# INFORME PROGRÉS 1

## 1 INTRODUCCIÓ:

En aquest segon informe de progrés, es detalla l'evolució del projecte, el qual ha experimentat canvis imprevistos a causa de la situació documentada anteriorment sobre la publicació de SchNetPackV2. Aquests imprevistos han obligat a adoptar una nova direcció en el projecte, amb l'establiment de nous objectius. Malgrat aquests canvis, s'ha mantingut la mateixa metodologia descrita en l'informe inicial, amb adaptacions realitzades a la fase 4, i la decisió de descartar la fase 5, que inicialment es considerava opcional.

## 2 PLANIFICACIÓ V2:

La nova planificació de la fase 4 del projecte consisteix en l'optimització dels models proporcionats per la toolbox SchNetPackv2 sobre la base de dades QM9, QM9 és una base de dades química que s'ha convertit en un referent en el camp de la Dinàmica Molecular (DM) i l'aprenentatge automàtic aplicat a la DM. Prové de l'estudi teòric i computacional de molècules orgàniques de baixa massa, i conté informació detallada sobre diverses propietats químiques clau.

En la primera fase, ens centrarem en l'ajust dels hiperparàmetres estrictament computacionals. Aquests hiperparàmetres, com la taxa d'aprenentatge (learning rate), el nombre d'èpoques d'entrenament (epochs) i la mida del lot (batch size), tenen un impacte directe en el procés d'entrenament i afecten el rendiment del model. Mitjançant la cerca i l'optimització d'aquests atributs explícitament computacionals, buscarem aconseguir el millor rendiment possible dels models.

En la segona fase, canviarem el focus de l'estudi als paràmetres interns de SchNetPack, com la llista de veïns neighbor list i els paràmetres dels models de representació entre d'altres.

Per dur a terme aquestes fases d'optimització, s'utilitza la plataforma Wandb, que proporciona eines eficaces per gestionar i monitorar els experiments d'aprenentatge automàtic. Amb Wandb, podrem visualitzar els resultats, comparar diferents experiments, fer un seguiment de mètriques clau i emmagatzemar models i resultats per a un accés posterior. Això ens permetrà realitzar proves iteratives i ajustar els hiperparàmetres i paràmetres interns dels models SchNetPackv2 amb l'objectiu d'aconseguir el millor rendiment possible.

Així doncs els nous objectius de la fase 4 queden definits en la següent taula.

| Objectiu | Descripció | | Prioritat |
|---|---|---|---|
| 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 |

# 3 ASSOLIMENTS:

En aquest apartat tractarem el progrés realitzat en els objectius de la última fase del projecte, incloent els objectius O4.1, O4.2 i els inicis de O4.3. També es tractarà en part la metodologia usada però no en el mateix format que l'anterior informe de progres.

## 3.1 BASE DE DADES:

En aquesta fase del projecte s'ha decidit fixar una base de dades sobre la qual fer tots els experiments, aquesta es la base de dades QM9[4]. La base de dades QM9 és una col·lecció de dades moleculars que proporciona informació essencial sobre propietats químiques i físiques de diverses molècules orgàniques petites. Aquesta base de dades s'ha convertit en una font de referència en els camps de la química computacional i la descoberta de materials.

Les dades de la base de dades QM9 provenen de simulacions quàntiques utilitzant mètodes tradicionals (completar amb ajuda del Jordi). Les simulacions s'han dut a terme per a diverses molècules orgàniques amb un nombre limitat d'àtoms. Aquestes simulacions proporcionen informació detallada sobre diverses propietats moleculars, com ara energies, capacitats calorífiques, dipols moleculars, polaritzabilitats, entre altres.

La base de dades QM9 empra el format de dades ASE (Atomic Simulation Environment) per emmagatzemar les seves entrades. ASE és una llibreria de Python amplament utilitzada per a càlculs i anàlisis de simulacions atòmiques i moleculars. El format ASE proporciona una estructura consistent i flexible per emmagatzemar informació sobre àtoms, molècules i càlculs associats.

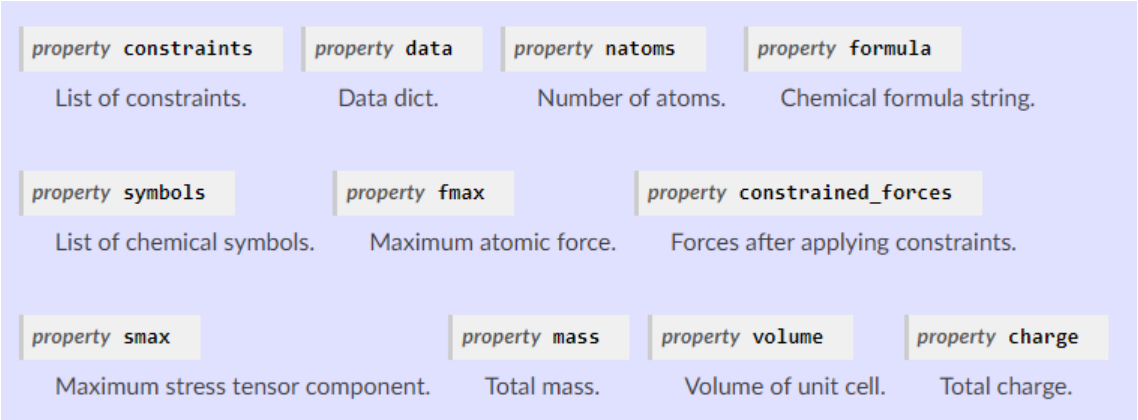
Cada entrada de la base de dades es un objecte de tipus `AtomsRow`[8] que representa una molècula, aquest te diccionari de valors amb numero variable però sempre hi ha els següents.

*Taula 1: Claus presents en cada entrada de la base de dades QM9. Taula extreta de la web oficial de ASE[8].*

| key | descripció | tipus de dada | forma |
|---|---|---|---|
| id | Local database id | int | |
| unique_id | Globally unique hexadecimal id | str | |
| ctime | Creation time | float | |
| mtime | Modification time | float | |
| user | User name | str | |
| numbers | Atomic numbers | int | (N,) |
| pbc | Periodic boundary condition flags | bool | (3,) |
| cell | Unit cell | float | (3, 3) |
| positions | Atomic positions | float | (N, 3) |

A part de això la classe també emmagatzema altres atributs però en forma de `property`, les propietats base estan en la Captura de pantalla 1.

*Captura de pantalla 1: Taula amb la llista de propietats que la classe AtomsRow te. Taula extreta de la web oficial de ASE[8].*




La `property` mes rellevant per al treball actual es `data`, la qual emmagatzema la majoria de informació respecte a la molècula representada.

```
rotational_constant_A          : [157.7118]
rotational_constant_B          : [157.70997]
rotational_constant_C          : [157.70699]
dipole_moment                  : [0.]
isotropic_polarizability       : [13.21]
homo                           : [-0.3877]
lumo                           : [0.1171]
gap                            : [0.5048]
electronic_spatial_extent      : [35.3641]
zpve                           : [0.044749]
energy_U0                      : [-40.47893]
energy_U                       : [-40.476062]
enthalpy_H                     : [-40.475117]
free_energy                    : [-40.498597]
heat_capacity                  : [6.469]
```

*Output generada al imprimir per pantalla la property data d'una row de la base de dades QM9.*

En conclusió, la base de dades QM9 utilitza el format de dades ASE per emmagatzemar informació sobre les molècules. A diferència d'una base de dades tradicional en format CSV, que utilitza taules amb files i columnes, ASE proporciona una estructura més flexible per emmagatzemar informació sobre àtoms, molècules i els càlculs associats. Mitjançant l'ús d'objectes de tipus AtomsRow, la base de dades QM9 pot emmagatzemar diverses propietats moleculars en forma de diccionaris, permetent un accés més eficient i organitzat a la informació. Aquesta estructura jeràrquica de dades ofereix una millor representació de la complexitat molecular i facilita l'anàlisi i la manipulació de les dades per a estudis teòrics i computacionals.

## 3.2 ESTUDI DE HIPERPARÀMETRES COMPUTACIONALS:

En la primera fase de l'experiment, ens centrem en l'ajust dels hiperparàmetres estrictament computacionals amb l'objectiu de millorar el rendiment dels nostres models. Aquesta fase és fonamental ja que els hiperparàmetres, com la taxa d'aprenentatge, el nombre d'èpoques d'entrenament i la mida del lot, tenen un impacte directe en el procés d'entrenament i poden afectar significativament el rendiment final del model.

Utilitzem la plataforma de monitoratge i optimització de models, wandb (Weights & Biases). Wandb ens permet explorar i ajustar els hiperparàmetres de manera sistemàtica fent l'ús de Sweeps, mantenir un seguiment de les execucions d'entrenament i analitzar els resultats per prendre decisions informades.

### 3.2.1 Paràmetres batch_size i epochs:

Ens hem centrat principalment en dues variables clau: `batch_size` i `epochs`. El nombre d'èpoques determina quantes vegades el model passa per tot el conjunt de dades d'entrenament, mentre que la mida del lot especifica la quantitat de mostres que es processen en cada pas d'actualització dels pesos del model.

Per dur a terme la cerca i l'optimització dels hiperparàmetres, hem utilitzat la funcionalitat de Sweeps de wandb. Amb aquesta característica, hem pogut definir un espai de cerca que inclou diferents valors per als hiperparàmetres de les èpoques i la mida del lot. Wandb s'encarrega de generar combinacions úniques de valors dins d'aquest espai i executa múltiples execucions d'entrenament per a cada combinació.

#### 3.2.1.1 Paràmetres

El `sweep_configuration` s'utilitza per indicar al sweep com ha de actuar, es defineix; els possibles valors que cada paràmetre pot prendre, quina mètrica es l'objectiu, si es vol minimitzar o maximitzar aquesta, etc. En aquest experiment batch_size i epochs estan fixats en una llista de possibilitats metres que lr es un rang per donar la màxima flexibilitat perquè els paràmetres que volem estudiar siguin els principals afectats. Fixar el valor de lr produiria una desavantatge per a convencions de paràmetres, per exemple si estem provant un batch size de 4096 i fem 10 epochs amb un valor de lr molt baix, el model no tindrà temps per aprendre.

```
sweep_configuration = {
    'method': 'random',
    'name': 'sweep',
    'metric': {
        'goal': 'minimize',
        'name': 'val_loss'
    },
    'parameters': {
        'batch_size': {'values': [512, 1024, 1536, 2048, 2560, 3072, 3584, 4096]},
        'epochs': {'values': [8, 10, 12, 14, 16, 18, 20]},
        'lr': {'max': 0.01, 'min': 0.0005}
    }
}
```

*Figura 1: Fragment del codi utilitzat per configurar el sweep. El sweep haurà de escollir de manera aleatòria entre els valors establerts en l'apartat de paràmetres, per a epochs i batch_size podrà escollir els valors dintre les llistes definides mentre que per el valor de lr, podrà triar el valor en el rang 0.01 fins a 0.0005.*

#### 3.2.1.2 Resultats:

Tots els resultats exposats provenen d'un conjunt de 6 experiments diferents, els quals sumen un total de 65 models entrenats.

Com mencionat anteriorment s'utilitza la plataforma Wandb per representar i monitoritzar els models. Mitjançant l'eina de report s'ha construït una sèrie de gràfics per a simplificar i sintetitzar els resultats. Les representacions per defecte (figura 2) no aporten massa informació, es necessari definir un mateix els gràfics que es busquen per a poder aconseguir informació mes polida.

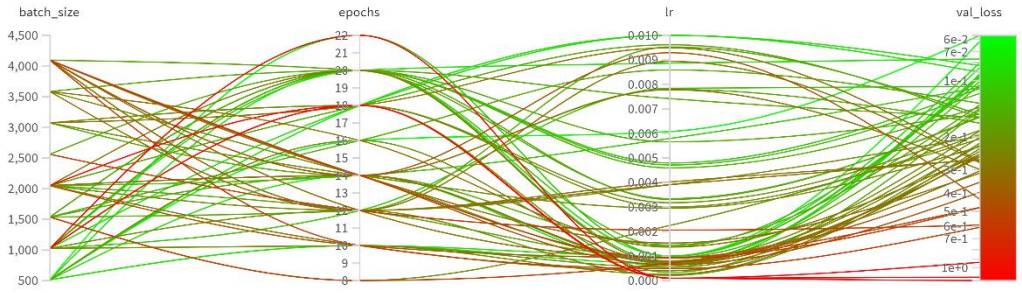



*Figura 2: Gràfic predeterminat proporcionat per Wandb per representar un sweep. Agrupa tots els models, cada un és representat per una línia que passa pels paràmetres que ha escollit i acaba al valor de val_loss que obté, aquest li dona un color obtingut d'un gradient definit per l'usuari.*

Un dels apartats mes importants del report de l'experiment es la correlació entre els paràmetres i la val_loss (mètrica objectiu). En aquest apartat, podem visualitzar en verd/vermell si la correlació es positiva o negativa respectivament, això implica una correlació directe o inversa. Tenint en compte que busquem minimitzar el valor de val_loss, partint de la Figura 3 podem inferir que molt probablement per batch_size serà millor mantenir els valors baixos i que per epochs serà el contrari, valors alts donaran millors resultats.

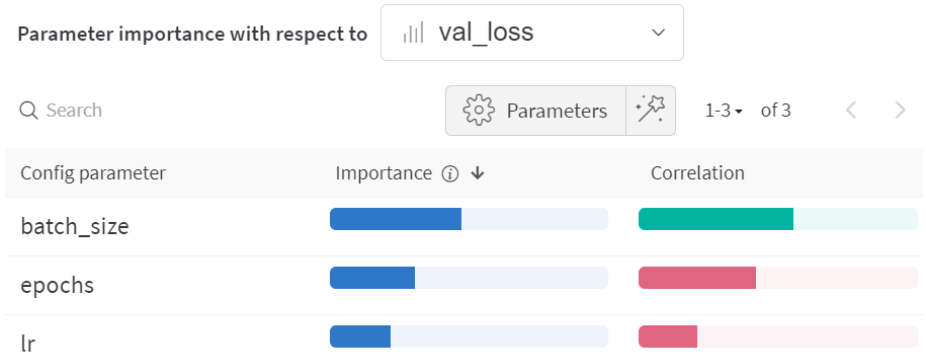



*Figura 3: Representació de la correlació entre els paràmetres i la mètrica de sortida val_loss, es pot apreciar que batch_size te una correlació positiva (en verd) i epochs la te negativa (en vermell).*

Es una evidencia bastant solida però per assegurar aquestes assumpcions es representaran les dades amb un altre mètode mes adequat. S'utilitzaran tres gràfiques dedicades per a cada paràmetre per a poder aïllar-lo de la resta.

***Hipòtesis 1: Valors de batch_size petits i numero de epochs grans minimitzarà el***

La idea principal per a les segones gràfiques es agrupar els models per la mètrica que es vol estudiar, per després seleccionar els grups mes rellevants i poder visualitzar clarament la importància de cada mètrica per al resultat final.

#### 3.2.1.2.1 Batch_size:

S'han escollit els valors 512, 2048 i 4096 (valors petit, mitja i gran). Aquests valors representen el rang de valors possibles a escollir molt adequadament i permeten visualitzar les teories comentades en la figura 3.

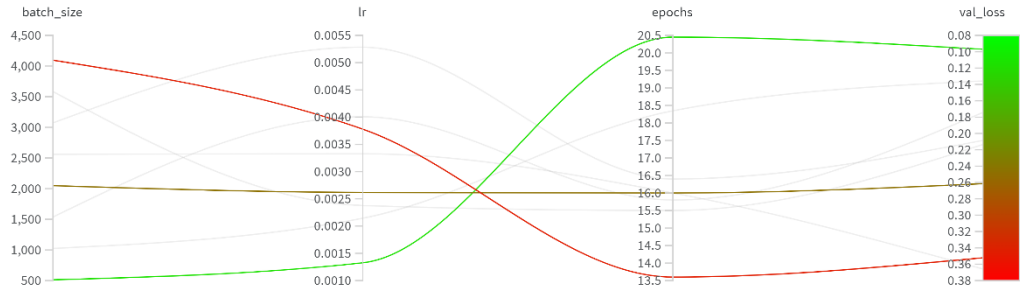



*Figura 4: Gràfic base dels paràmetres d'un sweep. S'observa clarament la correlació negativa que te batch_size amb val los. Valors mes grans de batch size porten a val_loss grans ≈ 0.35 valors mes petits (512) a valors petits ≈ 0.10, queda representat per el color de les línies (vermell gran, verd petit).*

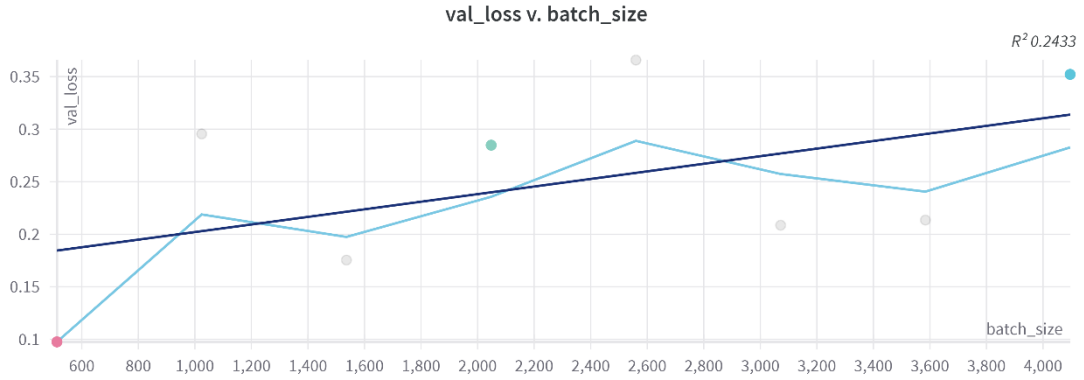



*Figura 5: Gràfic custom amb batch_size en l'eix de les x's i val_loss en el de les y's. Cada punt representa la mitja de val_loss del batch_size en el que es troba. La línia en blau fosc es la regressió lineal amb pendent 0.2433 (indicat en la cantonada superior dreta). De nou es pot interpretar clarament que per valors mes grans de batch_size obtenim valors grans de val_loss, s'aplica també a valors petits.*

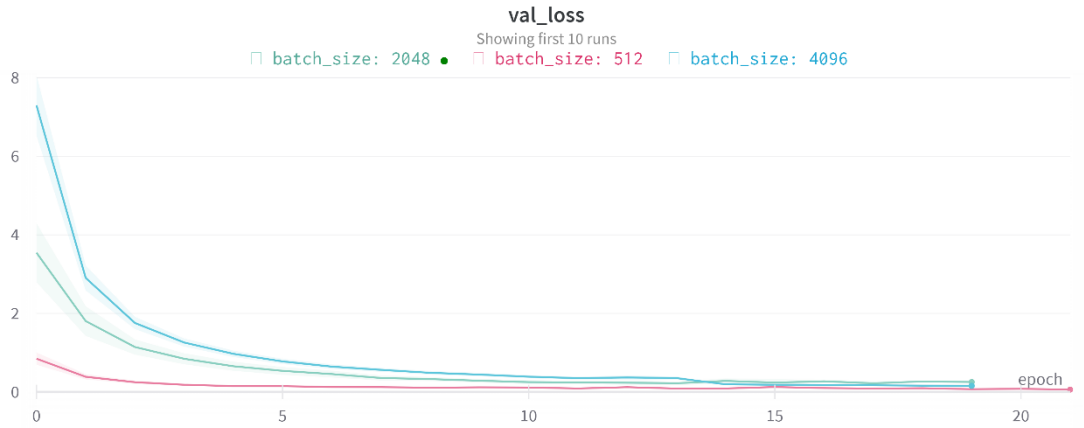



*Figura 6: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. Podem veure com per valors de batch_size petits la corba s'estabilitza en un valor de traint_loss amb un numero de epochs molt menor que la resta de batch_size. Com mes gran es la batch_size mes tarda en arribar a estabilitzar-se. Per visualitzar millor els valors petits als finals de les corbes es defineix la Figura 7.*

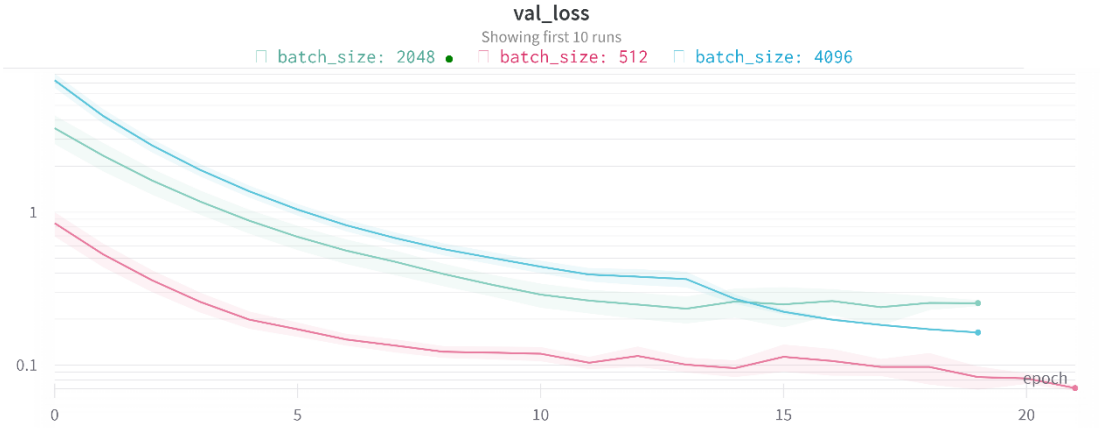



*Figura 7: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. L'eix de les y's esta en una escala logarítmica per visualitzar millor la etapa de estabilització dels models en el que traint_loss varia molt mes lent.*

#### 3.2.1.2.2 Epochs:

S'escull fer les representacions dels següents gràfics fent l'ús de 4 valors, donat que els models tenen mes opcions l'hora de escollir el numero de epochs. Els valors escollit son 8, 12, 18 i 22.

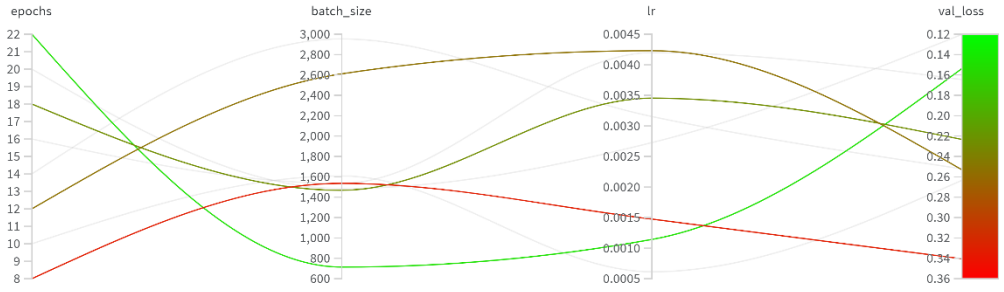



*Figura 8: Gràfic base dels paràmetres d'un sweep. S'observa clarament la correlació negativa que te epochs amb val los. Valors mes petits de epochs porten a val_loss grans ≈ 0.35 i valors mes grans a valors petits ≈ 0.15, queda representat per el color de les línies (vermell gran, verd petit).*

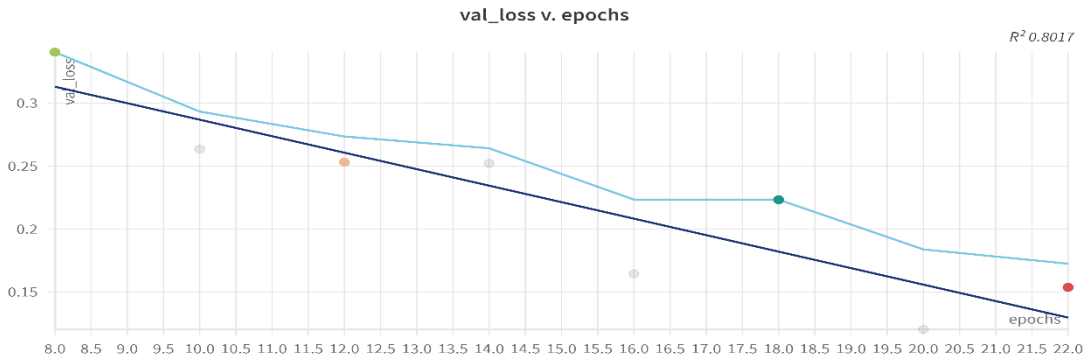



*Figura 9: Gràfic custom amb epochs en l'eix de les x's i val_loss en el de les y's. Cada punt representa la mitja de val_loss del batch_size en el que es troba. La línia en blau fosc es la regressió lineal amb pendent 0.8017 (indicat en la cantonada superior dreta). De nou es pot interpretar clarament que per valors mes grans de batch_size obtenim valors petits de val_loss, s'aplica també a valors petits.*

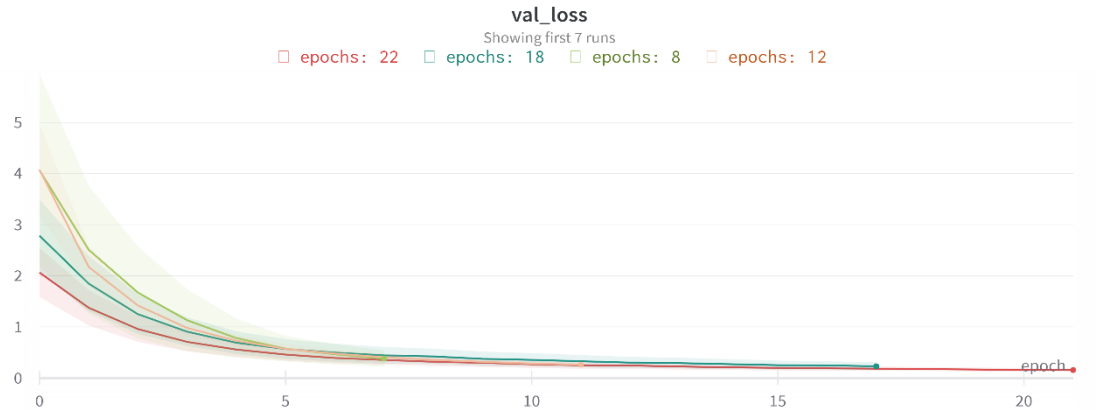



*Figura 10: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. En aquesta gràfica no es pot apreciar massa be en quin moment cada grup para de aprendre així que aplicarem les transformacions necessàries per fer-ho en la Figura 11*

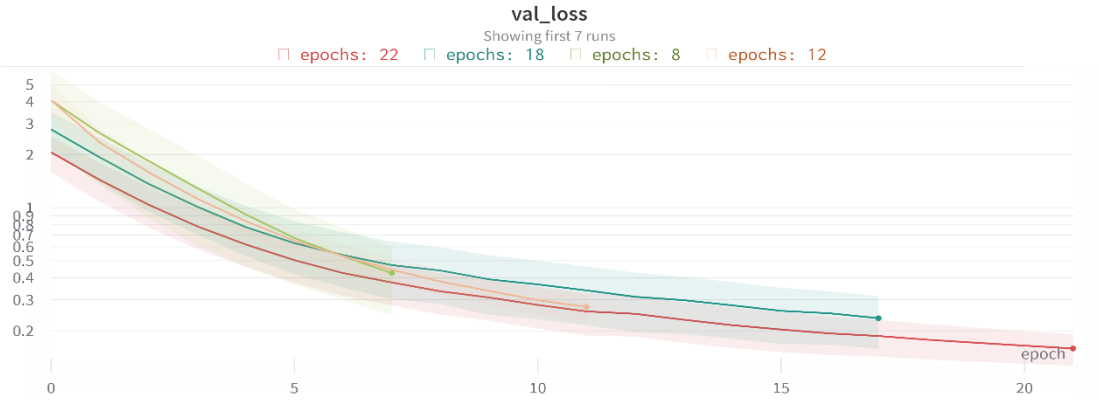



*Figura 11: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. L'eix de les y's esta en una escala logarítmica per visualitzar millor la etapa de estabilització dels models en el que traint_loss varia molt mes lent. Gracies a aquesta visualització podem veure que models amb valors de epochs petits no arriben a estabilitzar-se per falta de entrenament.*

***Hipòtesis 1: Valors de batch_size petits i numero de epochs grans minimitzarà el val_loss.***

Donades les evidencies en les Figues 3-11 podem determinar aquesta teoria com a certa. Tot i que mes epochs porten a millors resultats el cost temporal de pujar el numero de epochs no permetrà realitzar suficients experiments per a estudiar la resta de paràmetres. Així que queden establerts els valors (per a test i cerca de hyperparametres):

**`Batch_size: 512`**

**`Epochs: 22`**

#### 3.2.1.2.3 Learning rate (lr):

Per a l'estudi del valor de lr normalment es fa l'ús de potencies de 10, tot i això en els experiments realitzats per als paràmetres de batch_size i epochs es va usar un rang des de 0.01 fins a 0.0005, per a seguir els estàndards del camp s'ha realitzat un nou experiment amb 20 models en el que nomes s'usen els valors 0.01, 0.001 i 0.0001. En aquest s'ha fixat els valors de epochs entre 18 i 36 i batch_size entre 512 i 2048 per a permetre que tots els valors de lr disposin de paràmetres mes adequats. Tal i com s'ha fet en els altres paràmetres estudiats, s'han agrupat tots els models per lr utilitzat per a simplificar les gràfiques.

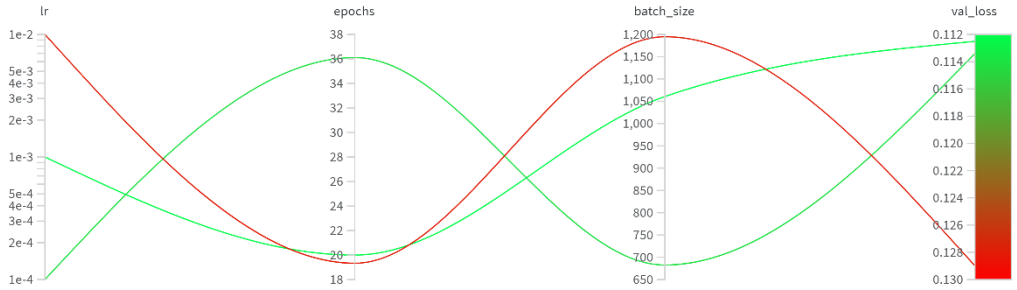



*Figura 12: Gràfic base dels paràmetres d'un sweep. L'eix de lr esta en escala logarítmica donat que el rang de valors que pren no es linear. S'observa que mentre que els valors de lr 0.001 i 0.0001 queden prou igualats, el valor de 0.01 es desmarca (negativament) per un marge considerable.*

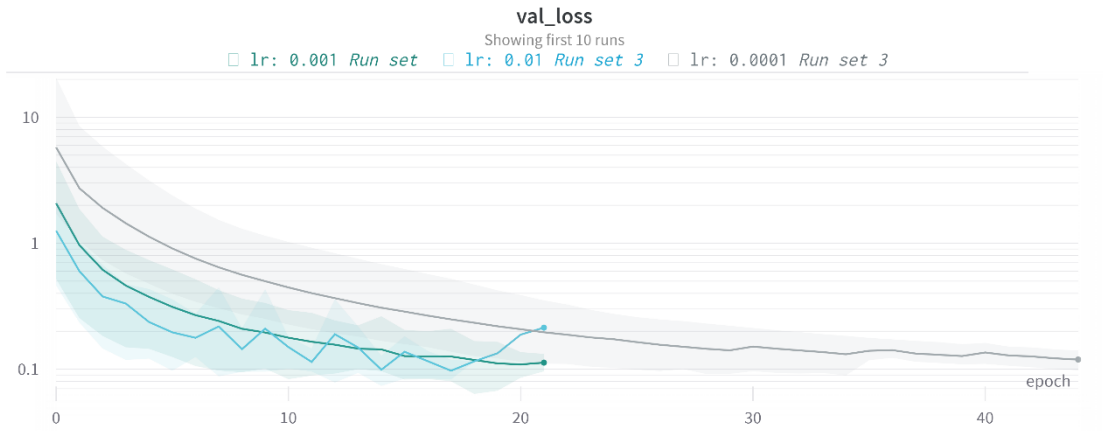



*Figura 13: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. L'eix de les y's esta en una escala logarítmica per visualitzar millor la etapa de estabilització dels models en el que val_loss varia molt mes lent. S'observen 3 conceptes importants, els models entrenats amb lr de 0.01 arriben molt ràpid a "estabilitzar-se" on comencen a incrementar i decrementar el valor de val_loss molt ràpid denotant que es un valor de lr massa gran, el models amb valor de lr 0.0001 tarden massa en arribar al seu mínim de val_loss, el models que usen lr 0.001 son el mixt perfecte entre els dos anteriors, arriben ràpid al mínim i s'estabilitzen sense variacions de val_loss extremes.*

***Hipòtesis 2: Els valors de lr petits donen mals resultats perquè els falta temps/epochs per entrenar. Si augmenta el numero de epochs la val_loss del model disminuirà.***

#### 3.2.1.2.4 Test amb èpoques extenses:

Per a aprovar/denegar la teoria 2 es fa un experiment on s'utilitzen els següents paràmetres:

**Lr:** **0.0001**

**Epochs:** **30, 35, 40 , 45, 60, 70, 80, 90 i 100**

**Batch_size:** **512**

Amb aquesta selecció s'espera obtenir evidencia sobre si amb mes epochs podem arribar als mateixos resultats que obtenim usant un lr de 0.001 amb 22 epochs.

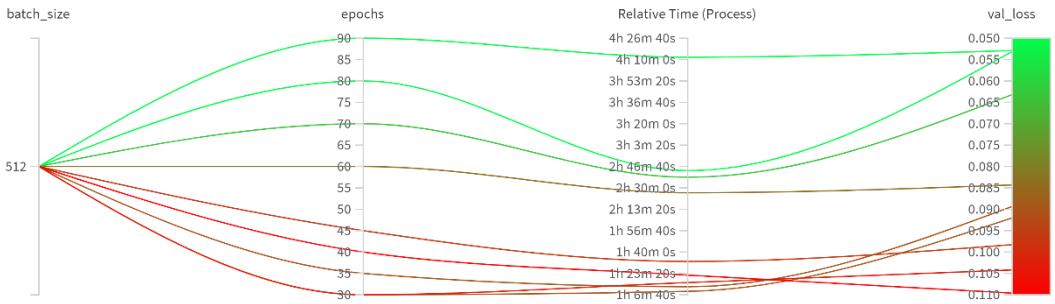



*Figura 14: Gràfic base dels paràmetres d'un sweep. Parteix d'un batch_size fixe i mostra l'efecte de les epochs i el runtime a la mètrica val_loss. Es pot veure clarament la correlació inversa/negativa entre els dos paràmetres i la mètrica val_loss, on un valor mes gran de epochs o runtime resulta en un valor menor de val_loss.*


*Figura 15: Gràfica de la evolució temporal de val_loss durant les epochs de l'entrenament dels models. L'eix de les y's esta en una escala logarítmica per visualitzar millor la etapa de estabilització dels models en el que val_loss varia molt mes lent. Queden representada la mitja per els millors models de cada valor de lr. S'observen dos conceptes importants, la velocitat per arribar a la convergència dels valors de lr = 0.001i la major estabilitat dels models amb lr = 0.0001.*

***Hipòtesis 2: Si augmenta el numero de epochs la val_loss del model disminuirà per valors de lr menors.***

La teoria queda confirmada com redactat en la Figura 14. Dit això els models amb el valor de lr = 0.001 i 0.0001 convergeixen en valors de val_loss molt pròxims com es pot veure en la Figura 15.

En conclusió es mantindrà el valor de lr = 0.001 per el seu rendiment (temporal) en comparació amb lr = 0.0001. S'estableix el valor de lr per els testos:

**`lr: 0.001`**

#### 3.2.1.3 Conclusions:

Tal i com s'ha establert en les Teories 1-2 els valors de batch_size, epochs, i lr queden fixats per a la següent etapa del projecte. En conclusió aquests valors han sigut escollits per una raons essencial, l'equilibri entre bons resultats i temps d'entrenament baixos. Es considerarà realitzar certs experiments futurs amb valors mes costosos temporalment per a millorar els resultats, com ara valors de epochs mes grans o lr mes petits ja que com demostrat en les Figures 14-15 donen resultats mes estables a canvi de temps d'execució mes llargs.

***Valors definitius per a la realització de l'objectiu O4.3:***

- **`Batch_size: 512`**
- **`Epochs: 22`**
- **`lr: 0.001`**

# 4 CONCLUSIONS:

En primer lloc, s'ha hagut d'adaptar-se als canvis imprevistos que han sorgit a causa de la situació documentada anteriorment sobre la publicació de SchNetPackV2. Això ha obligat a adoptar una nova direcció en el projecte, amb l'establiment de nous objectius. Malgrat aquests canvis, s'ha mantingut la mateixa metodologia descrita en l'informe inicial, amb adaptacions realitzades a la fase 4 i la decisió de descartar la fase 5.

En segon lloc, s'ha centrat en l'optimització dels models proporcionats per SchNetPackv2 sobre la base de dades QM9. QM9 és una base de dades química que s'ha convertit en un referent en el camp de la Dinàmica Molecular i l'aprenentatge automàtic aplicat a la DM. S'han ajustat els hiperparàmetres estrictament computacionals per obtenir millors resultats.

En el tercer apartat de l'informe de progrés, s'ha realitzat un estudi de les correlacions entre els paràmetres i la val_loss (mètrica objectiu) per inferir quins són els hiperparàmetres que donen millors resultats. Això ha permès obtenir una millor comprensió dels efectes que tenen els diferents paràmetres en el rendiment dels models.

En aquest sentit, s'ha observat que mantenir els valors baixos per batch_size i alts per epochs dona millors resultats. Això és degut al fet que un batch_size petit permet una actualització més freqüent dels pesos del model, mentre que un nombre elevat d'epochs permet una millor convergència del model.

A més a més, s'ha analitzat l'efecte del lr en el rendiment dels models. S'ha observat que una lr massa alt pot provocar una convergència ràpida però inestable, mentre que lr massa baix pot provocar una convergència lenta i estancada. Per tant, s'ha optimitzat la taxa d'aprenentatge per obtenir un equilibri entre velocitat de convergència i estabilitat.

El següent pas en l'estudi consistirà en l'optimització dels paràmetres interns de SchNetPack, com la llista de veïns neighbor list i els paràmetres dels models de representació. Gracies a l'estudi realitzat en aquesta fase permetrà obtenir una millor comprensió dels efectes que tenen aquests paràmetres en el rendiment dels models donat que quedaran aïllats dels paràmetres purament computacionals i per tant, podrem optimitzar-los per obtenir resultats més precisos i fiables.

# 5 BIBLIOGRAFIA:

[1] - 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.

[2] - 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.

[3] - MD17 (Molecular Dynamics 17): https://paperswithcode.com/dataset/md17 (accessed 5/3/2023).

[4] - Molecular Property Prediction on QM9: https://paperswithcode.com/sota/molecular-property-prediction-on-qm9 (accessed 05/3/2023).

[5] - Takeru Miyagawa, Kazuki Mori, Nobuhiko Kato , Akio Yonezu. **Development of neural network potential for MD simulation and its application to TiN.** *Computational Material Sciencie*, 15 April 2022, 111303.

[6] - Falcon, W. PyTorch Lightning. 2019, GitHub. https://github. com/PyTorchLightning/pytorch-lightning (accessed 24/3/2023).

[7] - Kristof T. Schütt, Stefaan S. P. Hessmann, Niklas W. A. Gebauer, Jonas Lederer, Michael Gastegger**; SchNetPack 2.0: A neural network toolbox for atomistic machine learning.** *J. Chem. Phys.* 14 April 2023; 158 (14): 144801

[8] – ASE oficial website, https://wiki.fysik.dtu.dk/ase/ase/db/db.html#row-objects (accessed 28/5/2023)