UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE CHIMICHE

CORSO DI LAUREA IN CHIMICA

MACHINE LEARNING PER LA PREVISIONE DELLA REATTIVITÀ MOLECOLARE IN FASE GAS E IN FASE CONDENSATA

Relatore: Prof. Mirco Zerbetto

Laureanda: Federica Giglio

1224754

Anno Accademico 2021/2022

Indice

Capitolo 1 – Introduzione

Nel corso del XX secolo, sono stati fatti grandi passi avanti per il potenziamento di metodi di calcolo della struttura elettronica e delle proprietà fisiche delle molecole attraverso lo sviluppo informatico: la chimica computazionale ha oggi un'estesa varietà di applicazioni, tra le più promettenti la determinazione delle attività di alcune molecole di interesse farmaceutico, che ne permette lo studio prima di intraprendere costose sperimentazioni cliniche, o la ricerca delle condizioni ottimali di una reazione, prima di procedere con le prove laboratoriali. L'oggetto di studio può spaziare dalle sintesi di nuovi materiali organici nel ruolo di fotocatalizzatori,[1] alla caratterizzazione della struttura elettronica e la reattività di molecole antiossidanti,[2] o ancora all'efficienza di inibitori nel campo della corrosione di metalli e leghe.[3]

Il fisico e chimico R. S. Mulliken, Nobel per la Chimica nel 1966, esponeva nel proprio discorso: "*In conclusion, I would like to emphasize my belief that the era of computing chemists, when hundreds if not thousands of chemists will go to the computing machine instead of the laboratory, for increasingly many facets of chemical information, is already at hand. There is only one obstacle, namely, that someone must pay for the computing time.*"[4] L'enorme difficoltà che rende questo studio spesso scomodo e impraticabile è il costo computazionale per la simulazione di sistemi molecolari medio-grandi. Se, infatti, nelle simulazioni classiche si incontrano le prime difficoltà con sistemi dell'ordine del milione di atomi, quando la trattazione diventa quantistica il limite si restringe al migliaio, soprattutto in presenza di atomi pesanti.

Una svolta è stata segnata dalla nascita del Machine Learning (ML), una branca dell'Intelligenza Artificiale interessata allo sviluppo di modelli informatici in grado di raggiungere i risultati desiderati, senza ricorrere a calcoli esatti, ma basandosi su pattern e previsioni. In chimica, è largamente sfruttato per la previsione di caratteristiche o comportamenti molecolari, permettendo di calcolare costanti cinetiche o valori termodinamici, fino alla reattività delle specie in un dato sistema. Un esempio è lo studio che si sviluppa in merito alle energie di solvatazione delle molecole in soluzione,[5] utile poi per previsioni più accurate della reattività molecolare in fase condensata.[6] L'aspetto che lo rende una soluzione comunemente apprezzata è la sua potenzialità nel veloce immagazzinamento di innumerevoli dati, risparmiando i pesanti costi computazionali delle procedure più datate.

Il presente elaborato si propone di riassumere i progressi degli ultimi anni in merito alla predizione della reattività delle molecole attraverso l'utilizzo del Machine Learning, in sinergia con la Meccanica Quantistica e i suoi principiali metodi di calcolo (Hartree-Fock, DFT), discutendo gli ultimi studi sui sistemi in fase gassosa e in fase condensata. Il Capitolo 2 è dedicato alla descrizione dei metodi del calcolo quantistico, soffermandosi su quello di Hartree-Fock e del Funzionale della Densità, con i quali sono costruiti i set di dati di molti studi di Machine Learning per la previsione della reattività molecolare. Nel terzo Capitolo, dopo un breve resoconto riguardante un progetto studiato personalmente, si presentano alcuni lavori incentrati sulla fusione delle simulazioni quantomeccaniche con il Machine Learning per il calcolo delle energie di solvatazione di molecole organiche in diversi solventi, essendo questo un fenomeno decisivo nella reattività in fase condensata. Infine, si riportano degli studi interessanti che affrontano la questione della reattività molecolare, analizzando le diverse modalità di approccio al problema, di rappresentazione molecolare e di ad-

destramento delle reti neurali, con i rispettivi limiti e i miglioramenti sostenuti. I risultati verranno riassunti nell'ultimo Capitolo, evidenziando gli approcci più efficienti, tra quelli menzionati, e i possibili prospetti per gli anni a venire.

Capitolo 2 – Metodi

Tutti i metodi utili per il calcolo quantistico della struttura elettronica di una molecola si basano sulla risoluzione dell'equazione di Schrödinger indipendente dal tempo: a ogni sistema atomico o molecolare è possibile associare un problema agli autovalori dell'operatore hamiltoniano, $\hat{H}$, attraverso cui si identificano gli autostati della matrice associata e i suoi autovalori, corrispondenti rispettivamente alle funzioni d'onda Ψ e all'energia del sistema.

Gli approcci principali per semplificarne la trattazione si dividono in metodi semi-empirici e metodi *ab initio*. Nei metodi semi-empirici, vengono approssimati gli elementi di matrice dell'Hamiltoniano con forme funzionali analitiche e parametriche. I parametri vengono adattati in modo che dal calcolo sia possibile riprodurre dati sperimentali della molecola (come, ad esempio, l'entalpia di formazione) o della classe di molecole in esame, e sono in grado di condurre studi su molecole con un numero virtualmente illimitato di atomi. I metodi *ab initio*, invece, non prevedono alcuna parametrizzazione, e calcolano le grandezze a partire esclusivamente dai principi fondamentali, assicurando, al costo di una complessità di calcolo più elevata, un risultato più accurato.

Due metodi *ab initio* capostipiti sono il metodo di Hartree-Fock (HF) e la teoria del funzionale della densità (*Density Functional Theory,* DFT). Entrambi sono metodi di campo medio, ossia trattano in maniera mediata le interazioni che subisce un elettrone da parte di tutti gli altri. Di conseguenza, l'operatore Hamiltoniano di ciascun elettrone contiene termini di energia potenziale che dipendono dagli autovalori dell'Hamiltoniano stesso. Usando il metodo dei campi autocoerenti (*Self Consistent*

Fields, SCF), si costruisce, a partire da dei valori di prova degli autostati, gli Hamiltoniani monoelettronici, dei quali si calcolano gli autovettori. Se questi sono identici a quelli di partenza, il processo si conclude. In caso contrario, i nuovi autostati vengono usati per la costruzione di ulteriori Hamiltoniani, e la procedura viene iterata fino alla convergenza.

È in unione a queste grandi teorie che il Machine Learning prende posto nella famiglia di sistemi risolutivi dei quesiti chimici. Di seguito, si introducono brevemente il metodo di Hartree-Fock e DFT, prima di immergersi nel mondo del ML e nelle sue applicazioni.

2.1 Hartree-Fock

Il lavoro condotto da D. R. Hartree alla fine degli anni '20, con il supporto degli studi indipendenti di J. C. Slater e V. Fock,[7] ha dato luogo al metodo *ab initio* di Hartree-Fock. Si tratta di una procedura computazionale iterativa, per risolvere l'equazione di Schrödinger indipendente dal tempo per sistemi polielettronici e, quindi, determinare una stima dell'energia di stato fondamentale.

In un sistema molecolare, gli elettroni sono immersi in un campo di energia potenziale, esercitato dai nuclei che, secondo l'approssimazione di Born-Oppenheimer, si considerano fissi nello spazio. L'equazione di Schrödinger per un sistema con N elettroni prende la seguente forma:

$$\widehat{H}\psi = [\sum_{i=1}^{N}\left(-\frac{\hbar}{2m_i}\nabla_i^2\right) + \sum_{i=1}^{N} V_{\mathrm{el-nuc}}(\boldsymbol{r}_i) + \sum_{i<j}^{N} V_{\mathrm{el-el}}(\boldsymbol{r}_i, \boldsymbol{r}_j)]\psi = E\psi \qquad (1)$$

in cui $\widehat{H}$ è l'operatore Hamiltoniano elettronico, $-\frac{\hbar}{2m_i}\nabla_i^2$ è l'energia cinetica dell'i-esimo elettrone, $V_{\mathrm{el-nuc}}$ è l'energia Coulombiana tra l'i-esimo elettrone e il

nucleo, e $V_{\mathrm{el-el}}$ l'energia Coulombiana tra l'i-esimo e il j-esimo elettrone. Gli orbitali molecolari vengono scritti come combinazione lineare di spin-orbitali monoelettronici, $\chi(\boldsymbol{\tau})$, rappresentata in modo compatto con un solo determinante di Slater:

$$\psi(\boldsymbol{\tau_1}, \dots, \boldsymbol{\tau_n}) = \frac{1}{\sqrt{N!}} |\chi_1(\boldsymbol{\tau_1})\chi_2(\boldsymbol{\tau_2}) \dots \chi_N(\boldsymbol{\tau_N})| \tag{2}$$

dove ψ è la funzione d'onda orbitalica, $\boldsymbol{\tau_i}$ le coordinate spaziali e di spin dell'i-esimo elettrone, N è il numero totale di elettroni, χ_N gli spin-orbitali. Questa forma della funzione d'onda implica che gli elettroni siano indistinguibili tra di loro e che ogni elettrone risenta di un campo esterno generato da tutti gli altri. Per questo motivo, il metodo HF viene definito a campo medio.

Gli spin-orbitali sono autofunzioni di un nuovo operatore, l'operatore di Fock $\hat{F}$, nell'equazione di Hartree-Fock, che costituisce la versione monoelettronica della complessiva di Schrödinger, e che restituisce come autovalore l'energia del singolo spin-orbitale:

$$\hat{F}_i = \hat{h}_i + \sum_{j=1}^{N} (\hat{J}_{i,j} - \hat{K}_{i,j}) \tag{3}$$

$$\hat{F}_i \chi_i = \epsilon_i \chi_i \tag{4}$$

dove $\hat{h}_i$ è l'Hamiltoniano monoelettronico e $\hat{J}_{i,j}$ e $\hat{K}_{i,j}$ sono rispettivamente l'integrale di Coulomb e l'integrale di scambio tra gli elettroni i e j. Ogni spin-orbitale può essere scritto come sviluppo sulla base ortonormale del loro spazio:

$$\chi_i(\boldsymbol{\tau}) = \Sigma_j c_j \varphi_j(\boldsymbol{\tau}) \tag{5}$$

I coefficienti c_j sono parametri scelti appositamente, tramite metodo variazionale, per ottenere la funzione orbitalica alla quale è associata l'energia più bassa e,

quindi, la configurazione dello stato fondamentale del sistema. Il calcolo ha inizio con una stima degli autostati χ_i che vengono usati, nell'equazione (4), per ricostruire gli operatori $\hat{J}$ e $\hat{K}$. Da questi, si calcolano nuovi autostati e autovalori, che a loro volta ridefiniscono gli operatori. Si aggiustano i parametri c_i, ripentendo l'operazione in maniera iterativa, fino a convergenza di questi ultimi, secondo il *SCF*.

Nonostante il metodo di Hartree-Fock riesca a riprodurre con accuratezza proprietà come il momento di dipolo o le intensità degli spettri IR e Raman, o le strutture di equilibrio della maggior parte delle molecole e le energie conformazionali, pecca di alcuni limiti, dal momento in cui trascura le correlazioni elettroniche, sovrastimando le energie.[8] Risulta pertanto impreciso nei processi in cui cambiano il numero o la natura dei doppietti elettronici di legame, come nelle ossidoriduzioni o nelle eccitazioni elettroniche; fallisce, inoltre, nella descrizione della dissociazione omolitica, fenomeno molto comune nella chimica in fase gas. Un'altra limitazione è l'inattuabilità di questo approccio a molecole che superano l'ordine della decina di atomi. Oggi, sono disponibili vari metodi computazionali che si basano sull'Hartree-Fock, ma che correggono alcuni di questi aspetti mancanti (metodi post-Hartree-Fock).

2.2 Density Functional Theory

La teoria del funzionale della densità tralascia la risoluzione delle equazioni di Schrödinger, incentrandosi piuttosto sul calcolo della densità mono elettronica; pertanto, riduce un problema 3N-dimensionale (se N è il numero di elettroni) in uno a tre sole dimensioni. Ha origine dai teoremi di W. Kohn e P. Hohenberg degli anni '60[9]: il primo sostiene che le proprietà dello stato fondamentale di un sistema sono univocamente determinate dalla sua densità elettronica $\rho(\mathbf{r})$, dipendente dalle tre

coordinate spaziali; il secondo introduce il concetto di energia come funzionale della densità, minimizzato per la densità elettronica dello stato fondamentale. Si evince, inoltre, che la somma dell'energia cinetica e potenziale tra gli elettroni sia un funzionale universale della densità di carica ($F[\rho(\mathbf{r})]$), dal momento che la stessa funzione d'onda dipende da $\rho(\mathbf{r})$. Conseguentemente, il funzionale dell'energia, dato un potenziale esterno $v(\boldsymbol{r})$, diventa:

$$E_v[\rho(\mathbf{r})] = F[\rho(\mathbf{r})] + \int v(\boldsymbol{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} \tag{6}$$

Per una densità di prova, $\tilde{\rho}(\mathbf{r}) \geq 0$, l'energia $E[\tilde{\rho}(\mathbf{r})]$ risulterà sempre maggiore rispetto a E_0, energia dello stato fondamentale. Le precedenti equazioni permettono di minimizzare l'energia, modificando il valore della densità. Purtroppo, l'espressione che mette in relazione queste due grandezze non è del tutto nota. La teoria venne tuttavia migliorata, grazie ai contributi di Kohn e L. J. Sham, i quali derivarono delle equazioni comprensive degli effetti di correlazione[10], combinando l'approccio delle funzioni d'onda e della densità: $F[\rho(\mathbf{r})]$ viene esplicitata come somma di $T_\mathrm{s}[\rho(\mathbf{r})]$, energia cinetica del sistema ad una data $\rho(\mathbf{r})$ senza interazioni elettrone-elettrone, ed $E_\mathrm{xc}[\rho(\mathbf{r})]$, energia di scambio e correlazione di un sistema con le interazioni. Se $\rho(\mathbf{r})$ varia lentamente, quest'ultimo termine, dapprima complicato da risolvere, diventa:

$$E_\mathrm{XC}[\rho(\mathbf{r})] = \int \rho(\mathbf{r})\epsilon_\mathrm{XC}\big(\rho(\mathbf{r})\big)\mathrm{d}\mathbf{r} \tag{7}$$

dove $\epsilon_\mathrm{XC}\big(\rho(\mathbf{r})\big)$ è l'energia di scambio e correlazione per singolo elettrone. Sulla base di questi concetti, si deriva l'energia potenziale di Kohn-Sham, v_eff, ossia il potenziale esterno in cui gli elettroni si muovono senza interazioni. La funzione

d'onda di Kohn-Sham è un determinante di Slater che soddisfa l'equazione (simile a quella di Hartree-Fock):

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + v_{\text{eff}}(\boldsymbol{r})\right)\varphi_i(\boldsymbol{r}) = \varepsilon_i\varphi_i(\boldsymbol{r}) \tag{8}$$

e che ha per autovalore l'energia minore (ε_i è l'energia dell'i-esimo orbitale φ_i).

Sebbene i calcoli DFT siano i più usati e di successo, in alcune aree di studio questo metodo incorre in errori significativi: il funzionale può essere trovato in maniera esatta per i sistemi piccoli, ma presto diventa più costoso dell'equazione di Schrödinger stessa. Inoltre, nei calcoli pratici, il contributo XC è frutto di approssimazioni. Un esempio è la sottostima dei gap delle bande di semiconduttori e isolanti[11], la difficoltà nel trattare le interazioni deboli, o l'eccessiva lentezza nelle simulazioni di dinamica molecolare dei liquidi e nella rottura e formazione di legami[12].

2.3 Machine Learning

Il fulcro del Machine Learning è disporre di una macchina in grado di generalizzare i propri apprendimenti sulla base di dati a sua disposizione. Il Machine Learning si divide in quattro tipi di approccio: l'Apprendimento Supervisionato (*Supervised Learning*), in cui si chiede al sistema di prevedere la risposta (*output*) incognita dato un quesito (*input*), sulla base di un insieme di coppie note di *input* e *output*; l'Apprendimento Non Supervisionato (*Unsupervised Learning*), in cui si dà al sistema il compito di raggruppare gli *input* sulla base di caratteristiche simili e di analizzare il nuovo quesito proposto classificandolo per quelle stesse proprietà; l'Apprendimento Semi-Supervisionato (*Semi-Supervised Learning*), una fusione dei

precedenti, per cui solo alcuni dei dati vengono presentati con *output* noti, mentre si lascia che il sistema etichetti i restanti usando le conoscenze acquisite dal primo gruppo; l'Apprendimento per Rinforzo (*Reinforcement Learning*), in cui il programma apprende guidato da un meccanismo di punizioni o premi, imparando la direzione corretta del proprio comportamento e delle proprie previsioni[13].

I dati da fornire al programma possono provenire da calcoli effettuati precedentemente: in questo modo, si coniugano i metodi citati, in grado di raccogliere le risposte a centinaia o migliaia di quesiti, e il Machine Learning, che servirà a prevederne altri milioni in pochissimo tempo. Si può, ad esempio, trovare la soluzione alla densità elettronica dello stato fondamentale, attraverso l'apprendimento di esempi di calcolo DFT per molecole reali, superando il costo computazionale della risoluzione delle equazioni di Kohn-Sham[14]. Inoltre, un programma in grado di aggiornare costantemente il proprio data set con i nuovi risultati da lui calcolati rende lo schema di previsione adattivo e molto più performante[15]. È facile, allora, prevedere come i nuovi approcci possano rappresentare un importante supporto al calcolo scientifico intervenendo per alleggerire parti del calcolo che in genere richiederebbero troppo tempo.

Capitolo 3 - QM/ML per predire reazioni chimiche

3.1 Progetto MMC

Allo scopo di esemplificare come il Machine Learning possa essere impiegato per predire la reattività chimica sostituendo la previsione di proprietà quantomeccaniche al loro calcolo esatto, è stato congegnato un semplice esperimento in cui l'intelligenza artificiale deve individuare quale sia l'atomo di carbonio con maggiore densità elettronica in un diene coniugato pluri-sostituito, al fine di individuare il sito più probabile per una sostituzione elettrofila. È stato effettuato l'addestramento supervisionato, sulla piattaforma MatLab, di una rete neurale costituita da due livelli, una funzione sigmoide nel livello nascosto, con numero di neuroni variabile, e una funzione di trasferimento lineare nel livello di output, composto da un unico neurone. I calcoli della densità elettronica per il raccoglimento dei dati sono stati effettuati attraverso il programma di calcolo quantistico GAMESS, secondo il modello AM1 (Austin Model 1), il quale, considerando gli elettroni dei gusci di valenza di tutti gli atomi della molecola, mostra la densità di carica su ogni singolo atomo.[16] Il set di dati è stato costruito con 90 esatrieni coniugati; ogni atomo della catena poteva avere al massimo un sostituente diverso dall'idrogeno, scelto casualmente tra F, Cl, Br, I, OH e NH_2. Per studiare l'efficacia dell'addestramento in funzione del tipo di descrittore, sono state usate diverse modalità per indicare la connettività delle molecole: in un file Excel, sei colonne consecutive rappresentavano le sei posizioni lungo la catena. Ogni casella della stessa riga, pertanto, veniva riempita con un valore, che indicava il tipo di sostituente legato all'atomo di carbonio situato in quella posizione. Il valore poteva essere un'etichetta numerica, per cui veniva assegnato un numero ad ogni atomo o gruppo funzionale; la massa, per cui si inseriva la massa dell'atomo diretta-

mente legato al carbonio, in g/mol; l'elettronegatività; il raggio di van der Waals, in pm.

Tabella 1 - Dati usati per descrivere i sostituenti nel progetto di ML.

	H	**F**	**OH**	**Cl**	**Br**	**I**	**NH_2**
Etichetta numerica	0	1	2	3	4	5	6
Elettronegatività	2.2	3.98	3.44	3.16	2.96	2.66	3.04
Massa (g/mol)	1.008	18.998	15.999	35.453	79.904	126.904	14.007
Raggio di vdW (pm)	120	147	152	175	185	198	155

A seguire, altre sei caselle della stessa riga riportavano le densità elettroniche calcolate da GAMESS per i rispettivi atomi di carbonio. Il file Excel veniva letto dalla rete neurale, che chiedeva quale tra le colonne conteneva i valori di output e quali colonne considerare valori di input. Dunque, procedeva usando il 70% delle molecole per l'addestramento, il 15% per la validazione, ossia la verifica che il programma stia generalizzando le informazioni, e il restante 15% come test, per determinare l'accuratezza della previsione. Al termine, restituiva il valore della regressione, indice dell'efficacia dell'addestramento. Si sono potuti confrontare, così, i diversi tipi di *training*. Come da supposizioni, quest'ultimo migliorava all'aumentare del numero di molecole componenti il set di dati, all'aumentare del numero di neuroni nel livello nascosto, e all'uso di un descrittore legato alle proprietà chimico-fisiche dei sostituenti, contro la casuale etichetta numerica. Si possono confrontare le regressioni ottenute al variare di uno di questi parametri: mantenendo lo stesso numero di neuroni e stesso tipo di descrittore, la regressione per l'addestramento con 20 molecole è 0.78, mentre con il set da 90 aumenta a 0.98; diminuendo il numero di neuroni da 10

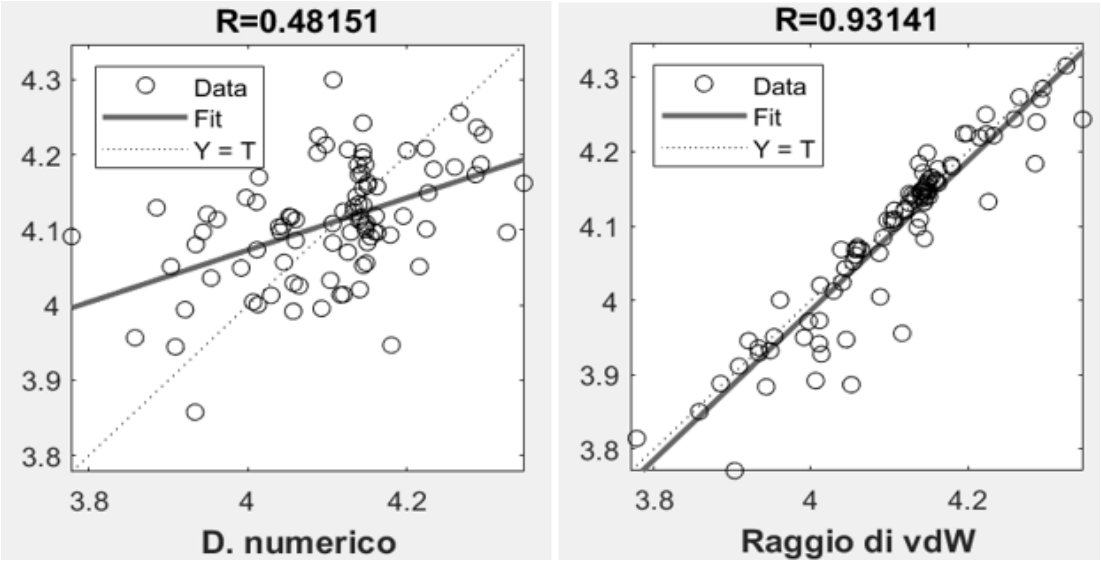

Figura 1 - Confronti delle regressioni con diversi descrittori.

a 5, quest'ultimo addestramento ottiene un R^2 di 0.91; si nota, infine, una grande differenza al variare del tipo di descrittore, per cui per lo stesso set di dati di 90 molecole processato da 10 neuroni, la regressione dell'addestramento con descrittore numerico è di 0.48, quando con la descrizione del raggio di van der Waals è di 0.93. Per testare la previsione della rete neurale su molecole esterne al dataset, sono stati effettuati i calcoli della densità elettronica, tramite GAMESS, di cinque nuovi esatrieni: il 4-cloro-1-iodio-5-idrossi-1,2,3-esatriene (1), il 3-ammino-1-fluoro-1,2,3-esatriene (2), l'1-bromo-4-iodio-1,2,3-esatriene (3), il 2,4-dicloro-1,2,3-esatriene (4), il 4,5-dibromo-2-idrossi-1,2,3-esatriene (5).

Tabella 2 - Previsioni della densità elettronica sugli atomi di carbonio di molecole di prova, effettuate dalla rete neurale (NN) e confrontate con il calcolo AM1 effettuato su GAMESS (AM1).

MOLECOLA	METODO	C1	C2	C3	C4	C5	C6
1	AM1	4.3432*	4.0804	4.1429	4.0501	3.9667	4.2220
	NN	4.2821*	4.0444	4.1574	3.9967	3.9851	4.2236
2	AM1	4.0020	4.1798	3.8802	4.2382*	4.1102	4.2334
	NN	4.0074	4.2849*	3.8603	4.2758	4.0872	4.2404
3	AM1	4.2217*	4.1123	4.1036	4.1916	4.1402	4.1937
	NN	4.2291*	4.0676	4.0515	4.2282	4.1464	4.2241
4	AM1	4.1953*	4.0540	4.1152	4.1153	4.0540	4.1953*
	NN	4.2042	4.0147	4.1111	4.0788	4.0280	4.2213*
5	AM1	4.3391*	3.8945	4.0711	4.1218	4.1531	4.1600
	NN	4.2414*	3.8753	4.1297	4.0575	4.1756	4.1632

L'errore quadratico medio risulta essere 0.032; la rete neurale riesce ad individuare correttamente il sito di attacco elettrofilo più probabile 4 volte su 5 (nella molecola (4), essendo questa simmetrica, i siti ugualmente probabili secondo il metodo AM1 sono i due terminali; si nota che la rete neurale ha previsto due densità diverse per i due atomi di carbonio, classificando ad ogni modo la densità in C1 come la seconda più alta nella molecola).

Questo progetto, sebbene elementare, può essere un buon esempio dei meccanismi che regolano i programmi di Machine Learning professionali descritti di seguito, poiché, in scala molto minore, è riuscito a mostrare chiaramente quanto la mole di dati e un uso appropriato dei descrittori influiscano sull'accuratezza delle previsioni.

3.2 ML per la previsione delle energie di solvatazione

3.2.1 Delfos

Delfos è un modello di *deep learning* del 2019 usato per la previsione delle energie di solvatazione di soluti organici in solventi organici.[5] Unisce il modello QSPR (*Quantitative Structure-Property Relationship*) con una rete neurale ricorsiva, per cui la propagazione delle informazioni può avvenire anche da un livello successivo a quello precedente, tra due neuroni appartenenti allo stesso livello, o da un neurone a se stesso. Il set di dati è stato prelevato dal database MNSOL,[17] e comprende 418 soluti e 91 solventi, in 2'495 combinazioni di soluzioni. Le specie chimiche sono descritte adoperando il programma Mol2Vec, un modello di Machine Learning ad apprendimento non supervisionato, che genera una rappresentazione vettoriale delle molecole.[18] La rete è suddivisa in tre parti: le prime due hanno il compito di codificare il solvente e il soluto, la terza è incaricata della previsione vera e propria. È stata impiegata una validazione incrociata in 10 parti, per cui l'intero set di dati viene sud-

diviso in 10 parti uguali, nove delle quali servono per l'addestramento, mentre la decima è dedicata alla validazione. Il coefficiente di regressione, per questo modello, è 0.96. L'errore quadratico medio nella previsione dell'energia libera di idratazione è di 0.64 kcal mol^{-1}, mentre per le energie di solvatazione in solventi organici risulta di 0.24 kcal mol^{-1}. Un grande limite di questo modello è la non generalizzabilità: la performance cala drasticamente al presentarsi di specie non presenti nel dataset (l'errore quadratico medio è di circa 0.60 kcal mol^{-1}).

3.2.2 QM9-solvation

Nel 2021 Ward et al. si sono impegnati nella costruzione di un set di dati di energie di solvatazione, QM9-solvation, per oltre 130'000 molecole in cinque diversi solventi (acetone, etanolo, acetonitrile, dimetilsolfossido, acqua), e nella programmazione di modelli di Machine Learning che implementassero il set per la previsione di ulteriori valori.[19] L'energia di solvatazione, ΔG_{sol}, è definita come segue:

$$\Delta G_{\mathrm{sol}} = E_{\mathrm{solv}} - E_{\mathrm{gas}} \tag{9}$$

dove E_{solv} è l'energia delle molecole all'interno del solvente specifico, e E_{gas} è la loro energia in fase gas. I calcoli sono stati effettuati sul modello DFT sul software Gaussian 16.[20] Sono state raccolte le energie di solvatazione di 130'258 molecole dal database QM9,[21] poste nei cinque solventi, arrivando ad un totale di 651'290 valori. Questi sono stati divisi tra addestramento, validazione e test. La struttura del modello ML è quella della rete neurale a passaggio di messaggi (MPNN), la quale funziona direttamente sulle rappresentazioni a grafi, che includono le proprietà delle molecole e dei solventi considerati.[22] Il modello impara, durante il *training*, ad interpretare correttamente le caratteristiche atomiche e dei legami descritte nei grafi, e a collegarle alle energie di solvatazione corrispondenti.

Questa rete neurale ha ottenuto, per le energie di idratazione, un errore assoluto medio di 0.44 kcal mol^{-1}. I composti per cui si presentano errori più significativi sono quelli con energie di solvatazione sensibilmente differenti rispetto a tutti gli altri, o quelli della classe degli zwitterioni, che nel dataset sono poco rappresentati.

Nel tentativo di generalizzare il programma a solventi non presenti nell'addestramento, gli autori hanno realizzato un modello che predicesse le energie di solvatazione con *input* delle rappresentazioni delle molecole e della costante dielettrica del solvente. La regressione rimane pressoché inalterata, per cui si può dire che il programma sia in grado di predire efficientemente anche nel caso di solventi esclusi dal set di addestramento. Di contro, le previsioni peggiorano all'aumentare delle dimensioni delle molecole prese in esame.

3.3 ML per le reattività molecolari

3.3.1 Reaction2Barrier

Gli autori del programma Reaction2Barrier (R2B) sfruttano il Machine Learning supervisionato per presentare un modello di previsione dei risultati di una reazione in presenza di meccanismi competitivi. Le reazioni considerate sono prelevate dal dataset QMrxn20,[23] e contano 1'286 reazioni di eliminazione E2 e 2'361 sostituzioni nucleofile S_N2, con le relative energie di attivazione. I reagenti di partenza sono molecole di etano sostituite (i gruppi funzionali possono essere idrogeni, gruppi nitro, cianuri, metili e ammine) alle quali sono legate diversi gruppi uscenti (fluoruri, cloruri, bromuri), e dei nucleofili (idruri, fluoruri, cloruri, bromuri) che possono attuare una sostituzione. I metodi di rappresentazione testati sono vari: BoB è una rappresentazione matriciale dei valori di repulsione Coulombiana tra i nuclei della molecola; SLATM descrive le dispersioni di London e i potenziali di Axilrod-Teller-Muto;

FCHL19 contiene le distanze interatomiche e gli angoli di legame; la codifica *one-hot* consiste in un vettore in cui gli elementi sono tutti zero, a meno di uno, per cui si usa un vettore per ogni sito di sostituzione occupato dai gruppi funzionali, uno per ogni nucleofilo e uno ogni gruppo uscente.

Anche in questo caso, l'addestramento è condotto con la validazione incrociata in *k* parti. La previsione risulta migliorare all'aumentare del set di dati: la regressione diventa sempre più accurata, registrando valori di 0.89 per le reazioni E2 e 0.94 per le S_N2. Inoltre, il confronto tra le rappresentazioni vede l'addestramento migliore con il descrittore *one-hot*, sebbene sia l'unico a non esprimere le caratteristiche geometriche delle molecole. Probabilmente, la sua semplicità permette una performance più efficace, perché il fattore principale che influisce sulla reattività in questo tipo di chimica è l'effetto induttivo, scollegato dalla geometria tridimensionale, ma espresso sufficientemente dalla sola connettività. Il programma, usando questo tipo di *input*, riesce a predire esaustivamente anche sistemi non presenti nel set di addestramento.

3.3.2 ml-QM-GNNs

Un approccio più recente è rappresentato dallo studio di T. Stuyver e C. W. Coley, una rete neurale per la previsione delle reazioni E2, S_N2 e di sostituzione aromatica, che sfrutta i grafi, ma include, come passaggio intermedio, la costruzione di una rappresentazione quantomeccanica delle molecole, prima della previsione vera e propria. Lo scopo di questa decisione è quello di scoperchiare la "scatola nera" dei modelli di intelligenza artificiale: il ragionamento che risiede dietro le previsioni delle reti neurali è spesso incognito. Nel momento, pertanto, in cui un modello mostra delle lacune, è estremamente difficile, se non impossibile, individuare la falla nel sistema. L'espediente di Stuyver e Coley prova a superare questo ostacolo.

I dati, in questo caso, sono estratti da due dataset, uno di natura computazionale, per le reazioni E2 e S_N2 in fase gas, derivante da QMrxn20,[23] l'altro esclusivamente sperimentale, per le sostituzioni aromatiche, dal database Pistachio.[24] Si contano 1'286 reazioni E2, 2'361 reazioni S_N2 e 3'242 reazioni aromatiche di sostituzione regioselettiva o funzionalizzazione. Il metodo di *input* iniziale è quello a grafi, usato poi dalla rete neurale per la previsione degli attributi quantomeccanici di atomi e legami presenti nelle molecole. Le informazioni raccolte vengono usate per produrre le energie di attivazione, nel caso di un problema di regressione, o per determinare la regioselettività di una reazione, nel caso di un problema di classificazione.

L'implementazione dei descrittori quantomeccanici assicura una performance migliore, ulteriormente perfezionata al progressivo aumentare dei dati inseriti per l'addestramento. Il modello di Stuyver e Coley non ottiene risultati migliori rispetto al R2B che usa la codifica *one-hot* ma, al contrario di quest'ultimo, si comporta più efficientemente al momento della generalizzazione. Per provare ciò, il set dell'addestramento è stato privato di uno dei quattro nucleofili, che rimaneva solo nei dati per il test. Mentre i classici modelli di rete neurale a grafi perdono sensibilmente di accuratezza, il modello ml-QM-GNN mantiene un errore quadratico medio di 8-9 kcal mol^{-1} all'esclusione di idruro o fluoruro, e 6-5 kcal mol^{-1} per cloruro o bromuro. Il passaggio per un descrittore quantomeccanico, prima della previsione della reattività, è inequivocabilmente un buon metodo per migliorare la generalizzabilità di un modello di apprendimento.

3.3.3 ReactionPredictor

Un progetto del 2012, condotto da Matthew A. Kayala e Pierre Baldi, presso l'*Insitute for Genomics and Bioinformatics and Department of Computer Science*

dell'Università di California, è un importante esempio di utilizzo del Machine Learning per la previsione delle reazioni chimiche, capace anche di descriverne i meccanismi elementari.[6] Già nel 2011, gli stessi autori, in collaborazione con Chloé-Agathe Azencott e Jonathan H. Chen, avevano provato a costruire un modello che apprendesse da degli esempi di reattività, piuttosto che da delle regole impostate manualmente.[25] Avevano così dimostrato che l'unico limite di un programma di Machine Learning fosse l'estensione del set di addestramento. È stato condotto un confronto con precedenti studi, tra cui Reaction Explorer, un sistema di previsione basato su 1'500 regole di trasformazione, inevitabilmente limitato nel tipo e nella mole di informazioni chimiche (include solo 80 modelli di reagente, tra i più comuni).[26] Inoltre, questo programma non scala ottimamente, in quanto aggiungere delle regole per predire nuove reazioni richiede la revisione di tutte le altre. Questo problema vuole essere superato, dagli ideatori di ReactionPredictor, con l'utilizzo del Machine Learning, per cui è necessario il solo inserimento di nuovi esempi per coprire ulteriori aspetti della reattività. Apprendendo da un set di dati, proveniente dallo stesso Reaction Explorer, il prototipo di Kayala, Baldi et al. riesce a dare una previsione corretta l'89.05% delle volte. Il lavoro del 2012 nasce dalla volontà di ampliare e ottimizzare il primo tentativo, arricchendo l'apprendimento della macchina con reazioni ricavate da libri di testo di chimica organica che spiegano in maniera esaustiva reazioni polari, radicaliche e pericicliche con una vasta tipologia di atomi, anche pesanti e ipervalenti. Il set finale conta 5'551 reazioni polari, 97 radicaliche e 294 pericicliche.

Gli input devono dare una descrizione delle molecole reagenti e delle condizioni di reazione. I descrittori molecolari sono costituiti da dei grafi per l'assegnazione elettronica, in cui ogni atomo viene etichettato come vuoto, se ha orbitali non occupati, pieno, se ha elettroni di non legame, o neutrale, se non ha nessuna delle prece-

denti caratteristiche; dei grafi per l'assegnazione dello stato di ossidazione, che etichettano ogni atomo con il proprio stato di ossidazione, calcolato automaticamente dal programma; inoltre, si etichettano i legami come singoli, σ, o multipli o aromatici, π. Le condizioni di reazione comprendono la temperatura, le condizioni di fotoeccitazione e, per implementare l'effetto solvente, i potenziali di solvatazione anionici e cationici, individuati per mezzo di un valore, che va da 0 a 1, e che quantifica il comportamento del solvente da apolare a polare, o da aprotico a protico.

La struttura della rete neurale artificiale è divisa in due livelli, uno nascosto e uno di output. Entrambi usano delle funzioni sigmoidali. L'addestramento è effettuato con la discesa stocastica del gradiente e uno schema di apprendimento adattivo dei pesi. I parametri della struttura, come il numero di neuroni nascosti o il numero di epoche per la convergenza, sono scelti in base al problema, di volta in volta.

Il sistema è programmato in modo leggermente diverso in base al tipo di reazione sottoposta, ma i modelli che sfrutta sono molto simili fra loro, e si dividono in tre categorie: il primo modello ha il compito di identificare, all'interno delle molecole reagenti, i siti che possono fungere da accettori o donatori di elettroni, per le reazioni polari e radicaliche, o delle proposte di chiusura ad anello, per le reazioni pericicliche, creando una lista di tutte le reazioni che potrebbero avvenire; un secondo modello è addestrato per una scrematura, ossia l'eliminazione di quelle reazioni che comprendono cattivi accettori e donatori di elettroni, quindi meno probabili rispetto ad altre; infine, un terzo modello individua, tra le proposte rimaste, la più favorita.

Usando il set di dati per l'addestramento e la validazione, è stata ottenuta un'accuratezza nella catalogazione dell'80.5% per le reazioni polari, del 78.7% per le radicaliche, e dell'88.1% per le pericicliche. Nota l'efficacia dell'addestramento, gli

autori hanno voluto implementare la possibilità di concatenare le previsioni e creare un meccanismo multistep che, dai reagenti e le condizioni di reazione, sia in grado di condurre al composto desiderato. Il programma tenta le reazioni più probabili per le specie reagenti, reiterando il procedimento finché non incontra il target tra i prodotti risultanti. Per tutti gli esperimenti effettuati, il modello riesce a visualizzare un meccanismo realistico, ponendo le reazioni corrette di ogni step nella top 3 delle sue classifiche. Un inconveniente di questo sistema, tuttavia, è l'elevato tempo computazionale: il sistema di identificazione dei siti reattivi è stato impostato al fine di rilevarne in eccesso, per non tralasciarne alcuno, e ogni combinazione, anche quando falsa positiva, viene esplorata come possibile percorso di sintesi.

Nel 2017, il ragionamento dietro ReactionPredictor viene ripreso da D. Foshee, A. Mood, et al. allo scopo di ampliarne il set di addestramento e conseguentemente la performance del modello.[27] Inoltre, dimostrano con ottimi risultati la validità del *training* basato sulle stringe SMILES, con l'utilizzo di un'architettura LSTM (*long short-term memory*). I dati di addestramento constano di 4'667 reazioni polari elementari provenienti dal dataset del RP del 2012, arricchite con altri 6'361 esempi dalla letteratura, mentre per il test sono state scelte, dalla letteratura, 289 reazioni polari multistep. In seguito, è stato adoperato un software per la generazione di altre decine di migliaia di reazioni elementari, sulla base del dataset. La rappresentazione delle molecole include le loro caratteristiche fisico-chimiche e topologiche, come la carica formale e parziale, la presenza di orbitali pieni o vuoti, di doppietti elettronici non condivisi, la connettività atomica.

Il progetto comprende dei modelli di filtraggio dei siti reattivi che usano un Percettrone multistrato (MLP), ossia una rete neurale che processa i dati di *input* insieme

a opportuni dati di *output*. L'addestramento previsto per questa rete neurale identifica il sito accettore e il sito donatore di ogni reazione elementare, e due siti che non partecipano. Una seconda rete neurale viene addestrata per ordinare le reazioni considerate dalla più alla meno favorita. I risultati di previsione sul set di test del 289 reazioni mostrano un'accuratezza dell'83%, per un totale di 10'812 proposte; il prototipo del 2012, sullo stesso gruppo, riesce, più lentamente, ad ottenere il 58.1% di precisione, pur proponendo 92'158 reazioni, in quanto tra queste figurano tante coppie donatore/accettore non plausibili. Questo risultato mostra un grande miglioramento del programma ReactionPredictor.

3.3.4 MCA e MAA per la reattività molecolare

Uno studio più recente, con i medesimi propositi, addestra una rete neurale attraverso un set di reazioni costruito usando i calcoli DFT.[28] Le simulazioni quantomeccaniche, seppur accurate, richiedono tempi piuttosto lunghi, anche di diverse ore, per molecole relativamente piccole, con qualche decina di atomi. M. Tavakoli, A. Mood, D. Van Vraken e P. Baldi si propongono, nel 2022, di sviluppare quattro modelli di Machine Learning, con il fine di smaltire i costi computazionali e trovare il miglior metodo di descrizione delle molecole per problemi di questo tipo. In primo luogo, viene specificata la metrica usata per misurare la reattività di un composto chimico: i calcoli si basano sulla determinazione della variazione di energia in seguito alla combinazione di una specie con un catione metile (affinità con il catione metile, MCA) o con un anione metile (affinità con l'anione metile, MAA). Questi valori sono strettamente correlati alle proprietà nucleofiliche ed elettrofiliche di una specie; pertanto, possono essere usati per quantificarne la reattività. Inoltre, possono essere calcolati implementando la solvatazione, caratterizzando esaustivamente il comportamento chimico dei composti anche in condizioni di reattività in fase condensata (in

questo caso, le sigle sono appuntate con un asterisco, MCA* e MAA*). Il set di dati, pertanto, è costituito dai valori di MCA* di 1'232 nucleofili, che includono idrocarburi funzionalizzati con ammine, eteri, ammidi e ioni ammide, carbanioni, aldeidi, chetoni, esteri, acidi carbossilici, enolati, anioni nitronato, composti diazotati, cianuri, immine, nitrili; e dai valori di MAA* di 1'189 elettrofili, funzionalizzati con ioni imminio e ossonio, immine, aldeidi, chetoni, esteri, ammidi, cationi alchilici, benzilici e allilici, carbonili, nitrili, nitro.

Le rappresentazioni usate sono quattro: un vettore di *fingerprint* con informazioni sulla reattività dei singoli atomi; uno di *fingerprint* esteso alla connettività (ECFP), con informazioni sulla reattività della molecola; una stringa SMILES per la reattività molecolare; una rappresentazione a grafo per la reattività atomica, dei gruppi funzionali e molecolare. Per ogni rappresentazione, si è sviluppato un modello di ML, analizzando a che livello l'apprendimento risulta più efficace. Ogni rete neurale usa la validazione incrociata in 10 parti.

I risultati mostrano come la modalità di rappresentazione vettoriale delle caratteristiche atomiche e molecolari, sebbene porti ad una buona performance, non permette una facile generalizzazione dell'apprendimento, essendo costituita da informazioni su caratteristiche scelte arbitrariamente. La rappresentazione tramite stringa SMILES non riporta una buona accuratezza, probabilmente perché è stata richiesta la trascrizione in testo a partire da grafi molecolari, che può portare ad errori nella traduzione. La rappresentazione a grafi si assicura pertanto la performance migliore, con un'accuratezza del 92%.

Capitolo 4 – Conclusioni

Nel presente lavoro di Tesi, si sono voluti evidenziare i più recenti sviluppi nello studio dei sistemi chimici attraverso Machine Learning, affiancato dalla chimica computazionale. Si può dedurre, confrontando tutti gli studi riportati, che raggiungere una migliore generalizzazione delle informazioni nel *dataset* permette di ottenere previsioni libere da *bias*. Questo si è visto possibile quando l'addestramento di una rete neurale, indipendentemente dal tipo di descrittore o dalla struttura del modello, può contare di un grande set di dati per assicurare la maggiore varietà di esempi. In secondo luogo, è importante scegliere correttamente le modalità di descrizione dei sistemi: delle etichette del tutto scollegate dagli attributi chimico-fisici delle molecole danno dei risultati meno accurati rispetto ad una descrizione delle geometrie e delle proprietà degli atomi.

Negli ultimi decenni si è riscontrata una veloce crescita nel campo del Machine Learning come soluzione all'elevato costo computazionale dei modelli tradizionali. La sinergia tra la meccanica quantistica e l'intelligenza artificiale permette di migliorare l'accuratezza dei calcoli senza aumentarne le tempistiche: i metodi computazionali possono adesso essere usati come buone basi per ricavare dei set di dati per l'addestramento di reti neurali, in grado di risolvere in minor tempo i quesiti richiesti. È facile immaginare come nei prossimi anni questo approccio possa riuscire ad interpretare modelli sempre più complessi, diventando un affidabile e pratico metodo di uso comune in ogni branca della chimica.

Riferimenti bibliografici

[1] Prentice, A. W.; Zwijnenburg, M. A. *Adv. Energy Mater.* **2021**, *11*, 2100709.
[2] Spiegel, M. *J. Chem. Inf. Model.* **2022**, *62*, 2639-2658.
[3] Abeng, F. R.; Nyong, B. E.; Ikpi, M. E. Obeten M. E. *Portugaliae Electrochimica Acta* **2022**, *40*, 243-248.
[4] Mulliken, R. S. *Nobel Lecture*, **1966**.
[5] Lim, H.; Jung, Y. *Chem. Sci.* **2019**, *10*, 8306-8315.
[6] Kayala, M. A.; Baldi, P. *J. Chem. Inf. Model.* **2012**, *52*, 2526-2540.
[7] Slater, J. C. *Phys. Rev*. **1951**, *81*, 385.
[8] Rauk, A. *Orbital Interaction Theory of Organic Chemistry. John Wiley & Sons, Inc*. **2001**.
[9] Hohenberg, P.; Kohn, W. *Phys. Rev*. **1964**, *136*, B864.
[10] Kohn W.; Sham, L. J. *Phys. Rev*. **1965**, *140*, A1133.
[11] Bagayoko, D. *AIP Advances* **2014**, *4*, 127104.
[12] Burke, K.; *J. Chem. Phys*. **2012**, *136*, 150901.
[13] El Naqa, I.; Murphy, M. J. *What is Machine Learning?* **2015**.
[14] Brockherde, F.; Vogt, L.; Li, L; et al. *Bypassing the Kohn-Sham equations with machine learning. Nat Commun* **2017**, *8*, 872.
[15] Botu, V.; Ramprasad, R. *International Journal of Quantum Chemistry* **2015**, *115*, 1074–1083.
[16] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902-3909.
[17] Marenich, A. V., Felly, C. P.; Thompson, J., D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database* Version 2012; University of Minnesota: Minneapolis, 2012.
[18] Jaeger, S.; Fulle, S.; Turk, S. *J. Chem. Inf. Model.* **2018**, *58*, 27-35.
[19] Ward, L.; Dandu, N.; Blaiszik, B.; Narayanan, B.; Assary, R. S.; Redfern, P. C.; Forser, I.; Curtiss, L. A. *J. Phys. Chem. A* **2021**, *125*, 5990-5998.
[20] Frisch, M. J.; et al. *Gaussian 16*, Revision C.01; Gaussian Inc.: Wallingford CT, 2016.
[21] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; et al. *Sci Data* **2014**, *1*, 140022
[22] Gilmer, J.; Schoenholts, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *34th International Conference on Machine Learning*, **2017**, 2053-2070.
[23] von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026
[24] N. Software, Pistachio, 2021.
[25] Kayala, M. A.; Azencott, C.; Chen, J. H.; Baldi, P. *J. Chem. Inf. Model.* **2011**, *51*, 2209-2222.
[26] Chen, J. H.; Baldi, P. *J. Chem. Inf. Model.* **2009**, *49*, 2034-2043.
[27] Fooshee D., Mood A., Gutman E., Tavakoli M., Urban G., Liu F., Huynh N., Van Vranken D., Baldi P., *Mol. Syst. Des. Eng.* **2018**, *3*, 442.
[28] Tavakoli, M.; Mood, A.; Van Vranken, D.; Baldi, P.; *J. Chem. Inf. Model.* **2022**, *62*, 2121−2132.