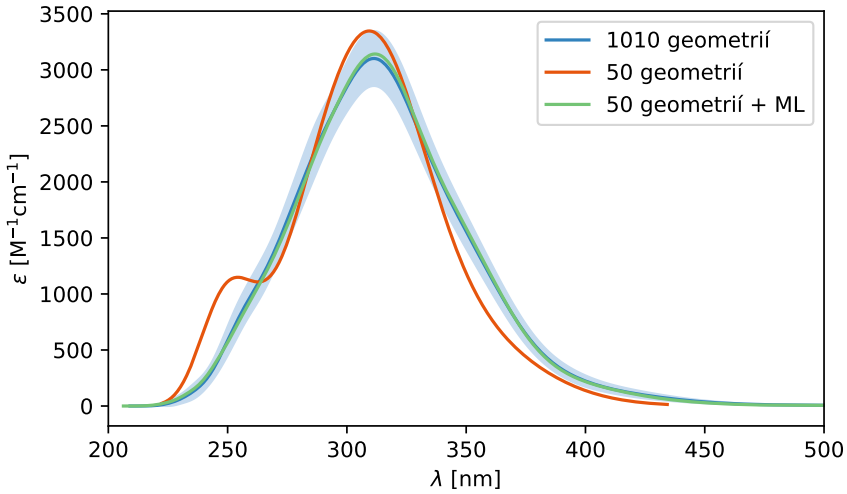


Obrázek 5.8: Porovnání průměrné střední absolutní chyby nejlepších reprezentací s a bez použití ΔML pro molekulu CH_2OO.

V tuto chvíli již mohu přistoupit k modelování spektra. Vzhledem k přesnosti predikcí by mělo být možné použít k učení pouze 50 geometrií z 1010 při zachování přesnosti. Excitační energie a tranzitní dipólové momenty pro zbytek geometrií pak predikuji pomocí naučeného modelu a společně všechna data použiji k modelování spektra (50 trénovacích geometrií + 960 predikovaných). Pro predikci excitačních energií jsem v tomto případě použil orbitální energie, které dávají nejlepší výsledky bez použití ΔML přístupu. Tranzitní dipólové momenty byly dle výsledků pro malé množství učících dat predikovány pomocí CM s mocninným koeficientem +1.

Výsledné spektrum je znázorněno na obrázku 5.9 společně s plným spektrem modelovaným ze všech 1010 geometrií, ke kterému se chceme co nejvíce přiblížit. Pro srovnání je v obrázku také znázorněno spektrum získané pouze z 50 geometrií, na kterých byl model učen, bez použití ML. Je patrné, že spektrum z predikovaných dat je prakticky identické se spektrem, kdy byly použity všechny geometrie. Stejně tak je patrné, že ML model má velikou přidanou hodnotu, jelikož pouhé vybrání 50 geometrií vede k velmi špatným výsledkům.

Na obrázku jsou také podbarvením kolem plného spektra znázorněny 95% intervaly spolehlivosti získané *bootstrap* metodou. Vidíme, že i pro tvarem velmi jednoduché spektrum a více než tisíc geometrií jsou tyto intervaly nezanedbatelné. Z toho vyplývá, že pro konvergované spektrum je nutné použít opravdu veliké množství geometrií, v čemž nám může pomáhat právě strojové učení. Míru odlišnosti predikovaného spektra s referenčním spektrem jsem kvantifikoval pomocí KL divergence, přičemž jsem obdržel hodnotu



Obrázek 5.9: Srovnání spektra s a bez použití ML pro molekulu CH_2OO.

0,00063, která značí extrémně dobrou shodu. Pro srovnání, KL divergence znázorněného spektra modelovaného z 50 geometrií s plným spektrem je 0,042.

5.4 Molekula azobenzenu

Po úspěšném nasazení ML pro modelování spektra molekuly CH_2OO přejděme k velmi složitému případu molekuly azobenzenu. Pro tuto molekulu vyhodnocuji 7 excitovaných stavů, u kterých se navíc mění pořadí (viz sekce 3.2). Jedná o jakýsi zatěžkávací test.

Nejprve si opět srovnejme vliv mocniny vzdáleností v relevantních reprezentacích a redukce orbitalů u orbitálních energií. Pro excitační energie jsou data shrnuta v tabulce 5.7. Stejně jako v případě molekuly CH_2OO hrají prim mocniny -1 a +1. Protože druhý excitovaný stav má suverénně nejvyšší průměr tranzitních dipólových momentů, hraje největší roli i ve výsledném spektru a tudíž je mu věnována speciální pozornost. Z tabulky je patrné, že kladná i záporná mocnina poskytují v tomto případě vesměs stejné výsledky. Dále vidíme, že redukce u orbitálních energií zde má relativně veliký význam, hlavně pro menší množství dat. Bohužel z tabulky také vyplývá, že se pro excitační energie tentokrát pohybujeme v o řád menší přesnosti než pro předchozí molekulu. Stále se však jedná o rozumné hodnoty.

Tabulka 5.8 poskytuje stejné srovnání pro tranzitní dipólové momenty. Do porovnání byly tentokrát zahrnuty i BoB reprezentace a matice elektronegativit, jelikož poskytují v tomto případě podobné výsledky. Ukazuje se, že ani zde nemá znaménko mocniny valný vliv. Oproti excitačním energií nedochází ani ke zlepšení výsledků z reprezentace orbitálních energií při použití redukce orbitalů. Všechny výsledky vycházejí přibližně stejně a to velmi nelibě v

Průměr excitovaných stavů

množství trénovacích dat	32	64	128	256	512	1024
CM (-1)	0,128	0,108	0,094	0,071	0,061	0,054
CM (+1)	0,153	0,124	0,092	0,072	0,064	0,058
Orbitální energie	0,123	0,096	0,078	0,053	0,047	0,041
Orbitální energie (red.)	0,083	0,063	0,054	0,047	0,043	0,041

2. excitovaný stav

množství trénovacích dat	32	64	128	256	512	1024
CM (-1)	0,174	0,133	0,145	0,070	0,051	0,043
CM (+1)	0,248	0,152	0,097	0,066	0,053	0,043
Orbitální energie	0,159	0,105	0,079	0,026	0,022	0,020
Orbitální energie (red.)	0,071	0,038	0,031	0,023	0,022	0,021

Tabulka 5.7: Porovnání středních absolutních chyb excitačních energií pro různá nastavení reprezentací pro molekulu azobenzenu.

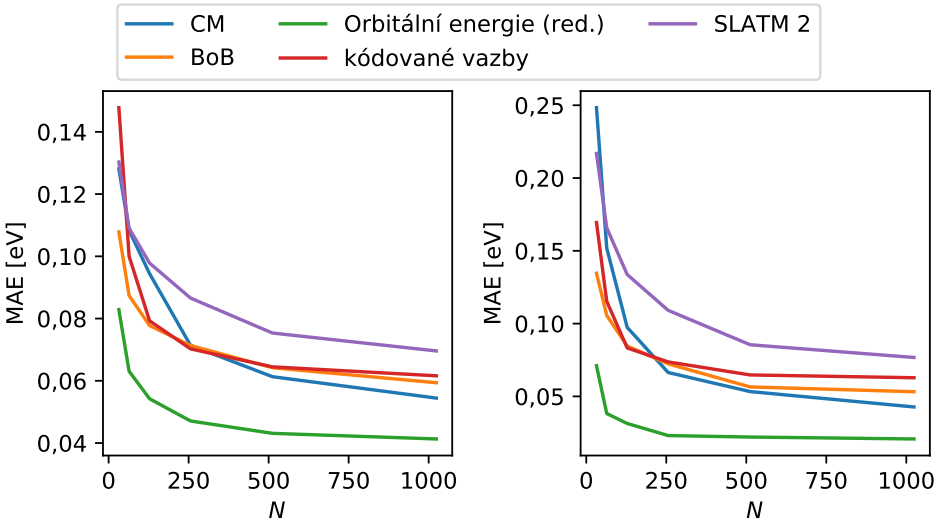
množství trénovacích dat	32	64	128	256	512	1024
CM (-1)	0,46	0,42	0,41	0,36	0,35	0,32
CM (+1)	0,46	0,40	0,39	0,34	0,33	0,30
Matice elektroneg. (-1)	0,47	0,44	0,43	0,37	0,36	0,33
Matice elektroneg. (+1)	0,45	0,42	0,41	0,37	0,35	0,33
BoB (-1)	0,45	0,38	0,36	0,33	0,32	0,31
BoB (+1)	0,41	0,35	0,34	0,33	0,31	0,30
Orbitální energie	0,41	0,38	0,36	0,33	0,32	0,30
Orbitální energie (red.)	0,43	0,35	0,34	0,32	0,31	0,30

Tabulka 5.8: Porovnání středních absolutních chyb průměru tranzitních dipólových momentů pro různá nastavení reprezentací pro molekulu azobenzenu.

desetinách D^2. Je nutné však zmínit, že tentokrát je průměr středních hodnot 1,49 D^2, přičemž střední hodnota nejintenzivnějšího 2. excitovaného stavu je 7,56 D^2.

Grafické srovnání predikcí excitačních energií různých reprezentací je zachyceno na obrázku 5.10. Pro porovnání jsou zde zařazeny i reprezentace založené na distribucích a BoB. Je zjevné, že jak celkově, tak pro 2. excitovaný stav suverénně vítězí reprezentace orbitálních energií v redukované verzi.

Na dalším obrázku 5.11 je pak obdobné srovnání pro tranzitní dipólové momenty. Zde je nutné zdůraznit velmi úzkou škálu grafu, která je nastavena pro lepší rozlišení reprezentací. Ve skutečnosti však všechny reprezentace vycházejí velmi podobně, jak je patrné i z výše uvedené tabulky 5.8. Nejlépe opět vychází reprezentace orbitálních energií, nicméně pouze s malým rozdílem. Například BoB reprezentace poskytuje prakticky identické výsledky a stejně

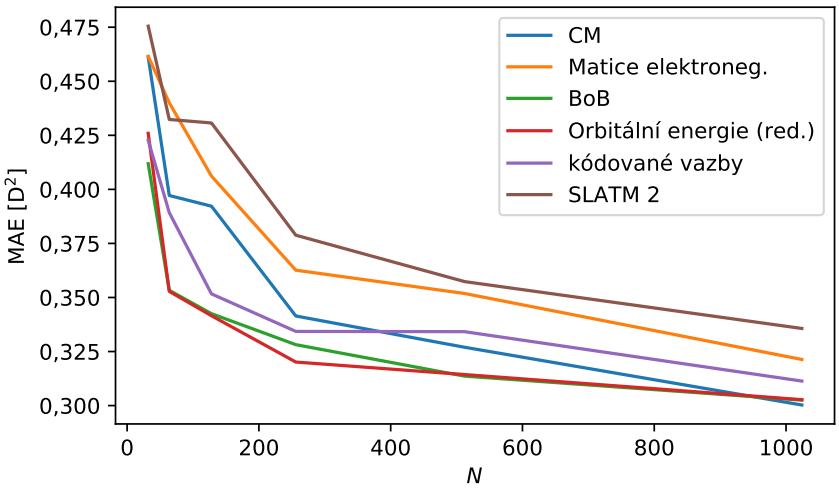


Obrázek 5.10: Porovnání průměrné střední absolutní chyby excitačních energií (vlevo) a střední absolutní chyby energií 2. excitovaného stavu (vpravo) pro molekulu azobenzenu.

tak i CM reprezentace pro větší množství bodů.

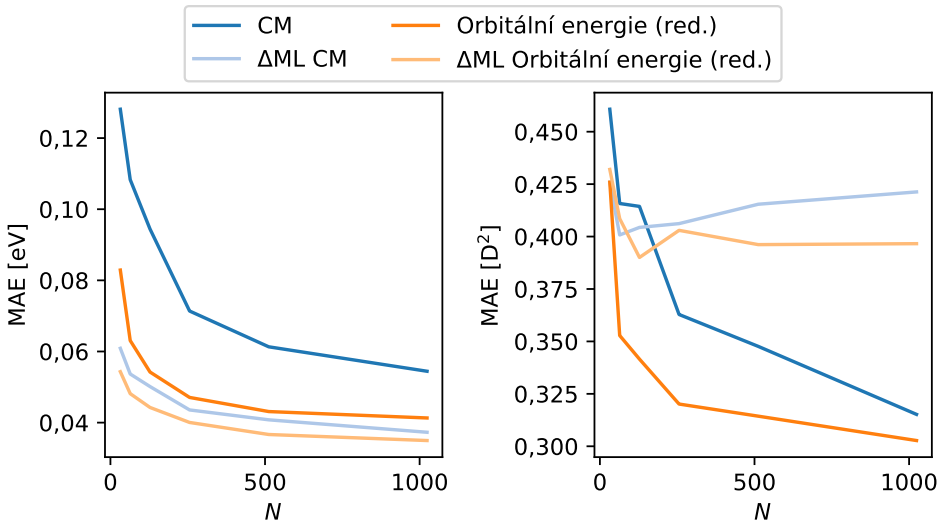
Dále se podíváme na přínos ΔML přístupu na dvou nejslibnějších reprezentacích (viz obrázek 5.12). Vidíme, že pro excitační energie zlepšuje tento přístup systematicky výsledky. Pro CM však výrazně více než pro orbitální energie, tento výsledek je konzistentní s předchozí molekulou. Orbitální energie nicméně stále zůstávají reprezentací s nejlepšími výsledky. Pro tranzitní dipólové momenty dochází s použitím ΔML k výraznému zhoršení. Je tak ještě prohloubena nefunkčnost tohoto přístupu z molekuly CH_2OO. Ještě větší chyba je nejspíše způsobena již popsaným jiným pořadím excitovaných stavů a to nejen mezi geometriemi, ale i pro jednu geometrii v rámci obou kvantově-chemických metod. BoB reprezentace v grafu není pro přehlednost znázorněna (i když je relativně slibná pro tranzitní dipólové momenty), jelikož pro energie poskytuje výrazně horší výsledky a pro tranzitní dipólové momenty se chová stejně jako ostatní metody.

Zhodnoťme nyní, jak moc veliké chyby se tedy dopustíme, pokud nejlepší z výše uvedených modelů použijeme pro predikci hodnot sloužících k modelování spektra reflexním principem. Při použití redukované reprezentace orbitálních energií s ΔML pro predikci excitačních energií a stejné reprezentace bez ΔML pro predikci tranzitních dipólových momentů obdržíme spektrum znázorněné na obrázku 5.13. Spektrum bylo získáno opět z pouhých 50 trénovacích geometrií, přičemž excitační vlastnosti pro zbývajících 3950 geometrií byli predikovány pomocí natrénovaného modelu. Toto spektrum porovnáváme znovu s plným spektrem (4000 geometrií) a se spektrem, kde používáme pouze oněch 50 geometrií bez ML modelu. V grafu je také pro srovnání zachyceno experimentální spektrum.[63]

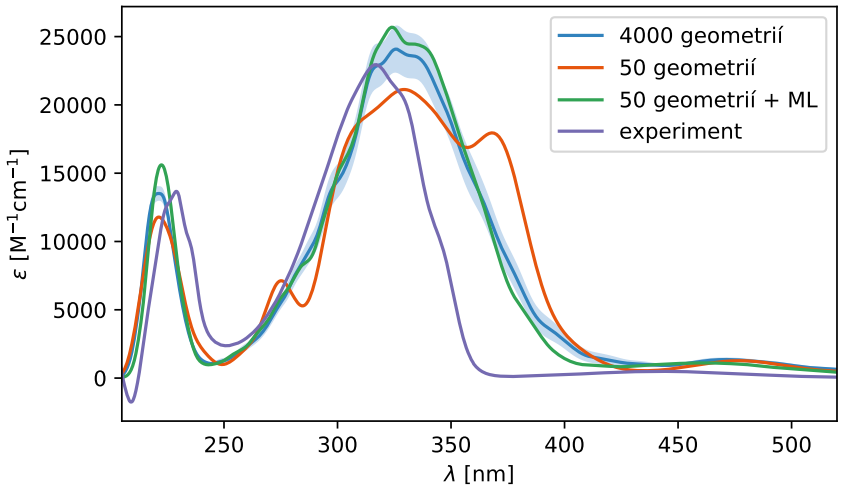


Obrázek 5.11: Porovnání průměrné střední absolutní chyby oscilátorových sil pro molekulu azobenzenu.

Vidíme, že i přes nedokonalé učení dostáváme spektrum víceméně spadající do 95% intervalů spolehlivosti plného spektra. Oproti tomu spektrum získané pro 50 geometrií bez ML modelu vychází značně odlišné. KL divergence od plného spektra vychází pro predikované spektrum 0,012, zatímco pro spektrum z 50 geometrií bez ML 0,028. Zároveň vidíme, že odlišnost od experimentu neboli chyba způsobená nedostatečností kvantové metody je výrazně větší než chyba způsobená použitím ML modelu. Získané výsledky pro molekulu azobenzenu jsou tedy nakonec vcelku povzbudivé. Navíc lze očekávat, že predikce tranzitních dipólových momentů by se výrazně zlepšila při vyřešení problému s pořadím excitovaných stavů.



Obrázek 5.12: Porovnání průměrné střední absolutní chyby nejlepších reprezentací s a bez použití ΔML pro molekulu azobenzenu.



Obrázek 5.13: Srovnání spektra s a bez použití ML pro molekulu (E)-azobenzenu.

Kapitola 6

Závěr

V rámci této práce jsem vyvinul a testoval řadu molekulárních deskriptorů, převážně pak za účelem predikce spektrálních dat. Zaměřil jsem se přitom na dva různé způsoby modelování spekter a s tím spojené dva druhy aplikace strojového učení. První variantou je trénování na velké sadě molekul a predikce pro nové molekuly, zatímco druhou možností je trénování na malém množství geometrií jedné molekuly a predikce pro další geometrie téže molekuly. Ukázalo se, že učení na geometriích v principu umožňuje velmi kvalitní predikce a pro molekulu nejmenšího Criegeeho intermediátu CH_2OO jsem tak dosáhl nevídané přesnosti modelovaného spektra i při redukci potřebných geometrií na pouhých 50 (z původních 1010). Nicméně existují popsané okolnosti, které znesnadňují učení v některých případech. I při problému s rozdílným pořadím excitovaných stavů napříč geometriemi však model učící se na 50 geometriích azobenzenu (z původních 4000) dosáhl vcelku úspěšných výsledků. Spektrum získané pomocí tohoto ML modelu je na hranici uvedených intervalů spolehlivosti pro cílové plné spektrum a to při teoretickém 80násobném zrychlení výpočtu. Při vyřešení problému s pořadím stavů pomocí diabatizace lze navíc očekávat další zlepšení.

I pro predikce na sadách molekul jsem dosáhl velmi dobrých výsledků, které obecně překonaly aktuálně nejlepší publikované výsledky v literatuře. S menším množstvím trénovacích dat jsem byl schopen vyrovnat přesnost na datové sadě QM7b a pro datovou sadu QM8 jsem publikované výsledky dokonce výrazně předčil, a to opět s menším počtem dat.

Některé zde vyvíjené deskriptory se ukázaly býti velmi užitečnými. Mezi nimi je nutno vyzdvihnout deskriptory založené na orbitální energiích, které představují vítanou alternativu k deskriptorům kódujícím strukturu molekuly. Diskrétní verze těchto deskriptorů předčila všechny ostatní při predikci excitačních energií azobenzenu, zatímco distribuce orbitálních energií se ukázala jako nejlepší v případě datové sady QM8, přičemž i v ostatních případech byly tyto deskriptory mezi nejlepšími. I vyvíjené reprezentace založené na distribucích interakcí poskytují velmi dobré výsledky pro učení na datových sadách překonávající alternativní implementace. Na druhou stranu se příliš neosvědčily navrhované reprezentace používající elektronegativity místo nábojů jader.

Jako velmi důležitá se ukázala volba mocninného koeficientu vzdáleností u strukturních reprezentací, obzvláště pak při učení na sadě molekul. Při použití velkých záporných mocnin (-4 až -6) dosahovaly reprezentace výrazně lepších výsledků než s běžnou mocninou -1. Relativně běžná a základní BoB reprezentace se díky tomuto dostala mezi nejlepší reprezentace, i když s běžnou mocninou poskytuje, ve shodě s literaturou, podstatně horší výsledky. Pro učení na sadě geometrií naznačují výsledky kvalitnější predikce pomocí mocniny +1, nicméně rozdíly nejsou tak velké.

Také jsem testoval přínos použití ΔML přístupu. Ukázalo se, že přínos pro predikci excitačních energií je obecně větší než u intenzit absorpce, kde má ΔML mnohdy i negativní efekt. Přínos zároveň typicky klesá s rostoucí přesností modelu. Vše však ovlivňuje korelace mezi použitými kvantově-chemickými metodami a ΔML tak lze systematicky vylepšovat volením vhodnější referenční metody. Obecně lze použití ΔML přístupu doporučit pro predikci excitačních energií při učení na sadě molekul, v ostatních případech je potřeba jeho přínos otestovat na konkrétních datech.

Literatura

[1] Afzal, M. A. F., Cheng, C., and Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *The Journal of Chemical Physics 148*, 24 (6 2018), 241712.

[2] Atkins, P. W. *Fyzikální chemie*, vyd. 1. ed. v Praze, Vysoká škola chemicko-technologická, 2013.

[3] Axilrod, B. M., and Teller, E. Interaction of the van der Waals Type Between Three Atoms. *The Journal of Chemical Physics 11*, 6 (6 1943), 299–300.

[4] Blum, L. C., and Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society 131*, 25 (7 2009), 8732–8733.

[5] Bohm, G., and Zech, G. *Introduction to statistics and data analysis for physicists.* DESY, 2010.

[6] Brown, J. W., Henderson, B. L., Kiesz, M. D., Whalley, A. C., Morris, W., Grunder, S., Deng, H., Furukawa, H., Zink, J. I., Stoddart, J. F., and Yaghi, O. M. Photophysical pore control in an azobenzene-containing metal–organic framework. *Chemical Science 4*, 7 (6 2013), 2858.

[7] Ceriotti, M., Bussi, G., and Parrinello, M. Colored-Noise Thermostats à la Carte. *Journal of Chemical Theory and Computation 6*, 4 (4 2010), 1170–1180.

[8] Ceriotti, M., Manolopoulos, D. E., and Parrinello, M. Accelerating the convergence of path integral dynamics with a generalized Langevin equation. *Journal of Chemical Physics 134*, 8 (2 2011), 84104.

[9] Chatterjee, S., and Hadi, A. S. *Regression analysis by example.* Wiley-Interscience, 2006.

[10] Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K.-R., and Tkatchenko, A. sGDML: Constructing Accurate and Data Efficient Molecular Force Fields Using Machine Learning, 12 2018.

[11] Christensen, A., Faber, F., Huang, B., Bratholm, L., Tkatchenko, A., Muller, K., and von Lilienfeld, O. QML: A Python Toolkit for Quantum Machine Learning, 2017.

[12] Christiansen, O., Koch, H., Jørgensen, P., and Helgaker, T. Integral direct calculation of CC2 excitation energies: singlet excited states of benzene. *Chem. Phys. Lett. 263*, 3-4 (12 1996), 530–539.

[13] Collins, C. MolML.

[14] Collins, C. R., Gordon, G. J., von Lilienfeld, O. A., and Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *The Journal of Chemical Physics 148*, 24 (6 2018), 241718.

[15] Cramer, C. J. *Essentials of Computational Chemistry : Theories and Models*. Chichester, Wiley, 2002.

[16] Dikmen, O., Yang, Z., and Oja, E. Learning the Information Divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 7 (7 2015), 1442–1454.

[17] Eickenberg, M., Exarchakis, G., Hirn, M., Mallat, S., and Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *The Journal of Chemical Physics 148*, 24 (6 2018), 241732.

[18] Faber, F. A., Christensen, A. S., Huang, B., and von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics 148*, 24 (6 2018), 241717.

[19] Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., Vinyals, O., Kearnes, S., Riley, P. F., and von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation 13*, 11 (11 2017), 5255–5264.

[20] Finlayson-Pitts, B. J., and Pitts Jr., J. N. *Chemistry of the Upper and Lower Atmospher: Theory, Experiments, and Applications*, 1st ed. ed. San Diego, Academic Press, 1999.

[21] Gastegger, M., Behler, J., and Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science 8*, 10 (9 2017), 6924–6935.

[22] Graser, J., Kauwe, S. K., and Sparks, T. D. Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons. *Chemistry of Materials 30*, 11 (6 2018), acs.chemmater.7b05304.

[23] Grimme, S. Calculation of the Electronic Spectra of Large Molecules. In *Reviews in Computational Chemistry*, K. B. Lipkowitz, R. Larter, and T. R. Cundari, Eds. John Wiley & Sons, Ltd, 8 2004, ch. 3, pp. 153–218.

[24] Gubaev, K., Podryabinkin, E. V., and Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *The Journal of Chemical Physics 148*, 24 (6 2018), 241727.

[25] Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K.-R., and Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters 6*, 12 (6 2015), 2326–2331.

[26] Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O. A., Tkatchenko, A., and Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *Journal of Chemical Theory and Computation 9*, 8 (8 2013), 3404–3419.

[27] Hehre, W. J. *A Guide to Molecular Mechanics and Quantum Chemical Calculations.* Wavefunction, Inc., 2003.

[28] Hilborn, R. C. Einstein coefficients, cross sections, f values, dipole moments, and all that. *Los Alamos Natl. Lab., Prepr. Arch., Phys.* (2002), 1–12.

[29] Huang, B., and von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics 145*, 16 (10 2016), 161102.

[30] Huang, B., and von Lilienfeld, O. A. The "DNA" of chemistry: Scalable quantum machine learning with "amons", 7 2017.

[31] Huo, H., and Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning, 4 2017.

[32] Klán, P. *Organická fotochemie.* Masarykova univerzita, Brno, 2001.

[33] Lee, S. Y., Brown, R. C., and Heller, E. J. Multidimensional reflection approximation: Application to the photodissociation of polyatomics. *Journal of Physical Chemistry 87*, 12 (6 1983), 2045–2053.

[34] Li, X., Curtis, F. S., Rose, T., Schober, C., Vazquez-Mayagoitia, A., Reuter, K., Oberhofer, H., and Marom, N. Genarris: Random generation of molecular crystal structures and fast screening with a Harris approximation. *The Journal of Chemical Physics 148*, 24 (6 2018), 241701.

[35] Lusci, A., Pollastri, G., and Baldi, P. Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. *Journal of Chemical Information and Modeling 53*, 7 (2013), 1563–1575.

[36] Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, Cambridge, 2004.

[37] Mercer, J. Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 83*, 559 (11 1909), 69–70.

[38] Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R., and Anatole von Lilienfeld, O. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics 15*, 9 (9 2013), 095003.

[39] Muto, Y. Force between nonpolar molecules. *J. Phys. Math. Soc. Jpn 17* (1943), 629–631.

[40] Narsky, I., and Porter, F. C. *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 11 2013.

[41] Ončák, M., Šištík, L., and Slavíček, P. Can theory quantitatively model stratospheric photolysis? Ab initio estimate of absolute absorption cross sections of ClOOCl. *Journal of Chemical Physics 133*, 17 (2010), 174303.

[42] Parrish, R. M., Burns, L. A., Smith, D. G. A., Simmonett, A. C., DePrince, A. E., Hohenstein, E. G., Bozkaya, U., Sokolov, A. Y., Di Remigio, R., Richard, R. M., Gonthier, J. F., James, A. M., McAlexander, H. R., Kumar, A., Saitow, M., Wang, X., Pritchard, B. P., Verma, P., Schaefer, H. F., Patkowski, K., King, R. A., Valeev, E. F., Evangelista, F. A., Turney, J. M., Crawford, T. D., and Sherrill, C. D. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *Journal of Chemical Theory and Computation 13*, 7 (7 2017), 3185–3197.

[43] Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist. 33*, 3 (1962), 1065–1076.

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, Oct (2011), 2825–2830.

[45] Pronobis, W., Schütt, K. T., Tkatchenko, A., and Müller, K.-R. Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *The European Physical Journal B 91*, 8 (8 2018), 178.

[46] Pronobis, W., Tkatchenko, A., and Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *Journal of Chemical Theory and Computation 14*, 6 (6 2018), 2991–3003.

[47] Pyzer-Knapp, E. O., Simm, G. N., and Aspuru Guzik, A. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Materials Horizons 3*, 3 (5 2016), 226–233.

[48] Ramakrishnan, R., Hartmann, M., Tapavicza, E., and Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *Journal of Chemical Physics 143*, 8 (8 2015), 084111.

[49] Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling 52*, 11 (11 2012), 2864–2875.

[50] Rupp, M. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry 115*, 16 (8 2015), 1058–1073.

[51] Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters 108*, 5 (1 2012), 058301.

[52] Rupp, M., von Lilienfeld, O. A., and Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *The Journal of Chemical Physics 148*, 24 (6 2018), 241401.

[53] Russew, M.-M., and Hecht, S. Photoswitches: From Molecules to Materials. *Advanced Materials 22*, 31 (4 2010), 3348–3360.

[54] Samanta, K., Beames, J. M., Lester, M. I., and Subotnik, J. E. Quantum dynamical investigation of the simplest Criegee intermediate CH2OO and its O-O photodissociation channels. *Journal of Chemical Physics 141*, 13 (10 2014), 134303.

[55] Schölkopf, B., and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT Press, 2002.

[56] Silverman, B. W. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1986.

[57] Sita, J. Zrychlení simulací elektronových spekter, 2019.

[58] Sršeň, Š., Hollas, D., and Slavíček, P. UV absorption of Criegee intermediates: quantitative cross sections from high-level ab initio theory. *Physical Chemistry Chemical Physics 20*, 9 (2 2018), 6421–6430.

[59] Taatjes, C. A., Welz, O., Eskola, A. J., Savee, J. D., Scheer, A. M., Shallcross, D. E., Rotavera, B., Lee, E. P. F., Dyke, J. M., Mok, D. K. W., Osborn, D. L., and Percival, C. J. Direct Measurements of Conformer-Dependent Reactivity of the Criegee Intermediate CH3CHOO. *Science 340*, 6129 (4 2013), 177.

[60] Ting, W.-L., Chen, Y.-H., Chao, W., Smith, M. C., and Lin, J. J.-M. The UV absorption spectrum of the simplest Criegee intermediate CH 2 OO. *Phys. Chem. Chem. Phys. 16*, 22 (2014), 10438–10443.

[61] Tuckerman, M. E. Path Integration via Molecular Dynamics, 2002.

[62] Vaníček, J., and Cohen, D. Path integral approach to the quantum fidelity amplitude. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 374*, 2069 (6 2016).

[63] Vetráková, L., Ladányi, V., Al Anshori, J., Dvořák, P., Wirz, J., and Heger, D. The absorption spectrum of cis -azobenzene. *Photochemical & Photobiological Sciences 16*, 12 (12 2017), 1749–1756.

[64] Wardle, B. *Principles and Applications of Photochemistry.* Wiley, 2009.

[65] Yang, G., Wu, J., Chen, S., Zhou, W., Sun, J., and Chen, G. Size-independent neural networks based first-principles method for accurate prediction of heat of formation of fuels. *The Journal of Chemical Physics 148*, 24 (6 2018), 241738.

Seznam použitých zkratek

F_{2B} molekulární reprezentace zahrnující dvouatomové násobné interakce

F_{3B} molekulární reprezentace zahrnující tříatomové násobné interakce

BAML molekulární reprezentace *Bonds and Angles Machine Learning*

BoB molekulární reprezentace *Bag of Bonds*

CC2 aproximativní metoda spřažených klastrů druhého řádu, z angl. *Second-Order Approximate Coupled-Cluster*

CC3 aproximativní metoda spřažených klastrů třetího řádu, z angl. *Second-Order Approximate Coupled-Cluster*

CM coulombovská matice

DFT teorie funkcionálu hustoty, z angl. *Density Functional Theory*

FCHL molekulární reprezentace *Faber-Christensen-Huang-Lilienfeld*

GPR regrese gaussovskými procesy, z angl. *Gaussian Process Regression*

KL Kullback–Leibler

KRR hřebenová regrese s jádrovou transformací, z angl. *Kernel Ridge Regression*

MAE střední absolutní chyba, z angl. *Mean Absolute Error*

MBTR molekulární reprezentace *Many-Body Tensor Representation*

ML strojové učení, z angl. *Machine Learning*

SLATM molekulární reprezentace *Spectrum of London and Axilrod-Teller-Muto potential*

SVM metoda podpůrných vektorů, z angl. *Support Vector Machines*

TDDFT metoda časově závislé teorie funkcionálu hustoty, z angl. *Time-Dependent Density Functional Theory*

ZINDO/S semiepirická metoda, z angl. *Zerner's Intermediate Neglect of Differential Overlap for Spectroscopy*