PLACEHOLDER

Joan Tibau Terma

**Resum** — La simulació de dinàmica molecular (DM) és una tècnica computacional que estudia sistemes moleculars, proporcionant informació sobre l'estructura, l’energia, la dinàmica i altres propietats. Té aplicacions en la indústria farmacèutica, biologia, enginyeria de materials, física i química. S'utilitzen mètodes com Verlet per a resoldre les equacions del moviment simultàniament a mètodes de mecànica quàntica per a calcular les forces i energies d'interacció entre àtoms. Amb sistemes complexos, es requereixen tècniques d'optimització i paral·lelització. Les noves tècniques d'aprenentatge computacional, com les xarxes neuronals, ofereixen alternatives per abordar els problemes de complexitat de la DM. Aquest projecte té com a objectiu de l’estudi i l'aplicació de xarxes neuronals a simulacions físiques per a la predicció de propietats moleculars. S'utilitzat la toolbox SchNetPack2 que ptoporciona eines per a aplicar xarxes neuronals per a la predicció de propietats de bases de dades com ara QM9. S’ha realitzat una anàlisi detallada dels resultats obtinguts i s'han extret conclusions rellevants per a la millora del model.

**Paraules clau** — Analítica de Dades, xarxes neuronals, dinàmica molecular, base de dades QM9, SchNet, predicció de propietats, anàlisi de resultats, optimització.

**Abstract** — Molecular dynamics (MD) simulation is a computational technique that studies molecular systems, providing information about structure, energy, dynamics, and other properties. It has applications in the pharmaceutical industry, biology, materials engineering, physics, and chemistry. Numerical methods like Verlet or Gear are used to solve the equations of motion and calculate interaction forces. With complex systems, optimization and parallelization techniques are required. New computational learning techniques, such as neural networks, offer alternatives to address the complexity of MD. This project aims to study and apply neural networks to physical simulations for the prediction of molecular properties. The SchNetPack2 toolbox is used, providing tools for applying neural networks to predict properties from databases like QM9. A detailed analysis of the obtained results has been conducted, drawing relevant conclusions for model improvement.

**Index Terms** — Data Science, neural networks, molecular dynamics, QM9 database, SchNet, property prediction, result analysis, optimization.

—————————— ◆ ——————————

# 1 Introducció:

L

A simulació de dinàmica molecular (MD) és una tècnica computacional amplament utilitzada per a l'estudi del comportament de sistemes moleculars. Per dur a terme aquestes simulacions es resolen les equacions de moviment de grans quantitats de molecules, tenint en compte per cada pas de temps les energies i forces internes de les molecules. Aquesta tècnica té una àmplia gamma d'aplicacions en àrees com la indústria farmacèutica, la biologia, l'enginyeria de materials, la física i la química.

————————————————

1. E-mail de contacte: jotite19@gmail.com
2. Menció realitzada: Computació
3. Treball tutoritzat per: Ramon Baldrich
4. Tutoria externa per: Jordi Faraudo
5. Curs 2022/23

# 2 Motivació i objectius:

Un dels principals problem~~e~~s del camp de la MD és que a mesura que els sistemes es fan més complexos augmenta la quantitat de dades a processar i, per tant, augmenta també el temps requerit per a cada pas de la simulació. Per exemple, el temps de càlcul per obtenir les propietats necesaries d’una molecula de 30 àtoms (molécula orgànica relativament petita) per realizar un pas en el temps en una simulacio, pot variar des de diverses hores fins a dies, aquest fet, es un problema, donat que els sistemas estudiats poden arribar als milions de molecules. Això ha conduït a la necessitat de tècniques d'optimització i paral·lelització per accelerar el procés d’obtenció de les propietats moleculars. Recentment, s'han proposat noves estratègies basades en l'aprenentatge computacional, com l'ús de NN.

En aquest projecte es plantegen els següents objectius per millorar l'estudi i l'aplicació de NN en simulacions físiques per a la predicció de propietats moleculars:

* **Objectiu 1**: Desenvolupar un coneixement profund dels fonaments teòrics i pràctics de les NN i la simulació de la MD.
* **Objectiu 2**: Desenvolupar una comprensió crítica dels avantatges i les limitacions de les NN en la simulació de la DM, comparant-les amb altres tècniques i abordatges existents.
* **Objectiu 3**: Estudiar els treballs més recents en el camp de la DM que apliquen NN, posant el focus en els següents aspectes: bases de dades utilitzades en els treballs recents relacionats amb la simulació de la DM i les NN; mètodes de representació dels sistemes de molècules usats en aquests treballs; arquitectures de les NN utilitzades en els treballs recents i la seva eficàcia i avaluar el rendiment i els resultats obtinguts amb els mètodes de simulació de la DM basats en les NN.
* **Objectiu 4**: Realitzar un estudi del rendiment d’un model de la toolbox de SchNetPack2 aplicat a la base de dades QM9.

**Estructura del document**

Taula 1: Frecuencia d’aparicio de elements en les molècules de la base de dades QM9. C i H apareixen a un 100% de les molècules i O i N un 85% i 62% respectivament, el F apareix només en 2% del total.

El document està organitzat amb els següents apartats:

* Exposició de les motivacions junt amb un breu context del camp de les Xarxes Neuronals (XN) aplicades a la DM i objectius del treball (punt 2).
* Revisió l’*state of the art* del camp de les XN aplicades a la DM, tant bases de dades com models, *frameworks* o *toolbox* existents (punt 3).
* Explicació de la metodologia emprada i exposició dels resultats obtinguts de les simulacions analitzats en detall (punt 4 i 5).
* Presentació de les conclusions rellevants i les possibles accions futures per millorar i ampliar aquest treball (punt 6).

# 3 State of the Art:

## 3.1 Bases de dades:

Donat el gran cost del calcul les propietats d’una molécula, en el camp de la MD, s’utilitzen les bases de dades com a biblioteques per emagatzemar les propietats de les molecules ja calculades. Un referent en aquest aspecte, es la base de dades GB17[[1]](#_7_Bibliografia:), que conté un registre de 166 mil milions de molècules orgàniques.

Figura 1: Histograma de la freqüència d'aparició per nombre d’àtoms a les molècules de la base de dades QM9. Les freqüències formen una distribució normal, amb un dèficit de molècules de més de 25 àtoms. *Gràfic d’autoria pròpia.*

## 3.2 QM9:

La base de dades QM9[[6]](#_7_Bibliografia:) és una col·lecció de molècules seleccionades de la base de dades GB17. La base de dades QM9 utilitza el format de dades Jarvis Atoms, un format estandarditzat d’emagatzematge de molècules i les seves propietats.

La base de dades QM9 disposa de aproximadament 134,000 molècules orgàniques, les quals tenen entre 6 i 29 àtoms cada una, que poden ser C, H, O, N i F. Entre les propietats emmagatzemades per cada molecula hi ha:

* **Nnumero atomic:** llista dels números atòmics **(Z)** dels atoms que composen la molecula.
* **Geometria de la molecula**: llista de les posicions **(R)** representades per coordenades cartesianes, de cada àtom de la molecula. Les coordenades **x, y, z** es representen en unitats de longitud com àngstroms (Å) o nanòmetres (nm).
* **Energia fonamental (U0):** indica l'energia total mínima d'una molècula en l'estat fonamental, és a dir l'estat d'energia més baixa que pot tenir el sistema. Aquesta energia es dona en electrons-volts (eV) o bé amb kcal/mol. Es una propietat fonamental a l’hora fer simulacions de MD.

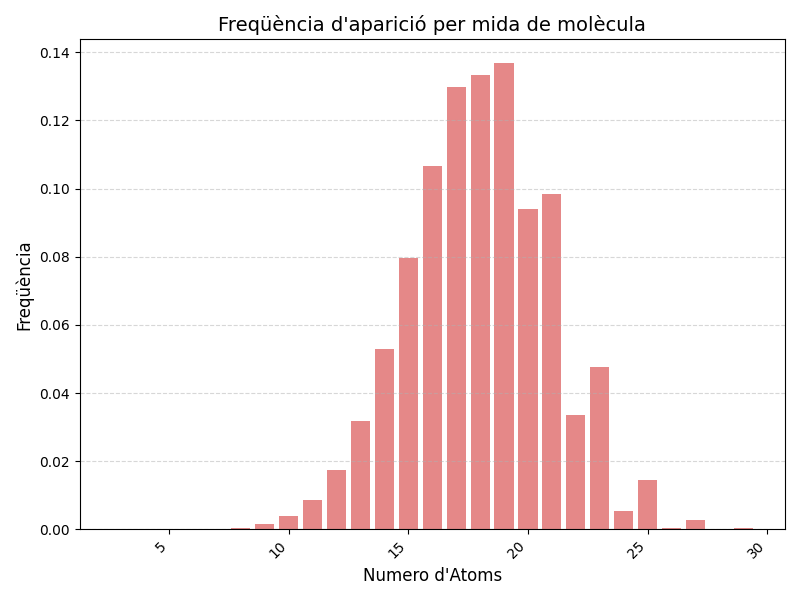
### 3.2.1 Detalls tècnics de les dades:

Per estudiar la distribució de les molècules es defineixen dos clases de referencia:

* **Elements:** els elements presents en la molecula.
* **Mida:** el numero d’atoms de de la molecula.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Element | C | H | O | N | F |
| Frequencia | 1 | 1 | 0,85 | 0,62 | 0.02 |

Com queda representat en la ***Taula 1*** la frecuencia d’aparicio de elements en les molècules de la base de dades QM9. On el carboni (C) i hidrogen (H) apareixen a un 100% de les molècules, Oxigen (O) i Nitrogen (N) en un 85% i 62% respectivament i el Fluor (F) apareix només en 2% del total.



En el cas de la mida de les molècules, trobem una distribució normal on les mides de molècula més freqüents són 17, 18 i 19 àtoms, com es pot apreciar en la ***Figura 1*** i la ***Figura Apendix 5***.

Es pot concloure que la base de dades QM9 presenta alguns aspectes que poden afectar negativament l'entrenament de models de XN. En primer lloc, hi ha una desigualtat en la distribució dels elements, amb una major prevalença de C i H i una presència limitada de F. Aquest desequilibri pot introduir biaixos en les prediccions per a molècules amb poca representació a la base de dades. A més, les dimensions (nombre d’àtoms) de les molècules mostren una distribució normal, amb dimensions més freqüents que poden limitar la capacitat del model per generalitzar bé en altres dimensions.

## 3.3 Treballs prèvis:

Com mencionat anteriorment, recentment el el camp de MD, han sorgit nous treballs fent l’us de NN, entre ells els mes utilitzats son TorchMD i SchNetPack, per aquest treball s’ha escollit utilitzar SchNetPack donat que es el mes pioner dels dos.

### 3.3.1 SchNetPack:

SchNetPack 2.0[[2]](#_7_Bibliografia:) és una *toolbox* dissenyada per al desenvolupament i desplegament de NN per a simulacions de MD. Proporciona un *framework* flexible i modular per a la construcció de models complexos que poden predir diverses propietats de molècules, com ara forces d’interacció intermoleculars, energies fonamentals, entre d’altres. Les principals aportacions de SchNetPack 2.0 són: una *data pipeline* flexible, modularitat a l’hora de construir els models de NN, implementació de PyTorch per a MD, una interfície de comandes basada en Hydra per simplificar-ne l’ús i integració de PyTorch Lightning que permet gestionar i realitzar entrenaments fàcilment.

La *data pipeline* de SchNetPack 2.0 permet processar i preparar les dades per als models de NN. Està composta per 2 components principals, una interficie per carregar i procesar les dades directament desde la base de dades i un agrupador de lots que junta les dades procesades per acelerar el proces d’entrenament.

Amb SchNetPack, es pot treballar amb models predefinits o desenvolupar nous models personalitzats. Això ho fa possible l’ús de les classes AtomisticModel i NeuralNetworkPotential.

* **AtomisticModel** és la base de SchNetPack i hereta de la classe nn.Module de PyTorch. Utilitza mòduls predefinits i personalitzats per definir arquitectures de XN, aquests mòduls poden ser des de capes per a la representació de les dades a convolucions (llista detallada dels mòduls disponibles en l’article[[2]](#_7_Bibliografia:)).
* **NeuralNetworkPotential** simplifica la creació de models MLP (Machine Learning Potentials). Aquesta subclasse aplica seqüencialment les funcions definides en AtomisticModel, amb l'afegit dels paràmetres *input\_modules* i *output\_modules*. El mètode *forward* és responsable de passar el diccionari d'inputs per les diferents funcions del model.

Figura 2: Diagrama dels mòduls del model utilitzat. A la esquerra queda lestructura general mentres que indicat amb fletxes hi ha els moduls interns per cada modul general. Les entrades i sortides dels moduls representen el resultat de les operacions a tots els valors de lot. Els moduls amb una multiplicacio implica que es repeteixen X vegades, per exemple Interaction o Atomwise.

El model utilitzat en el treball be definit per la ***Figura 2,*** la esquerra queda lestructura general del model, desde les dades prepocesades per la *data pipeline* fins la prediccio del model. Els moduls del model son els seguents:

* **Input module:** s’utilitza el modul PairwiseDistance, un modul no entrenable. PairwiseDistance pren per input les posicions de cada àtom (**R)** i la llista de veins (**V**) i calcula una llista de distancies entre veins i la emagatzema en un nou camp en el lot anomenat **Rv**, que es la distancia de cada àtom amb els seus veins**.**
* **Representation:** s’utilitza el modul SchNet[[5]](#_7_Bibliografia:), els parametres d’aquest modul es veuen modificats durant l’entrenament, SchNet pren per entrada, els numeros atomics (**Z**), la llista de veins (**V**) i la llista de distancies entre veins (**Rv**) i retorna un vector de *features* (**X**) que representa la molecula. El output d’aquest modul es unicament el vector de features (**X**). L’estructura interna del modul es basa en primer un embedding dels números atomics que despres es pasa per els 3 moduls Interaction, generant aixi l’output.
* **Interaction module:** (a la dreta de la figura) pren per input el vector de *features* (**X**) i les posicions dels àtoms (**R**) i retorna el vector de *features* modificat. L’estructura interna del modul es basa en 4 capes denses i un filtre de convolució continua[[5]](#_7_Bibliografia:), aquest filtre té com a objectiu capturar les interaccions locals entre àtoms de la molècula de manera continua.
* **Output module**: el modu Atomwise és una capa densa que realitza una transformació lineal al vector de features (**X**) per obtenir la propietat desitjada. En el cas del model utilitzat hi han 10 moduls atomwise.

Una vegada els models han estat definits i les dades han estat processades, PyTorch Lightning entra en joc per simplificar el procés d'entrenament i avaluació dels models de SchNetPack. PyTorch Lightning proporciona una interfície d'entrenament de nivell superior que gestiona automàticament tasques com la configuració de l'entrenament, la gestió dels dispositius de càlcul, el càlcul dels gradients i l'actualització dels paràmetres del model.

# 4 Metodologia:

Tenint en compte els objectius definits en lapartat 2, s’ha pres de referència la metodologia àgil utilitzada regularment en el camp del desenvolupament de software, se separarà el projecte en fases. En cada fase s’ha fet servir GitHub com a sistema de control de versions, podent establir cicles de treball de durada flexible (entre una i dues setmanes) per assegurar un seguiment adequat del progrés del projecte. Per a cada cicle, s’han predefinit objectius específics (***Figura apendix 6***) a assolir i s’han generat informes de cicle per verificar si s'han assolit tots els objectius desitjats i explicar les raons si no s'han assolit. També s’ha contemplat la possibilitat de canviar l'ordre dels cicles de treball sempre que es justifiqui adequadament o no afecti negativament a altres tasques pendents. En finalitzar cada fase s’ha redactat un informe de progrés que recull els continguts dels informes dels cicles que componen la fase. Les fases es divideixen en dos grups: les dues primeres teòriques i la tercera pràctica.

* **Fase de formació**: En aquesta fase s'ha desenvolupat un coneixement profund dels fonaments teòrics i pràctics de les XN i la simulació de DM.
* **Fase d’exploració**: En aquesta fase s’han explorat els avantatges i les limitacions de les XN en la simulació de DM, comparant-les amb altres tècniques i abordatges existents.
* **Fase d'avaluació**: En aquesta fase s’ha centrat en l’estudi i optimizacio del model definit en la ***Figura 2*** aplicat a la base de dades QM9.

### **4.1 Fase de formació:**

Per a l'assoliment del primer objectiu, s'han fet servir diverses metodologies, incloent-hi tutories amb el tutor extern Jordi Faraudo (ICMAB) per al camp de la DM, i amb el tutor acadèmic Ramon Baldrich pel coneixement de XN; complementat amb recerca pròpia a través de la lectura d'articles i altres fonts de documentació, així com l'aplicació d'assignatures relacionades com Aprenentatge Computacional (APC) per entendre els fonaments teòrics de les XN. A més, s'han realitzat reunions amb estudiants especialitzats en temes com Intel·ligència Artificial (IA) o les Matemàtiques Computacionals i Analítica de Dades (MatCAD) per obtenir una perspectiva diferent.

### **4.2 Fase d’exploració:**

Per al desenvolupament d'una comprensió crítica dels avantatges i les limitacions de les XN en la simulació de DM, s'han dut a terme tutories amb el tutor extern Jordi Faraudo i lectures d'articles científics i treballs relacionats que comparaven les XN amb altres tècniques i abordatges existents.

### **4.3 Fase d’avaluació:**

Per l’estudi del model s’ha establert que la propietat objectiu del model es l’energia fonamental (**U0**). La *pipeline* utilitzada per dur a terme l'entrenament, validació i obtenció de mètriques està representada en la ***Figura apèndix 1***, al ser una fase pràctica els resultats d’aquesta es trobaran en l’apartat de Resultats (punt 5). Aquesta fase es divideix en 3 subfases:

* **Cerca d’hiperparàmetres**: per tal d’optimitzar els models proporcionats per la *toolbox* SchNetPack2 usant la base de dades QM9. Aquesta subfase s’ha dut a terme amb l’ajuda de Weights and Biases[[4]](#_7_Bibliografia:)(a partir d’ara WandB), una plataforma de monitoratge online que permet la fàcil visualització d’informació rellevant respecte l’entrenament, validació i resultats de models d’aprenentatge computacional, com ara corbes d’aprenentatge o consum de recursos. Especificament s’ha fet servir la funcionalitat ***sweep*** de WandB per a fer cerques d’hiperparàmetres. La primera etapa de l’optimització s’ha centrat en la cerca d’hiperparàmetres estrictament computacionals com el *learning rate*, el nombre d’*epochs* i el *batch size*, amb l'objectiu de millorar el rendiment dels nostres models. Aquesta fase és fonamental, ja que els hiperparàmetres anomenats tenen un impacte directe en el procés d'entrenament i poden afectar significativament el rendiment final del model. En la segona etapa, es canvia el focus de la cerca als hiperparàmetres que s’ocupen de modelitzar les propietats físiques de les mol·lècules.
* **Anàlisi profunda dels resultats**: s’han dut a terme una sèrie d’experiments amb l’objectiu de millorar els resultats dels models. Aquests experiments tenen tres fases: definició de la hipótesi, realització de tests i extracció de conclusions. S’han dut a terme 3 tests principals en els quals s’entra en més detall en l’apartat de Resultats (punt 5).
* **Conclusions**: per a concluir la fase s’ha entrenat un model tenint en compte les conclusions extretes dels anteriors experiments amb l’objectiu de comparar els resultats d’aquest model amb els resultats exposats en el *paper* de SchNetPack 2[[2]](#_8_Bibliografia:).

# 5 Resultats:

Abans de profunditzar en els resultats, és essencial establir alguns conceptes fonamentals. Tot i que en la *toolbox* SchNetPack2 s’hi tracten molts tipus de simulacions de DM, en aquest apartat es contemplaran exclusivament les simulacions de DM que busquen calcular l'energia de l'estat fonamental (U0) d'una molècula. L'atribut U0 indica l'energia total mínima d'una molècula en l'estat fonamental, és a dir, l'estat d'energia més baixa que pot tenir el sistema. Aquesta energia es dona en electrons-volts (eV) o bé amb kcal/mol.

Per a realitzar una simulació de DM d’aquest tipus, és necessari definir un conjunt de condicions inicials, que inclouen les posicions (R) i el número atòmic (Z) dels àtoms del sistema. Aquestes condicions inicials se solen obtenir a partir d'una configuració estructural coneguda del sistema, com ara una estructura cristal·lina, o bé a partir d'una configuració aleatòria que segueixi les restriccions imposades per a les interaccions del sistema.

Les posicions (R) són comunament representades per coordenades cartesianes, que són les més senzilles, i consisteixen en la definició de la posició de cada àtom en un sistema de coordenades tridimensional, que es representa generalment en unitats de longitud com àngstroms (Å) o nanòmetres (nm). Així, una molècula es pot representar per la seva posició en l'espai, definida per les coordenades x, y i z de cada àtom que la conforma.

Tradicionalment, el càlcul de l’U0 es fa resolent les equacions de Schrödinger per aconseguir la funció d'ona electrònica i la seva energia associada[9]. Aquests mètodes són computacionalment costosos ja que han de tenir en compte la naturalesa quàntica dels electrons i el moviment dels nuclis atòmics. Això implica el càlcul de les funcions d'ona de tots els electrons de la molècula i, per tant, requereixen un gran nombre de càlculs matemàtics i computacionals. Com a resultat, aquests mètodes són limitats per la seva capacitat de resoldre problemes en molècules grans i complexes.

### **5.1 El Model:**

Pel que fa al model usat, s’ha seguit una guia subministrada en el GitHub de SchNetPack2, s’ha escollit no modificar massa el model estàndard per poder comparar els resultats amb els que asseguren haver obtengut els creadors de la *toolbox*. Dit això, en la ***Figura 4*** queda representat el model, a l’esquerra de la figura hi ha els mòduls pels quals passa un lot de dades (data batch) durant un pas de l’entrenament, la estructura es la següent:

* **Mòdul de representació**: ***SchNet*** a partir de les posicions (R) i els nombres atòmics (Z) i la llista de veïns (*neighbour list*) genera una llista de *features* que representa la molècula, és un mòdul entrenable i per tant, la representació es va millorant durant l’entrenament.
* **Mòdul predictor (input module)**: ***pairwiseDistancies.*** Es el mòdul responsable d’obtenir les prediccions de l’atribut objectiu a partir de les representacions realitzades pel mòdul de representació.
* **Mòdul de postprocessament:** s’ocupa de desfer les transformacions prèvies a l’entrenament**.**

A la dreta de la figura hi ha els mòduls interns del mòdul SchNet, siguen aquests, per més detalls de les capes internes d’aquest mòdul es pot trobar informació en la documentació original del treball SchNetPack v0[[5]](#_8_Bibliografia:), on s’explica el perquè de cada capa del model de representació.

### **5.2 Cerca d’hiperparàmetres:**

Per aquest apartat s’ha decidit deixar els detalls fora de l’informe final i es troben en la secció d’apèndix A2, es proporciona una extensa explicació respecte al procés realitzat i les conclusions obtingudes. Seguidament es dona un resum d’aquest apèndix.

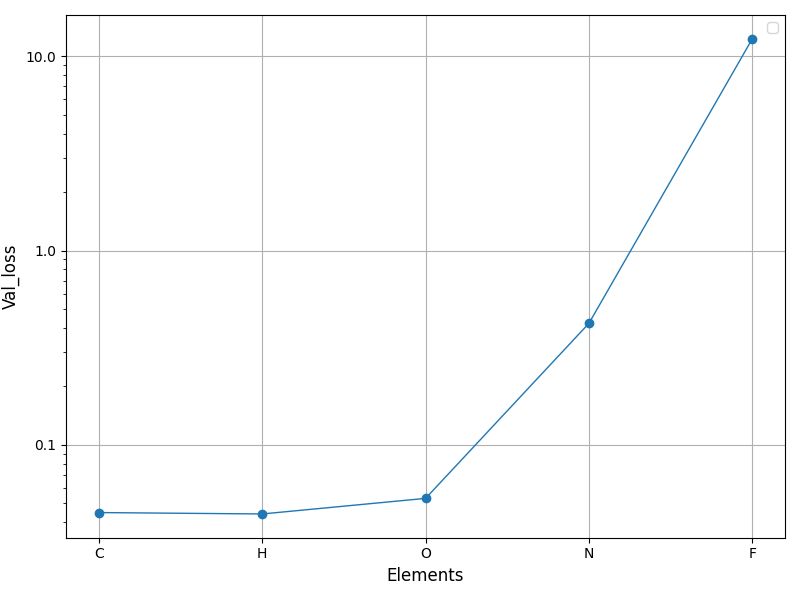
### 5.2.1 Hiperparàmetres Computacionals:

S’ha centrat en tres variables clau: *batch\_size*, *epochs* i *lr*. El nombre d'èpoques (*epochs*) determina quantes vegades el model passa per tot el conjunt de dades d'entrenament, mentre que la mida del lot (*batch size*) especifica la quantitat de mostres que es processen en cada pas d'actualització dels pesos del model i per últim la taxa d'aprenentatge *learning rate* (lr) que indica la quantitat d'ajustament que es fa als pesos del model durant el procés d'entrenament. L’entrenament s’ha realitzat sobre el 80% de les dades mentres que la validacio pren el 20% restant.

Al final del procés de la cerca d’hiperparàmetres s’ha arribat a la següent conclusió: donada la situació que es compta amb temps limitat per fer l’estudi, s’escullen els valors que proporcionen major l’equilibri entre bons resultats i temps d’entrenament baixos.

* **Epochs**: 22
* **Batch\_size**: 512
* **Lr**: 0.001

*5.2.1 Hiperparàmetres Interns de SchNetPack2:*

S’ha centrat en 3 paràmetres interns, *dataCutoff*, el nombre de veïns que té en compte a l’hora d’aplicar les transformacions prèvies a l’entrenament, *trainingCutoff*, el valor de Cutoff de la funció gaussiana utilitzada per representar les distàncies i n\_atomBasis, el nombre de *features* usat per representar la molècula.

Al final de la cerca s’ha arribat a la conclusió: els paràmetres escollits no han representat tanta variació en els resultats finals com s’esperava, tot i això s’han obtingut els següents paràmetres com als òptims:

* **DataCutoff**: 4
* **TrainingCutoff**: 5
* **N\_atom\_basis**: 38

Figura 6: Model entrenat amb la base de dades QM9 default i validat amb els *dataframes* separats per elements. L’escala de l’eix de les ordenades és logarítmic per visualitzar millor els resultats. Es pot veure la clara correlació inversa entre la freqüència d’aparició en la base de dades dels elements i la *validation loss*. **Confirma la hipòtesi 2.**

### **5.3 Anàlisi profunda dels resultats**:

Els resultats mostrats en aquest apartat s’han obtingut entrenant el model amb la base de dades QM9 default o modificada (indicat en cada resultat), i després validat el model amb la base de dades QM9 separada en *dataframes* per nombre d’àtoms de les molècules o separada per elements presents en les molècules. En el cas del nombre d’àtoms de la molècula s’ha decidit descartar els tamanys amb menys representació que un 0,1%, és a dir, s’utilitzaran per a la validació molècules d’entre 10 i 24 àtoms.

Vistos els resultats obtinguts per molècules amb més de 19 àtoms i donada la major complexitat de calcular l’energia per aquestes molècules, s’ha establert com a objectiu: **Millorar prediccions per molècules amb més de 19 àtoms.**

*5.3.1 Baseline:*

Abans d’entrar als tests individuals s’ha fet una *baseline* amb el model sense cap modificació per poder comparar els resultats dels experiments.

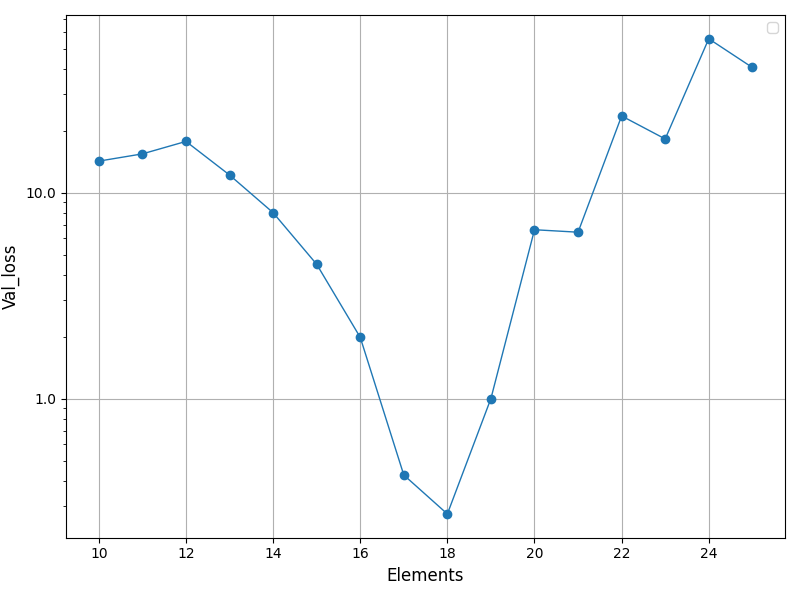
* **Hipòtesi 1:** com s’ha mencionat en l’apartat 3.2.1 es teoritza que mides de molècules poc representades en la base de dades donaran resultats pobres.
*  **Hipòtesi 2:** seguint la hipòtesi 1, elements poc representats en la base de dades donaran resultats pobres

Figura 5: Model entrenat amb la base de dades QM9 default i validat amb els *dataframes* separats per nombre d’àtoms (entre 10 i 25). L’escala de l’eix de les ordenades és logarítmic per visualitzar millor els resultats. Es pot veure la clara correlació inversa entre la freqüència d’aparició en la base de dades i la *validation loss*. **Confirma la hipòtesi 1**

Es conclou que amb **la hipòtesi 1 i la hipòtesi 2 confirmades** hi ha molt marge de millora respecte al model base que queda completament esbiaixat a favor dels elements i mides de molècules més representats.

*5.3.2 Database Subsampling:*

La resposta més utilitzada en casos de bases de dades desequilibrades és el *subsampling/oversampling*, en el camp de la DM no és possible dur a terme *oversampling* així que s’han fet els tests amb *subsampling*.

* **Hipòtesi 3**: forçant que totes les mides de molècula tinguin la mateixa freqüència d’aparició s’ equilibraran els resultats.

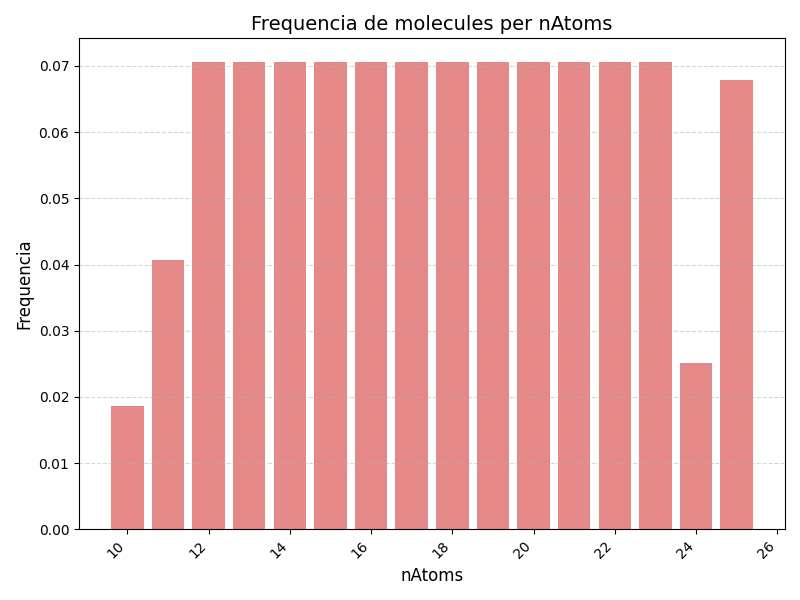


Figura 7: Histograma de la freqüència d'aparició per nombre d’àtoms en les molècules de la base de dades amb *threshold* de 2.000 molècules representades.

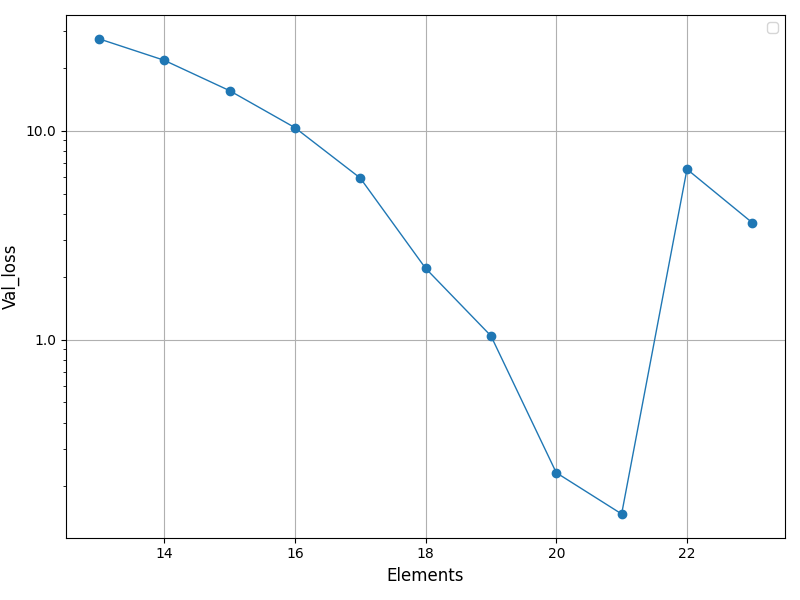
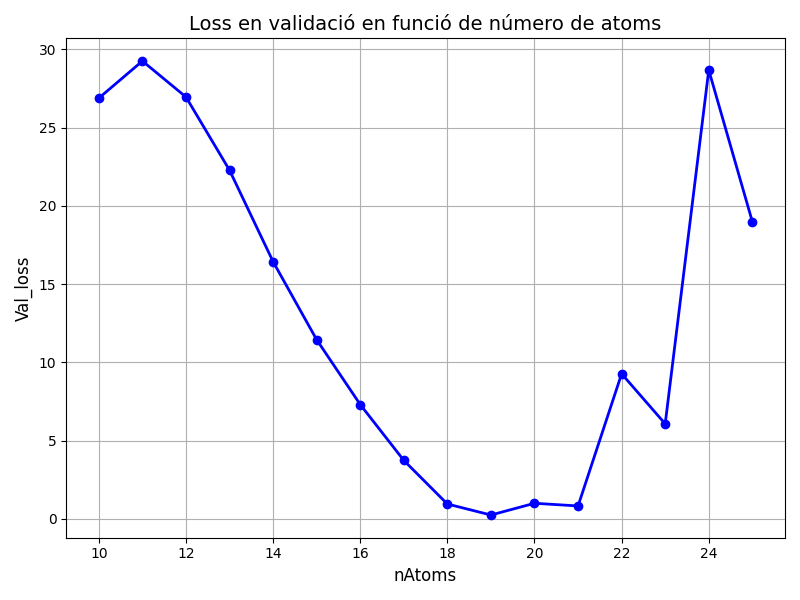
S’han fet 3 tests cada un amb un *threshold* de representació per nombre d’àtoms 2.000, 4.000 i 6.000 molècules.

Figura 8: Models entrenats amb les bases de dades QM9 *subsampled* i validat amb els *dataframes* separats per nombre d’àtoms (entre 10 i 25). L’escala de l’eix de les ordenades és logarítmic per visualitzar millor els resultats. Es pot visualitzar l’efecte del *subsampling* en els resultats en forma d’una millora per molècules més grans en el cas més allunyat de la base de dades *default* (*Threshold* de 2000).

Figura 10: Models entrenats amb les bases de dades QM9 splited per molècules amb més de 17 àtoms i validat amb els *dataframes* separats per nombre d’àtoms.

Donat que l’efecte en els resultats no és l’esperat **la hipòtesi 3 queda descartada** el que porta a pensar que el principal causant del desequilibri en els resultats no és la representació en la base de dades.

*5.3.3 Database Spliting:*

A vista del poc impacte aconseguit amb el *subsampling* s’ha considerat necessari utilitzar altres mètodes. El principi del mètode de *data base Spliting* és senzill, dividir la base de dades, entrenar amb una de les meitats i validar amb el total.

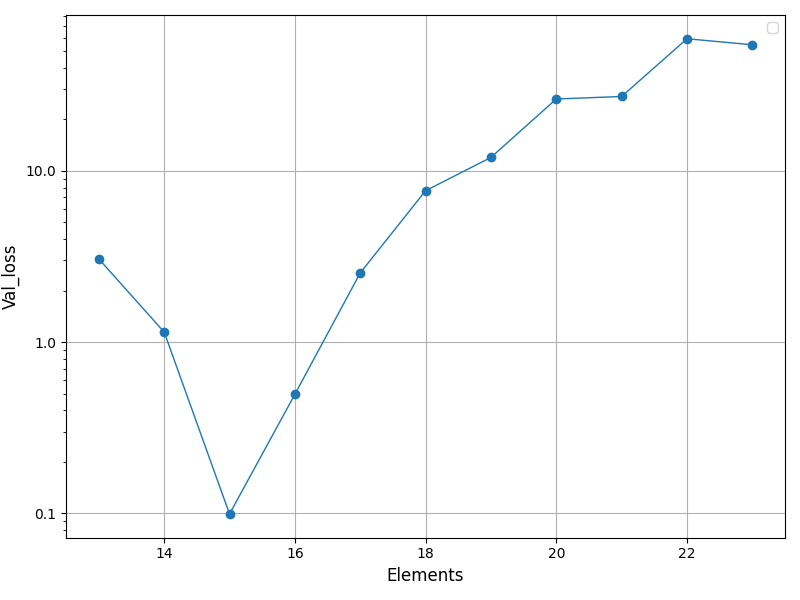
* **Hipòtesi 4**: els resultats obtinguts a partir de realitzar l’entrenament amb una meitat seran millors per a les molècules que pertanyin a aquesta meïtat. Amb això quedarà demostrada la falta de capacitat de generalització del model.

Figura 9: Models entrenats amb les bases de dades QM9 splited per molècules amb menys de 18 àtoms i validat amb els *dataframes* separats per nombre d’àtoms (entre 13 i 23).

Tant en la ***Figura 9*** com en la ***10*** s’aprecia el fenomen descrit en **la hipòtesi 4**, per tant, queda **confirmada**. A part de les implicacions negatives mencionades en la hipòtesi que té aquest resultat també n’hi ha de positives, es redueix significativament l’error en molècules del rang d’àtoms amb el que s’ha entrenat el model, informació molt rellevant en cas de voler entrenar un model per a una mida de molècules específic.

*5.3.4 Loss Rate Functions:*

Un dels altres possibles mètodes per a compensar una base de dades desequilibrada és canviar l’impacte que tenen les dades en la funció del càlcul del *loss rate* durant l’entrenament que de base s’utilitza *mean square error* (MSE). S’han definit 2 possibles funcions de *loss* modificades, siguent **v** el factor pel qual es multiplica el valor de *loss*, **s** el nombre d’àtoms de la molècula, **f** la freqüència d’aparició en la base de dades i **r** el *rate* d’impacte.

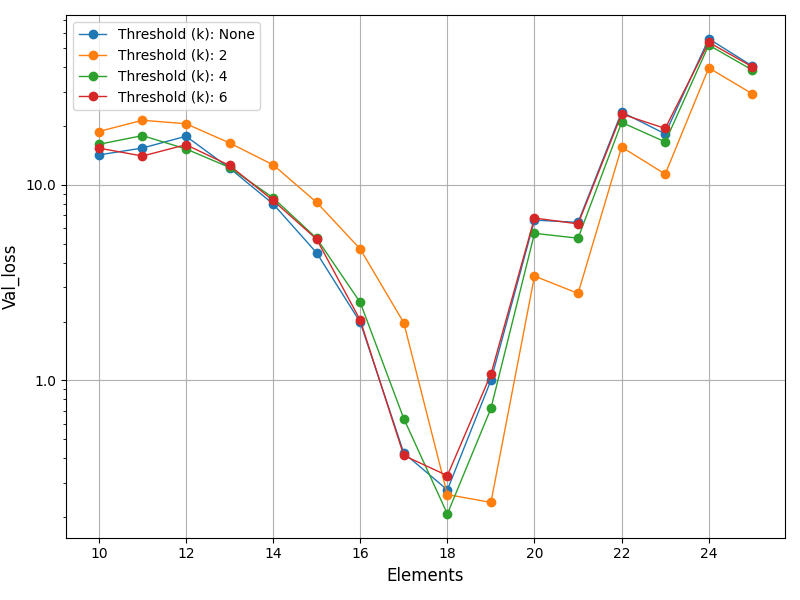
* Un reloj de aguja

  Descripción generada automáticamente con confianza media**Size based**: amb l’objectiu de potenciar les molècules més grans:
* **Un reloj de aguja

  Descripción generada automáticamente con confianza bajaSize and Frequency based**: busca un equilibri entre la mida de les molècules i la freqüència d’aparició:

Per a visualitzar millor s’han realitzat la ***Figura*** ***Apèndix 2*** on queda representat l’evolucio del *factor* (v) per determinats números de *rate* (r).

* **Hipòtesi 5**: valors de **r** grans donaran més importància a l’error en les molècules més grans, millorant així les prediccions per molècules grans.



Una vegada definides les funcions s’ha dut a terme un test entrenant amb la base de dades amb nombre d’àtoms entre 13 i 23 per obtenir els millors *rates* per a les funcions.

Gráfico, Gráfico de líneas

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamenteS’extreu de les ***Figures 10*** i ***11*** que els *rates* més rellevants per minimitzar la *loss* en molècules amb nombre d’àtoms elevat són per ambdues funcions, el *rate* 10. Fet que té sentit considerant la naturalesa de les funcions.

Figura 12: En taronja el model Default i en blau fosc el model Custom. Especificacions dels models explicades al punt 5.3.5. Mentre que per valors inferiors a 19 el model Default obté resultats superiors, el model Custom millora significativament el rendiment del Default per molècules amb és de 19 àtoms.

Figura 11: Models entrenats amb les bases de dades QM9 splited amb nombre d’àtoms entre (13 i 23) utilitzant la modificació **Size and Frequency Based** i validat amb els *dataframes* separats per nombre d’àtoms. Per referència la línea de *rate* = 0 representa la funció *default* de *loss* del model.

Figura 11: Models entrenats amb les bases de dades QM9 splited amb nombre d’àtoms (entre 13 i 23) utilitzant la modificació **Size Based** i validat amb els *dataframes* separats per nombre d’àtoms. Per referència la línea de *rate* = 0 representa la funció *default* de *loss* del model.

Amb això **s’ha confirmat la hipòtesi 5,** donat que els millors resultats per molècules grans s’obté amb els valors de *rate* grans (en aquest cas 10)

*5.3.5 Experiment final:*

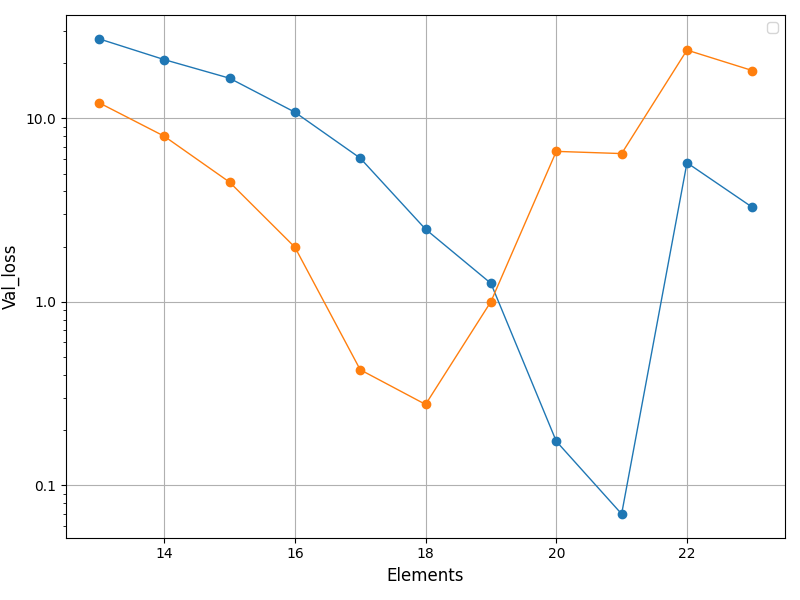
Per mesurar la millora que representen els canvis proposats respecte el model base de SchNetPack 2, s’han establert els següents criteris:

* **MSE general:** mètrica MSE del model, validant amb el 20% de les entrades de la base de dades (seleccionades aleatòriament).

Taula 1: Comparació de les mètriques MSE i *Runtime* dels models Default i Custom.

* **MSE per nombre d’àtoms**: s’ha mesurat la mètrica MSE que obté el model per cada mida de molècula amb representació superior a l’1% en la base de dades.
* **Estabilitat en les corbes d’aprenentatge:** s’ha implementat una funcionalitat extra per poder emmagatzemar la mitjana dels 5 valors mínims i màxims de *loss* durant cada pas de l’entrenament. Amb això s’espera poder representar millor el *rang* d’error durant l’entrenament.
* **Runtime:** temps d’execució total tenint en compte l’entrenament i la validació amb els *dataframes* per nombre d’àtoms.

Una vegada establerts els criteris, s’han definit els paràmetres pels models:

* **Model Default**: s’ha entrenat amb els paràmetres del model base subministrat utilitzant la base de dades sencera.
* **Model Custom**: s’ha entrenat amb la base de dades QM9 splited per molècules amb més de 17 àtoms. Els hiperparàmetres escollits són:
* **Batch\_size**: 128
* **Epochs**: 30
* **Lr**: 0.0001
* **DataCutoff**: 4
* **TrainingCutoff**: 5
* **N\_atom\_Basis**: 38
* **Loss Size based Rate**: 10

|  |  |  |
| --- | --- | --- |
| Model | General MSE  (EV) | Runtime  Train + Val |
| Default | 0.0643 | 1h 46m |
| Custom | 0.0925 | 1h 28m |

Llista de resultats de les mètriques usades per a la comparació entre el model Default i el model Custom:

* **MSE general**: en la ***Taula 1*** es pot apreciar un millor resultat pel model Default. Aquest fet prové de les bases de dades emprades per entrenar els models. En el cas del model Default s’ha entrenat amb la base de dades sensera donant major importància a les mides de molècules amb més representación, i per tant, a l’hora de calcular la métrica MSE s’obté menor error. En el cas del model Custom, s’ha entrenat amb la base de dades *splited* per molècules amb més de 17 àtoms, donant importància exclusivamente a les molècules de mides superiors a 17 àtoms, menys presents en la base de dades, que per tant, contribueixen menys en la mètrica MSE.
* **Runtime**: com es veu en la ***Taula 1***, el model Custom tarda 18 minuts menys en realitzar el test sencer. Aquesta diferència la provoca única i exclusivament la mida de la base de dades amb què s’ha entrenat el model: més petita que la base de dades QM9 sencera utilitzada per entrenar el model Default.
* **MSE per nombre d’àtoms**: com s’ha explicat en la ***Figura 12*** s’obté el resultat esperat, el model Default obté millors resultats en valors inferiors a 19 i el model Custom millora significativament el rendiment del Default per molècules més grans que 19 àtoms. Aquest fenomen ve donat per dos factors principals:

- Entrenament amb la base de dades *splited* per molècules amb més de 17 àtoms (explicat en l’apartat 5.3.3).

- Ús de la funció de *loss* modificada **Size based Rate** que com s’ha mencionat anteriorment (apartat 5.3.4) prioritza molècules amb nombre d’àtoms elevat**.**

* **Estabilitat en les corbes d’aprenentatge**: com es veu en les ***Figures*** ***Apèndix 3*** i ***4***, la corba d’aprenentatge del model Default arriba a la convergència més ràpid que en el cas de la del model Custom. També es pot apreciar que la diferència entre mínims i màxims de *train loss* durant l’entrenament és menor en el model Custom.

# 6 Conclusió:

..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... .......... ...... ........ ............ ........ ...... ..... ...... ..... ..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... ........ ...... .

# 7 Agraiments:

..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... .......... ...... ........ ............ ........ ...... ..... ...... ..... ..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... .... .......... ...... ........ ...... ..... ...... ..... ..... .... ........ ...... .

# 8 Bibliografia:

1. Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond, [**Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17**](https://pubs.acs.org/doi/10.1021/ci300415d)*,* ***Journal of Chemical Information and Modeling*** *2012 52 (11), 2864-2875*
2. K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller, [**SchNetPack: A Deep Learning Toolbox For Atomistic Systems**](https://pubs.acs.org/doi/10.1021/acs.jctc.8b00908)**, *Journal of chemical theory and computation***, 2019, 15, 448-455.
3. Stefan Doerr, Maciej Majewski, AdriàPérez, Andreas Krämer, Cecilia Clementi, Frank Noe,Toni Giorgino, and Gianni De Fabritiis, [**TorchMD: A Deep Learning Framework for Molecular Simulations**](https://pubs.acs.org/doi/full/10.1021/acs.jctc.0c01343), ***Journal of chemical theory and computation***, 2021, 17, 2355−2363.
4. [**Weights & Biases Documentation**](https://docs.wandb.ai/), (accedit el: 5/5/2023).
5. Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, Klaus-Robert Müller, [**SchNet: A continuous-filter convolutional neural network for modeling quantum interactions**](https://arxiv.org/abs/1706.08566), ***arXiv***: 1706.08566v5
6. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, [**Quantum chemistry structures and properties of 134 kilo molècules**](https://www.nature.com/articles/sdata201422), ***Scientific Data 1***, 140022, 2014.
7. Choudhary, K. et al. [**The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design.**](https://www.nature.com/articles/s41524-020-00440-1) npj Computational Materials, 6(1), 1-13 (2020).

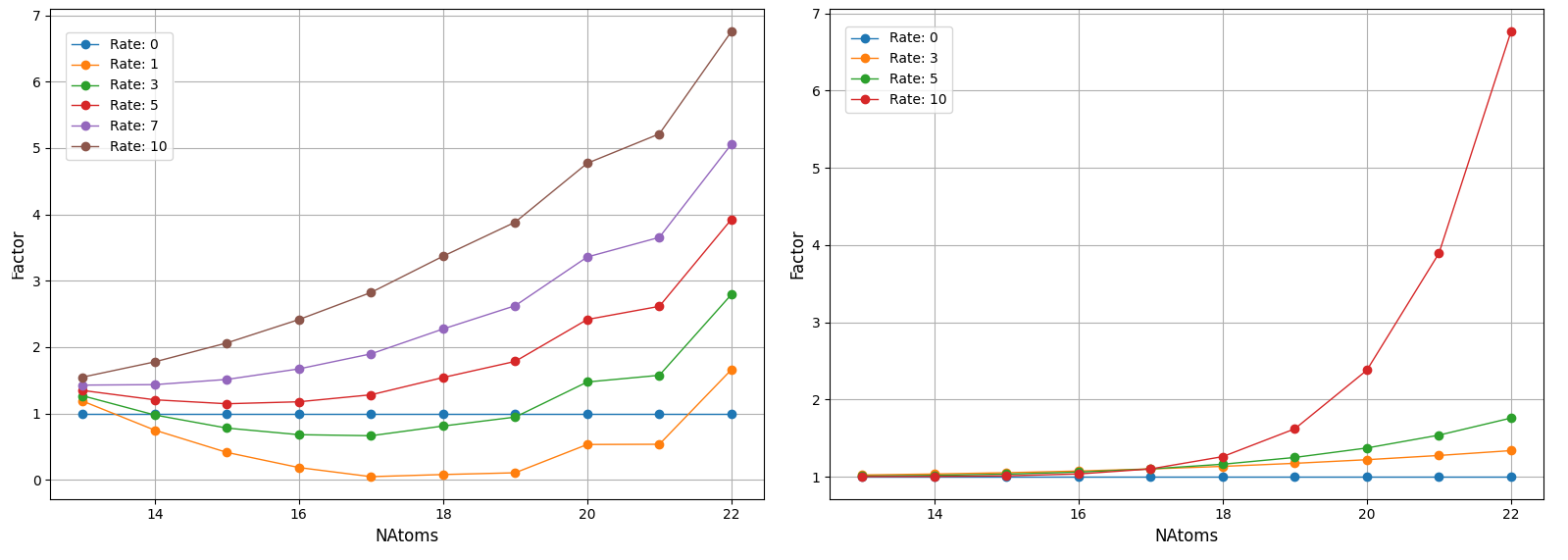
**APÈNDIX**

**Diagrama

Descripción generada automáticamenteA1. Figures:**

Figura Apendix 2: Grafic de l’evolucio del factor en relacio al rate per la funcio **Size and Frequency Based** (a la Esquerra) i **Size Based** (a la dreta), en d’abscisses el nombre d’elements de les molècules i en l’eix de les ordenades el factor multiplicador a la loss. Per referencia la línea de rate = 0 representa la funcio default de loss del model

Figura Apendix 1: Diagrama de flux de la pipeline utilitzada per l'entrenament, validació i obtenció de mètriques utilitzada durant tots els experiments del treball.

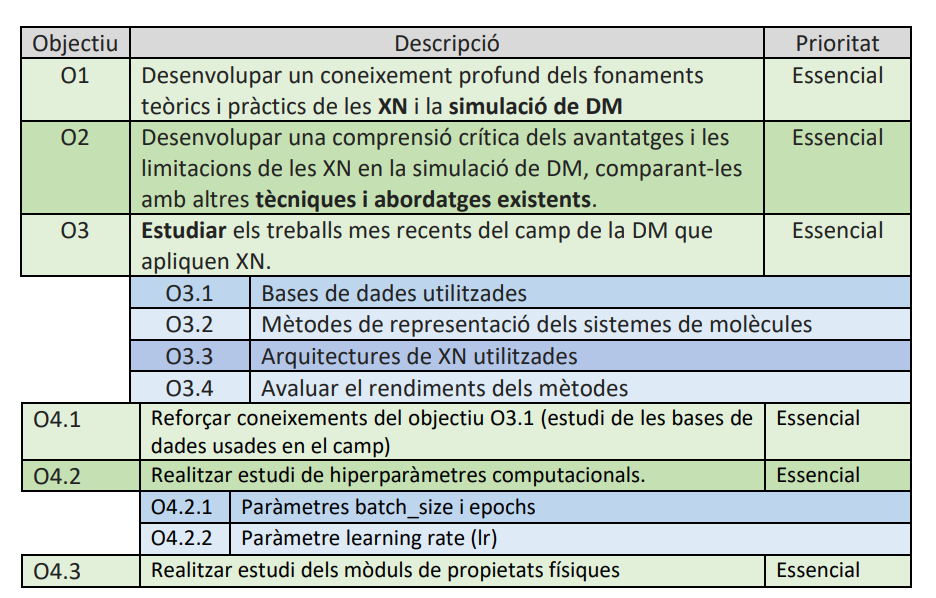
****

**Interfaz de usuario gráfica

Descripción generada automáticamenteA2. Secció d’apèndix**

Figura Apendix 4: A l’esquerra grafic de l’evolucio de la métrica train loss en relacio als trainer steps, a la dreta grafic amb l’eix de les ordenades en escala logarítmica. de l’evolucio de la métrica train loss, max train loss i min train loss en relacio als trainer steps. Grafics obtinguts durant l’entrenament del model custom. Denotar que els valors de train loss queden efectats per la funcio Size based Rate (explicada en l’apartat 5.3.4) i pertant son no comparables amb l’escala del model base.

Figura Apendix 3: A l’esquerra grafic de l’evolucio de la métrica train loss en relacio als trainer steps, a la dreta grafic amb l’eix de les ordenades en escala logarítmica. de l’evolucio de la métrica train loss, max train loss i min train loss en relacio als trainer steps. Grafics obtinguts durant l’entrenament del model default.

Aplicación, Tabla

Descripción generada automáticamente

Figura Apendix 6: Taula dels objectius definits per el treball.

Figura Apendix 5: *Heatmap* de la distribució de les entrades de la base de dades QM9. En l’eix d’abscisses el nombre d’elements de les molècules i per a cada fila de l’*heatmap* la distribució de les molècules amb l’element corresponent. S’aprecia que les molècules amb Fluor (F) present no segueixen la tendència general. *Gràfic d’autoria pròpia.*