What neural networks can do on molecular properties prediction?

Joan Tibau Terma

Resum — La simulació de dinàmica molecular és una tècnica computacional amplament utilitzada per a l'estudi del comportament de sistemes moleculars. Té aplicacions en la indústria farmacèutica, biologia, enginyeria de materials, física i química. El seu principal obstacle és la necessitat de càlcul ràpid de propietats moleculars com ara forces i energies. Recentment, s'han proposat noves estratègies per simulacions de dinàmica molecular basades en l'ús de xarxes neuronals. Aquest projecte té com a objectiu l'estudi i l'aplicació de xarxes neuronals per a la predicció de propietats moleculars. Es realitza un estudi de la base de dades QM9. S'usa un model de la toolbox SchNetPack2 aplicat a la base de dades QM9. S'elabora una anàlisi detallada de l'impacte que té en la qualitat de les prediccions el nombre d'àtoms per molècula i els elements que la componen i es suggereixen canvis per millorar els resultats. S'han extret les conclusions que cal millorar la base de dades QM9 i que s'ha d'establir un estàndard per exposar les mètriques dels models de xarxes neuronals aplicats a la dinàmica molecular

Paraules clau — Analítica de Dades, xarxes neuronals, dinàmica molecular, base de dades QM9, SchNetPack2, anàlisi de resultats, optimització.

Abstract -- Molecular dynamics simulation is a widely used computational technique for studying the behavior of molecular systems. It has applications in the pharmaceutical industry, biology, materials engineering, physics, and chemistry. Its main bottleneck is the need for fast calculation of molecular properties like forces and energies. Recently, new strategies based on the use of neural networks have been proposed for molecular dynamics simulations. This project aims to study and apply neural networks for predicting molecular properties. An analysis of the QM9 database is performed. The SchNetPack2 toolbox model is used on the QM9 database. A detailed analysis is conducted to assess the impact of the number of atoms per molecule and the elements composing it on the quality of predictions, and changes are proposed to improve the results. The conclusions drawn are that the QM9 database needs improvement and an established standard is required to expose metrics for neural network models applied to molecular dynamics.

Keywords -- Data analytics, neural networks, molecular dynamics, QM9 database, SchNetPack2, results analysis, optimization.

1 Introducció:

La 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.

2 Motivació i objectius:

Un dels principals problemes 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 necessàries d'una molècula de 30 àtoms (molècula orgànica relativament petita) per realitzar un pas en el temps en una simulació, pot variar des de diverses hores fins a dies, aquest fet, és un problema ja que els sistemes estudiats poden arribar als milions de molècules. 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 xarxes neuronals (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 MD, comparant-les amb altres tècniques i abordatges existents.
- **Objectiu 3**: Estudiar els treballs més recents en el camp de la MD que apliquen NN, posant el focus en els següents aspectes: bases de dades utilitzades en els treballs recents relacionats amb la simulació de la MD 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 MD basats en les NN.
- **Objectiu 4**: Realitzar un estudi del rendiment d'un

- *E-mail de contacte: jotite19@gmail.com*
- *Mención realitzada: Computació*
- *Treball tutoritzat per: Ramon Baldrich*
- *Tutoria externa per: Jordi Faraudo*
- *Curs 2022/23*

model de la toolbox de SchNetPack2 aplicat a la base de dades QM9.

Taula 1: Freqüència d'aparició d'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 un 2% del total.

Estructura del document

El document està organitzat en els següents apartats:

- Exposició de les motivacions junt amb un breu context del camp de les NN aplicades a la MD i objectius del treball (punt 2).
- Revisió l'*state of the art* del camp de les NN aplicades a la MD, 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 càlcul de les propietats d'una molècula, en el camp de la MD, s'utilitzen les bases de dades com a biblioteques per emmagatzemar les propietats de les molècules ja calculades.

3.2 QM9:

La base de dades QM9[6] és una col·lecció de molècules seleccionades del diccionari de molècules GDB17[1] que conté un registre de 166 mil milions de molècules orgàniques. La base de dades QM9 utilitza el format de dades Jarvis Atoms[7], un format estandarditzat d'emmagatzematge de molècules i les seves propietats (obtingudes amb càlculs de mecànica quántica).

La base de dades QM9 disposa d'aproximadament 134.000 molècules orgàniques, les quals tenen entre 6 i 29 àtoms cadascuna, que poden ser C, H, O, N i F. Entre les propietats emmagatzemades per cada molècula hi ha:

- **Número atòmic:** llista dels nombres atòmics **(Z)** dels àtoms que componen la molècula.
- **Geometria de la molècula**: llista de les posicions **(R)** representades per coordenades cartesianes, de cada àtom de la molècula. 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.

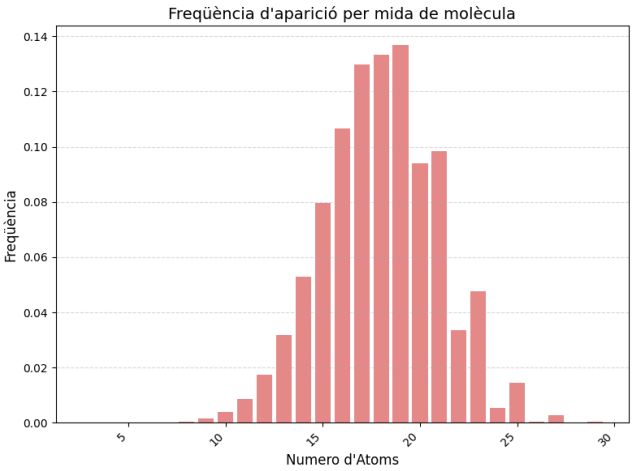
3.2.1 Detalls tècnics de les dades:

Per estudiar la distribució de les molècules es defineixen dues classes de referència:

- **Elements:** els elements presents en la molècula.
- **Mida:** el nombre d'àtoms de la molecula.

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

A la ***Taula 1*** es representa la freqüència d'aparició d'elements en les molècules de la base de dades QM9: el Carboni (C) i l'Hidrogen (H) apareixen a un 100% de les molècules, l'Oxigen (O) i el 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 Apèndix 5***.

Es pot concloure que la base de dades QM9 presenta

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.*

alguns aspectes que poden afectar negativament l'entrenament de models de NN. 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 previs:

Com s'ha mencionat abans, recentment en el camp de MD, han sorgit nous treballs fent l'ús de les NN, entre ells els més utilitzats són TorchMD i SchNetPack, per aquest treball s'ha escollit utilitzar SchNetPack donat que és el més pioner dels dos.

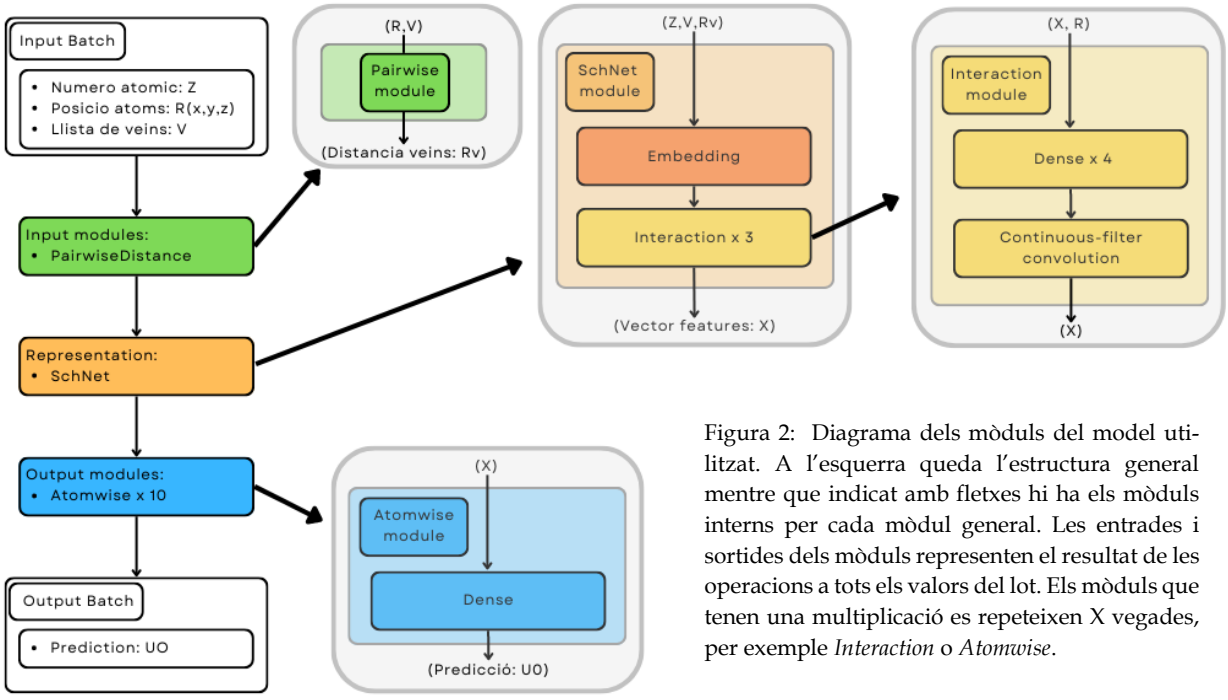


Figura 2: Diagrama dels mòduls del model utilitzat. A l'esquerra queda l'estructura general mentre que indicat amb fletxes hi ha els mòduls interns per cada mòdul general. Les entrades i sortides dels mòduls representen el resultat de les operacions a tots els valors del lot. Els mòduls que tenen una multiplicació es repeteixen X vegades, per exemple *Interaction* o *Atomwise*.

3.3.1 SchNetPack:

SchNetPack 2.0[2] 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 interfície per carregar i processar les dades directament des de la base de dades i un agrupador de lots que junta les dades processades per accelerar el procés 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 NN, 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]).
- **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.

El model usat en el treball ve definit per la ***Figura 2***, a l'esquerra queda l'estructura general del model, des de les dades prepocessades per la *data pipeline* fins la predicció del model. Els mòduls del model són els segents:

- **Input module:** s'utilitza el mòdul *PairwiseDistance*, un mòdul no entrenable. *PairwiseDistance* pren per *input* les posicions de cada àtom (**R**) i la llista de veïns (**V**) i calcula una llista de distàncies entre veïns i l'emmagatzema en un nou camp en el lot anomenat **Rv**, que és la distància de cada àtom amb els seus veïns.
- **Representation:** s'utilitza el mòdul SchNet[5], els paràmetres d'aquest mòdul es veuen modificats durant l'entrenament, SchNet pren per entrada, els números atòmics (**Z**), la llista de veïns (**V**) i la llista de distàncies entre veïns (**Rv**) i retorna un vector de *features* (**X**) que representa la molècula. L'*output* d'aquest mòdul és únicament el vector de *features* (**X**). L'estructura interna del mòdul es basa primer en un *embedding* dels números atòmics que després es passa per als 3 mòduls *Interaction*, generant així l'*output*.
- **Interaction module:** (a la dreta de la ***Figura 2***) 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 mòdul es basa en 4 capes denses i un filtre de convolució contínua[5], aquest filtre té com a objectiu capturar les interaccions locals entre àtoms de la molècula de manera contínua.
- **Output module**: el mòdul *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 ha 10 mòduls *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 el ***Punt 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 a assolir (***Figura Apèndix 6***) 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 NN i la simulació de MD.

- **Fase d'exploració**: En aquesta fase s'han explorat els avantatges i les limitacions de les NN en la simulació de DM, comparant-les amb altres tècniques i abordatges existents.

- **Fase d'avaluació**: Aquesta fase s'ha centrat en l'estudi i optimització 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 MD, i amb el tutor acadèmic Ramon Baldrich pel coneixement de NN; 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 NN. A més, s'han fet reunions amb estudiants especialitzats en temes com Intel·ligència Artificial (AI) 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 NN en la simulació de MD, s'han dut a terme tutories amb el tutor extern Jordi Faraudo i lectures d'articles científics i treballs relacionats que comparaven les NN 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 és 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***, en 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] (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, s'ha canviat el focus de la cerca als hiperparàmetres que s'ocupen de modelitzar les propietats físiques de les molècules.
- **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 3 fases: definició de la hipòtesi, 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 concloure 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].

5 RESULTATS:

Tots els resultats de l'apartat han sigut obtinguts seguint la *pipeline* representada en la ***Figura Apèndix 1***. S'ha escollit **MSE** com la mètrica utilitzada com a *loss* tant, per l'entrenament com per a la validació del model. Per últim l'entrenament s'ha realitzat sobre el **80%** de les dades mentre que la validació amb el **20%** restant. La mètrica MSE s'ha escollit perquè és una de les mètriques que s'utilitzen en el *paper* original de SchNetPack 2, per tant, facilitarà la comparació amb els resultats de referència.

5.1 Cerca d'hiperparàmetres:

Per aquest apartat s'ha decidit deixar els detalls fora de l'informe final i es troben a **l'Informe de progrés 2**, on es proporciona una extensa explicació respecte al procés realitzat i les conclusions obtingudes. Seguidament, es dona un resum d'aquest document.

5.1.1 Hiperparàmetres Computacionals:

S'ha centrat en tres variables clau: *batch_size*, *epochs* i *lr*. 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

5.1.2 Hiperparàmetres del model físic:

S'ha centrat en 3 paràmetres del model fisic:

- **dataCutoff:** el nombre de veïns que té en compte la *data pipeline* al generar les llistes de veins (**V**) els lots inicials l'hora d'aplicar les transformacions prèvies a l'entrenament,

- **trainingCutoff**: el valor de Cutoff de la funció gaussiana utilitzada en el filtre de de convolució contínua de *l'Interaction Module*.

- **n_atomBasis**: la dimensió del vector de features (**X**) generat per el mòdul de *Representation*.

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 a òptims:

- **DataCutoff**: 4
- **TrainingCutoff**: 5
- **N_atom_basis**: 38

5.2 Anàlisi profunda dels resultats:

Per aquest apartat s'han creat dos conjunts de *datasets* per validar el model (visualització en la ***Figura Apèndix 1***). L'objectiu és comprovar l'efecte de la mida de la molècula i dels elements presents en la molècula en el resultat del model:

- **Nombre d'àtoms**: conté un *dataset* per a cada mida de molècula amb més d'un 1% de representació en la base de dades, és a dir molècules d'entre 10 i 25 àtoms.
- **Elements:** conté un *dataset* per a cada element, on guarda totes les molècules que tenen aquel element en la fórmula.

5.2.1 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 a causa del desequilibri de les clases, 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

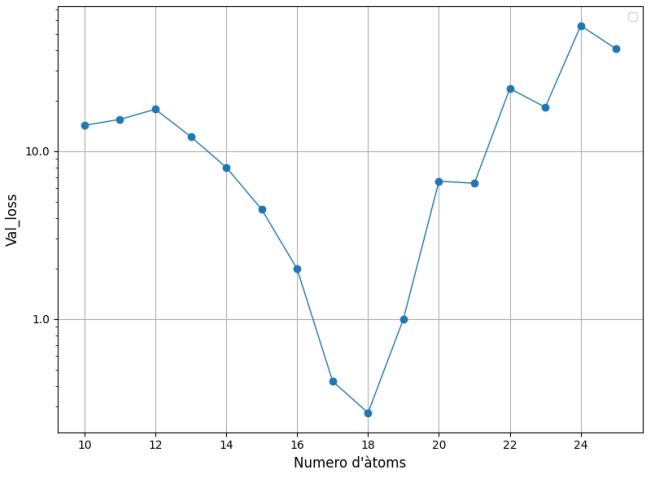


Figura 3: Model entrenat amb la base de dades QM9 *default* i validat amb el conjunt de *datasets* **Nombre d'àtoms**. L'escala de l'eix de les ordenades es logarítmic per visualitzar millor els resultats. Es pot veure la clara correlació inversa entre la freqüència d'aparició en la base de dades i la *validation loss*. **Confirma la hipòtesi 1**

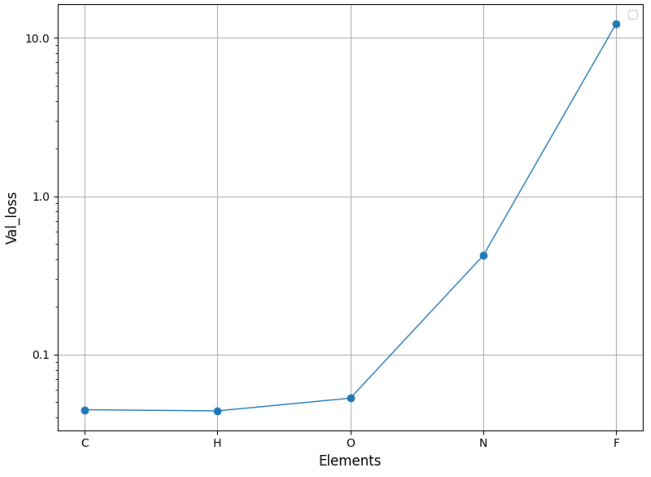


Figura 4: Model entrenat amb la base de dades QM9 *default* i validat amb el conjunt de *datasets* **Elements**. L'escala de l'eix de les ordenades és logarítmic per visualitzar millor els resultats. Es pot veure la clara correlació inversa entre la freqüència d'aparició en la base de dades dels elements i la *validation loss*. **Confirma la hipòtesi 2.**

Tenint en compte els resultats obtinguts en la ***Figura 3 i 4*** tant en el cas del **Nombre d'àtoms** i els **Elements** es pot veure una correlació inversa entre la freqüència d'aparició en la base de dades i el *validation loss*, es conclou doncs 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.

5.2.2 L'anàlisi del desequilibri de les classes:

Com s'ha evidenciat en confirmar les **Hipòtesis 1 i 2**, l'efecte del desequilibri de classes és significatiu, per tant, s'ha definit un mètode per mitigar-ne l'efecte. Aquest mètode s'ha pensat exclusivament pel **Nombre d'àtoms** no pels **Elements**.

Reducció de l'oversampling:
S'ha forçat que totes les mides de molècula tinguin la mateixa representació en la base de dades

- **Hipòtesi 3**: forçant que totes les mides de molècula tinguin la mateixa representació en la base de dades s'equilibraran els resultats.

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

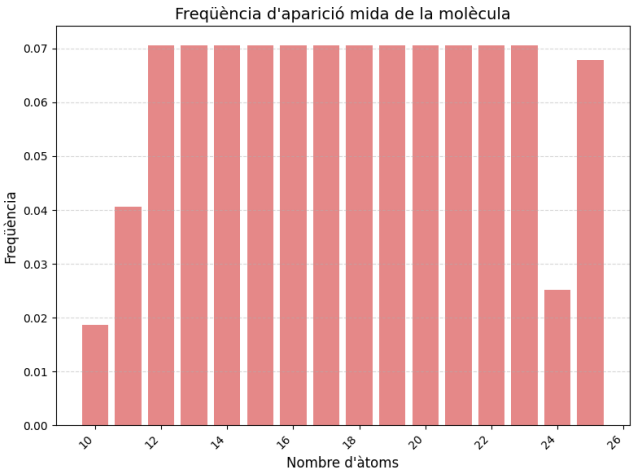


Figura 5: Histograma de la freqüència d'aparició per nombre d'àtoms en les molècules de la base de dades amb *threshold* de 2.000 molècules representades.

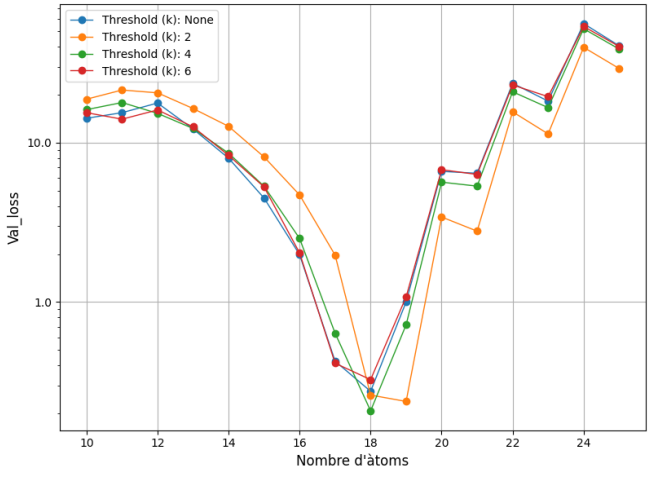


Figura 6: Models entrenats amb les bases de dades QM9 sense desequilibri de classes i validat amb el conjunt de *data-sets* **Nombre d'àtoms**. L'escala de l'eix de les ordenades és logarítmic per visualitzar millor els resultats.

Es pot visualitzar l'efecte de l'equilibració de les clases en els resultats en forma d'una millora per molècules amb més de 18 àtoms quan s'utilitza el *Threshold* de 2.000, el més equilibrat. Donat que aquest és l'únic efecte visible en els resultats, s'ha arribat a la conclusió que **la hipòtesi 4 queda descartada,** ja que no s'han equilibrat els resultats**.**

5.2.3 Millora del rendiment per Mides grans:

Vistos els resultats obtinguts en la ***Figura 3*** per molècules de més de 17 àtoms i donada la major complexitat per calcular l'energia d'aquestes molècules, s'ha establert com a objectiu: **Millorar les prediccions per a molècules de més de 17 àtoms.**

Modificació de la Funció de *loss*:
S'han definit 2 possibles funcions de *loss* modificades amb l'objectiu de beneficiar les molècules amb més de 17 àtoms. Sent **s** el nombre d'àtoms de la molècula, **f** la freqüència d'aparició en la base de dades, **r** el *rate* d'escalat de la funció i **v** el factor pel qual es multiplica el valor de *loss*.

- **Size based**: només té en compte la mida de la molècula.

$$v = \frac{(s-7)^r}{10^{r+1}} + 1$$

- **Size and Frequency based**: busca un equilibri entre la mida de les molècules i la freqüència d'aparició.

$$v = log\left(\frac{\frac{s^r}{10^{r+1}}}{f} - \frac{r}{4.5}\right)$$

Per a visualitzar millor s'ha realitzat la ***Figura Apèndix 2*** on queda representada l'evolució del *factor* (v) per determinats números de *rate* (r).

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.

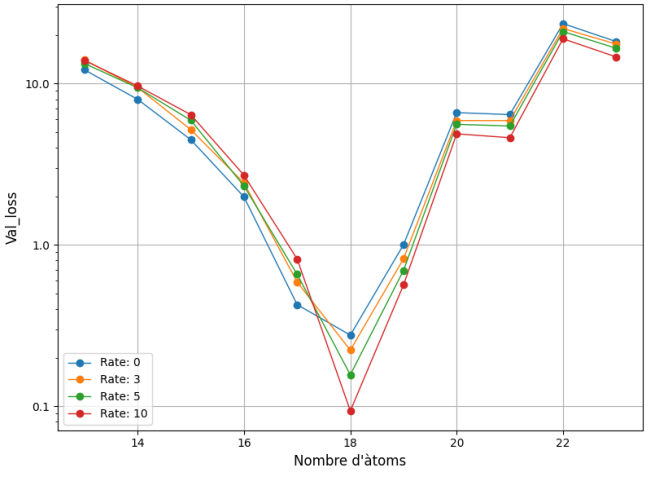


Figura 7: Models entrenats amb les bases de dades QM9 i validats amb el conjunt de *datasets* **Nombre d'àtoms** (del 13 al 23) utilitzant la modificació **Size Based**. Per referència la línia de rate = 0 representa la funció *default* de *loss* del model. S'observa que el *rate* que millors resultats dona és: **rate = 10**

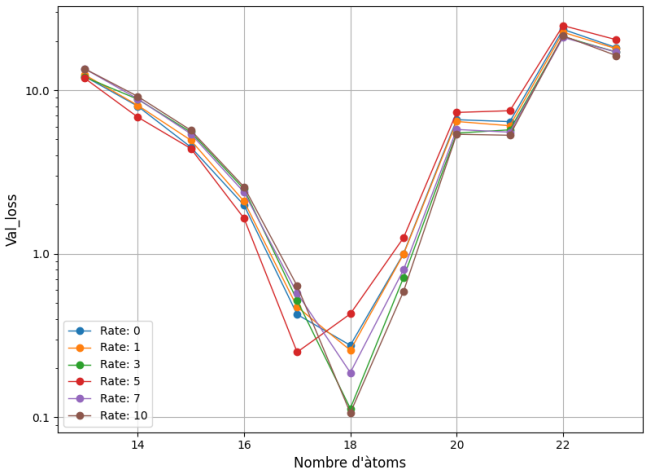


Figura 8: Models entrenats amb les bases de dades QM9 i validat amb el conjunt de *datasets* **Nombre d'àtoms (del 13 al 23)** utilitzant la modificació **Size and frequency Based**. Per referència la línea de *rate* = 0 representa la funció *default* de *loss* del model. S'observa que el rate que millors resultats dona és: **rate = 10**

S'extreu de les ***Figures 7*** i ***8*** 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.

Entrenament dirigit:
Per forçar el model a especialitzar-se en molècules de més de 17 àtoms, s'ha creat un nou *dataset* per l'entrenament amb només molècules amb més de 17 àtoms.

- **Hipòtesi 4**: el model no generalitza bé així que, si l'objectiu és millorar el rendiment per molècules de més de 17 àtoms, utilitzar tot el conjunt de la base de dades (incloent-hi molècules amb menys de 18 àtoms) és donar informació redundant al model.

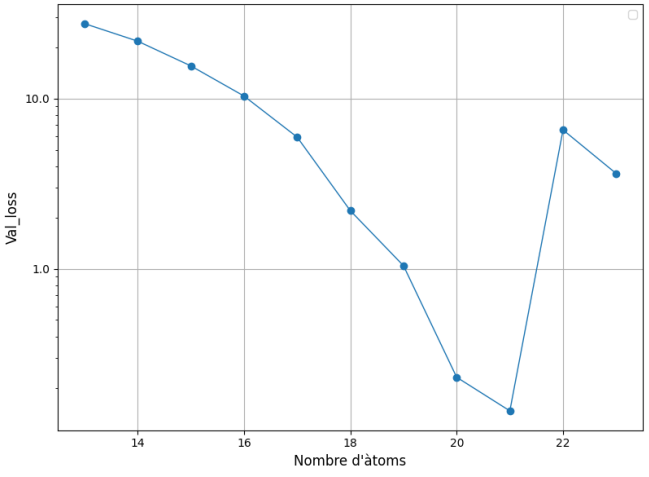


Figura 9: Models entrenats amb le base de dades QM9 splited per valors superiors a 17 número d'àtoms i validat amb el conjunt de datasets **Nombre d'àtoms.**

En ***la Figura 10*** s'aprecia l'efecte d'entrenar únicament amb les molècules amb més de 17 àtoms, els resultats per molècules amb més de 19 àtoms milloren significativament, mentre que per les molècules amb menys de 19 àtoms el rendiment empitjora. Amb això queda parcialment **confirmada la Hipòtesi 4**, donat que s'esperaven millors resultats per molècules 18 i 19 àtoms.

5.2.4 Experiment final:

Per mesurar la millora que representen els canvis proposats respecte el model base de SchNetPack 2, es compararan els models en els seguents camps:

- **MSE total:** mètrica MSE del model, validant amb el 20% de les entrades de la base de dades (seleccionades aleatòriament).
- **MSE per nombre d'àtoms**: s'ha mesurat la mètrica MSE que obté el model per a 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 el data set de 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

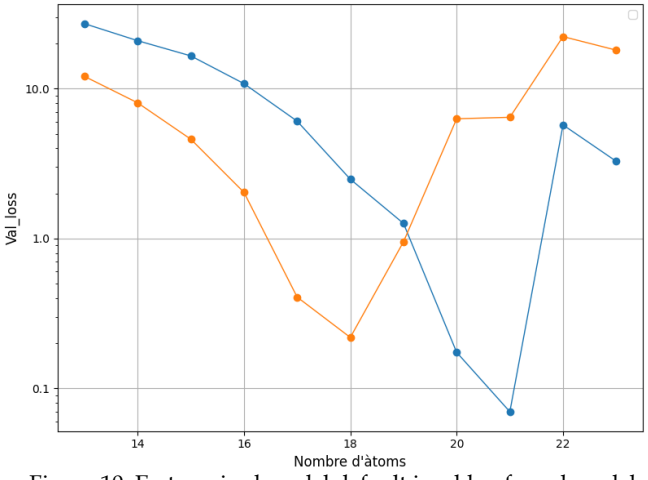


Figura 10: En taronja el model default i en blau fosc el model custom. Especificacions dels models explicat al punt 5.3.4. Mentres que per valors inferiors a 19 el model default obte resultats superiors, el model custom millora significativament el rendiment del default per molècules més grans que 19 àtoms.

Model	*Total MSE (eV)*	*Runtime Train + Val*
Default	0.0643	1h 46m
Custom	0.0925	1h 28m

Resultats de l'execució del test pels models Default i Custom de l'Experiment final, a la columna Total MSE el model custom obte un valor de MSE menor al model Default. En La columna del Runtime el model custom obté millor resultat.

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

- **MSE total**: en la ***Taula 2***, 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 sencera 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 el dataset de 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 Custom, 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 el dataset de molècules amb més de 17 àtoms (explicat en l'apartat de l'**Entrenament dirigit**).
  - Ús de la funció de *loss* modificada **Size based Rate** que com s'ha mencionat anteriorment (apartat de la **Modificació de la Funció de loss**) 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.

5.4 Conclusions dels resultats:

S'ha arribat dues conclusions respecte als models:

- **Exposició dels resultats**: utilitzar la mètrica MSE per representar els resultats del model no és adequat, ja que com s'ha vist, no és representativa dels resultats de totes les molècules sinó només de les que apareixen amb mes freqüència a la base de dades.
- **Arquitectura del model**: l'arquitectura utilitzada pel model no és capaç de generalitzar el coneixement obtingut de les molècules a altres mides i, per tant, s'hauria de considerar redissenyar el model..

6 Conclusions:

Amb aquest treball, s'ha realitzat un estudi i aplicació de xarxes neuronals per a la predicció de propietats moleculars. S'ha analitzat la base de dades QM9 i s'ha utilitzat la toolbox SchNetPack2 en aquest context.

En primer lloc, en relació amb les bases de dades de dinàmica molecular, és un gran avenç passar d'un conjunt de dades no estandarditzat, generat a partir de la contribució d'investigadors independents, a bases de dades com QM9 que segueixen l'estàndard com Jarvis Atoms. Aquesta transició representa un començament molt sòlid. No obstant això, per obtenir resultats més fiables, és necessari dur a terme un esforç addicional per millorar les bases de dades existents, evitant desequilibris tan pronunciats com els observats durant aquest estudi.

D'altra banda, és important establir un estàndard per exposar les mètriques dels models de xarxes neuronals aplicats a la dinàmica molecular. Com s'ha assenyalat a la secció 5.4, no és adequat fer servir una mètrica general i assumir que serà representativa per a tots els resultats del model. És necessari definir un marc de treball coherent per avaluar i comparar els resultats obtinguts amb aquests models.

En conclusió, l'aplicació de les xarxes neuronals a la dinàmica molecular és un camp amb un gran potencial de cara al futur, ja que permet accelerar considerablement el càlcul de les propietats moleculars. No obstant això, per aconseguir el màxim rendiment d'aquestes tecnologies, és imprescindible abordar les qüestions esmentades anteriorment i realitzar les correccions pertinents..

7 Agraiments:

M'agradaria agrair a en Jordi Faraudo i en Ramon Baldrich (tutor extern i tutor acadèmic respectivament).
Als meus pares Alexis Tibau i Judit Terma per fer això possible i a Juan Carretero, encara que ja no està amb nosaltres, el meu físic de referència.

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**, 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, 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, Journal of chemical theory and computation**, 2021, 17, 2355−2363.

[4] Weights & Biases Documentation, (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**, arXiv: 1706.08566v5

[6] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, **Quantum chemistry structures and properties of 134 kilo molècules,** Scientific Data 1, 140022, 2014.

[7] Choudhary, K. et al. **The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design.** npj Computational Materials, 6(1), 1-13 (2020).

APÈNDIX
A1. Figures:

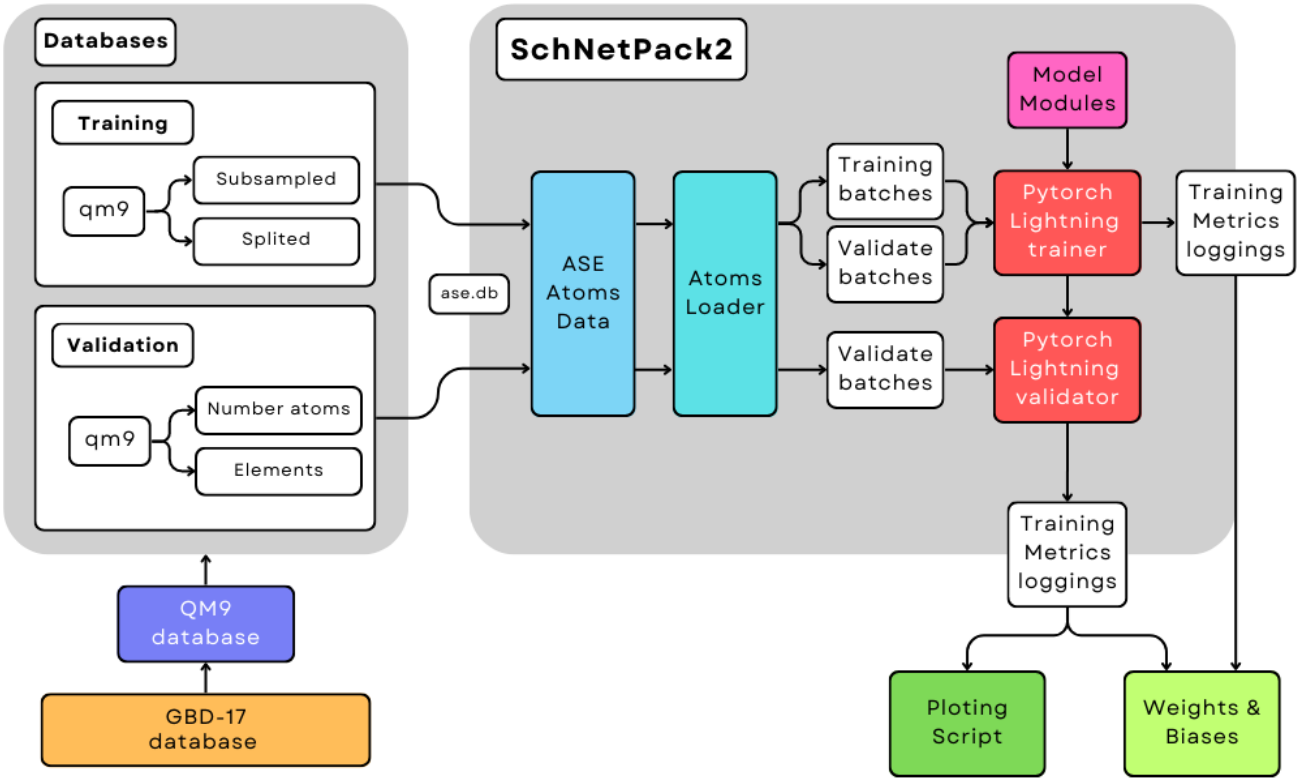


Figura apèndix 1: Diagrama de flux de la pipeline utilitzada per l'entrenament, validació i obtenció de mètriques utilitzada durant tots els experiments del treball.

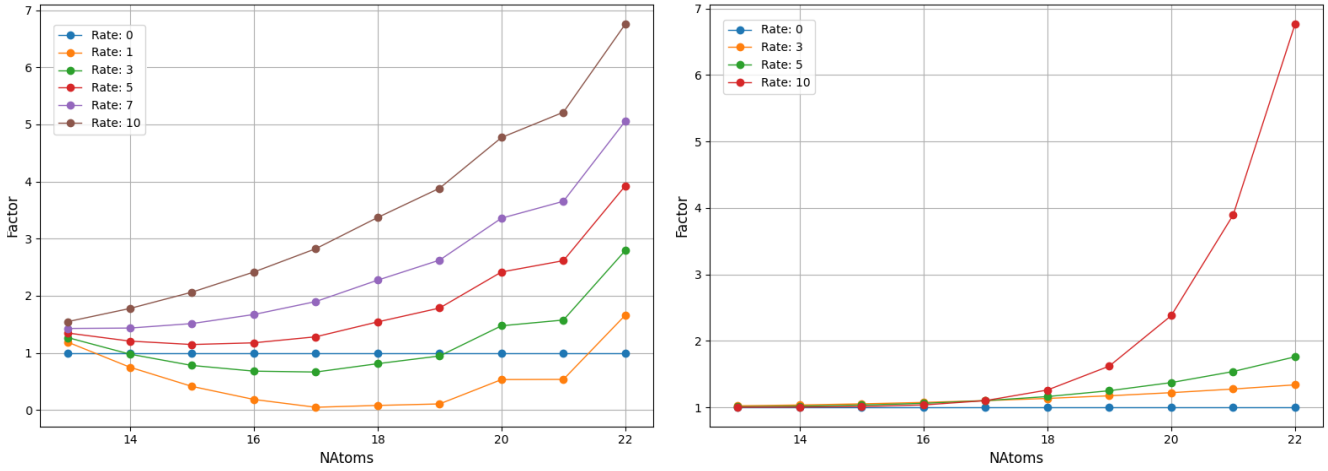


Figura Apèndix 2: Gràfic de l'evolució del factor en relació al rate per la funció **Size and Frequency Based** (a l'Esquerra) i **Size Based** (a la dreta), en l'eix d'abscisses el nombre d'elements de les molècules i en l'eix de les ordenades el factor multiplicador a la loss. Per referència la línia de rate = 0 representa la funció default de loss del model

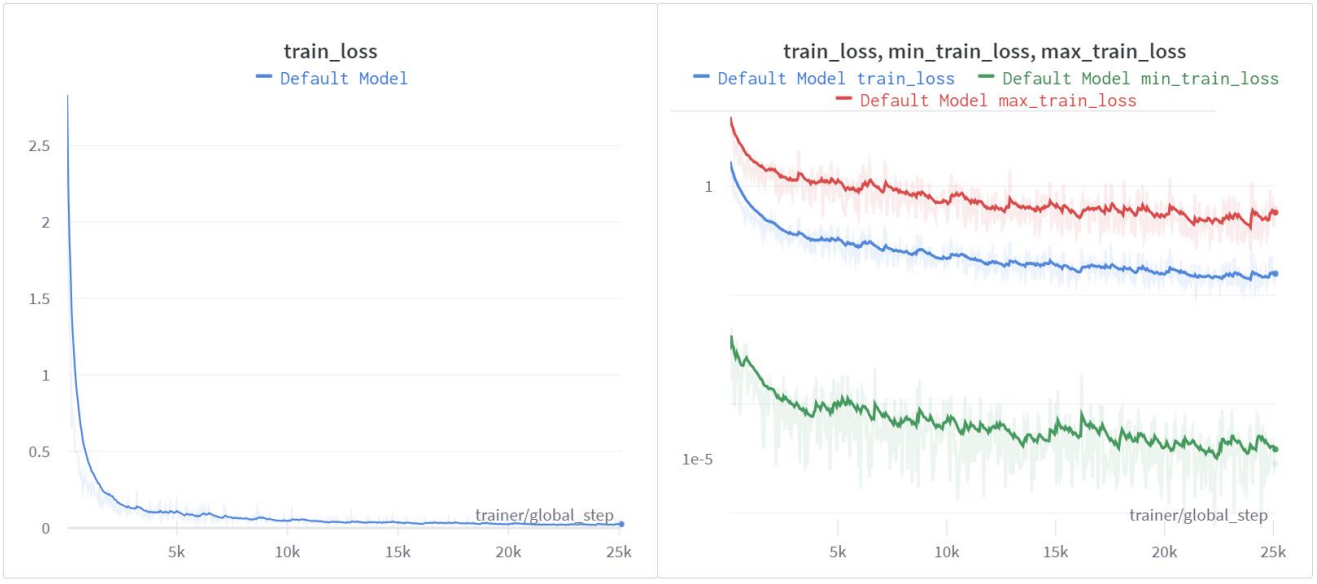


Figura Apèndix 3: A l'esquerra gràfic de l'evolució de la mètrica train loss en relació als trainer steps, a la dreta gràfic amb l'eix de les ordenades en escala logarítmica. A la dreta gràfic De l'evolució de la mètrica train loss, max train loss i min train loss en relació als trainer steps. Gràfics obtinguts durant l'entrenament del model default.

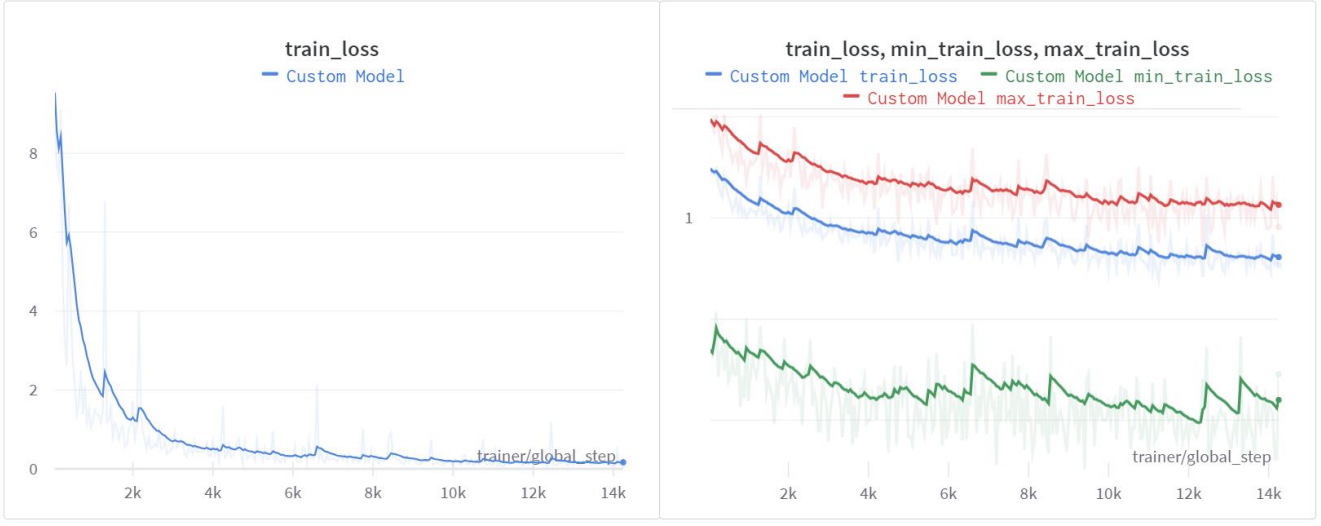


Figura Apèndix 4: A l'esquerra gràfic de l'evolució de la mètrica train loss en relació als trainer steps, a la dreta gràfic amb l'eix de les ordenades en escala logarítmica. A la dreta gràfic de l'evolució de la mètrica train loss, max train loss i min train loss en relació als trainer steps. Gràfics obtinguts durant l'entrenament del model custom. Denotar que els valors de train loss queden afectats per la funció **Size based** Rate (explicada en l'apartat 5.3.4) i per tant són no comparables amb l'escala del model base.

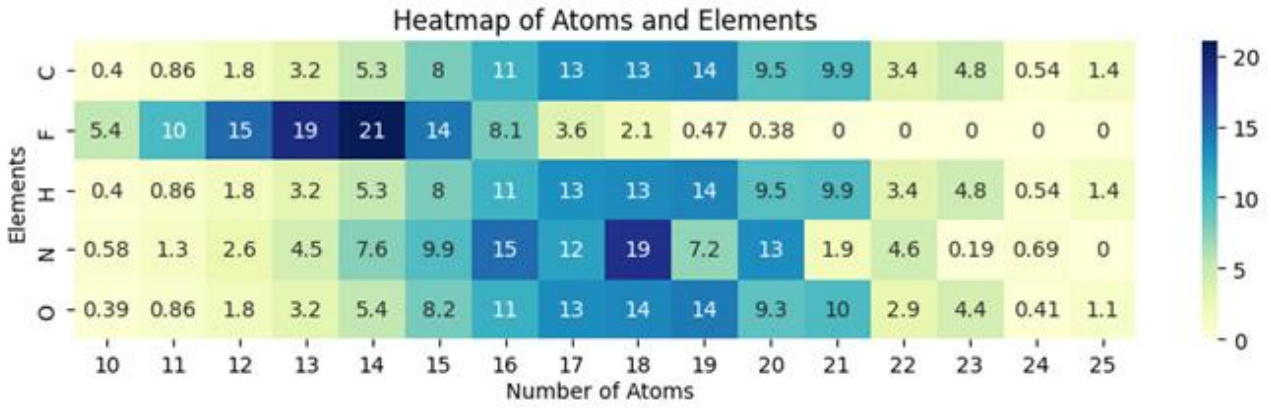


Figura Apèndix 5: Heatmap de la distribució de les entrades de la base de dades QM9. En l'eix d'abscisses el nombre d'elements de les molècules i per a cada fila de l'heatmap la distribució de les molècules amb l'element corresponent. S'aprecia que les molècules amb Fluor (F) present no segueixen la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

Objectiu	Descripció		Prioritat
O1	Desenvolupar un coneixement profund dels fonaments teòrics i pràctics de les **XN** i la **simulació de DM**		Essencial
O2	Desenvolupar una comprensió crítica dels avantatges i les limitacions de les XN en la simulació de DM, comparant-les amb altres **tècniques i abordatges existents**.		Essencial
O3	**Estudiar** els treballs mes recents del camp de la DM que apliquen XN.		Essencial
	O3.1	Bases de dades utilitzades	
	O3.2	Mètodes de representació dels sistemes de molècules	
	O3.3	Arquitectures de XN utilitzades	
	O3.4	Avaluar el rendiments dels mètodes	
O4.1	Reforçar coneixements del objectiu O3.1 (estudi de les bases de dades usades en el camp)		Essencial
O4.2	Realitzar estudi de hiperparàmetres computacionals.		Essencial
	O4.2.1	Paràmetres batch_size i epochs	
	O4.2.2	Paràmetre learning rate (lr)	
O4.3	Realitzar estudi dels mòduls de propietats físiques		Essencial

Figura Apèndix 6: Taula dels objectius definits pel treball.