Base de dades QM9:

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. Les dades de la base de dades QM9 provenen de simulacions quàntiques utilitzant mètodes tradicionals. 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.

Format de les dades:

La base de dades QM9 utilitza el format de dades Jarvis Atoms 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 deses un objecte de tipus AtomsRow[8], aquest disposa de diversos atributs com ara:

- *`Formula`*: la formula de la molècula, exemple: 'CH4'
- *`Id`*: Id de la entrada en la base de dades local.
- *`Numbers`*: llista dels números atòmics de la molècula, exemple: '6,1,1,1,1'
- *`Natoms`*: el numero de àtoms de la molècula, exemple: '5'
- *`Positions`*: les posicions de cada àtom en coordenades cartesianes
- *`Properties`*: llista de propietats de la molècula (Figura Annex 1), aquestes propietats acostumen a ser l'objectiu de les prediccions.

Detalls tècnics de les dades:

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. En les figures 1 i 2 queda representada la distribució de les molècules respecte el numero de àtoms i els àtoms que les componen. Respecte la distribució dels elements, mentre que els elements C i H apareixen a un 100% de les molècules i O i N un 80% i 60 % respectivament, el fluor apareix nomes en aproximadament 3000 molècules de les 134000 totals. En el cas de les dimensions de les molècules trobem una distribució normal on els números de molècules mes freqüents son 17, 18 i 19.

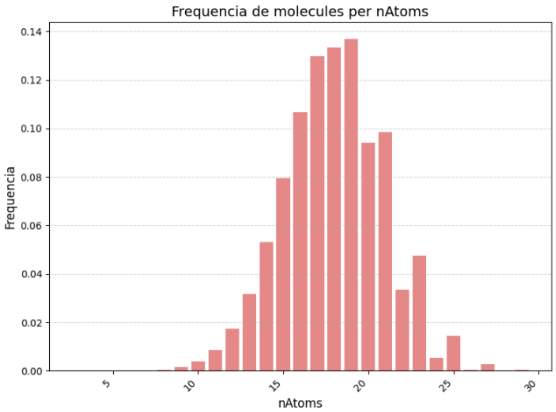


Figura 1: Histograma de la freqüència d'aparició per element en les molècules de la base de dades QM9.

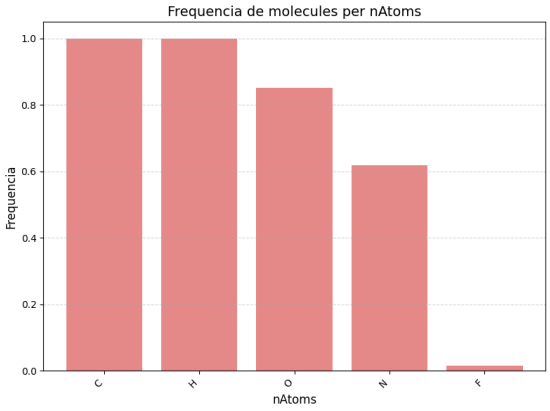


Figura 2: Histograma del percentatge de molècules per numero de àtoms en la base de dades QM9.

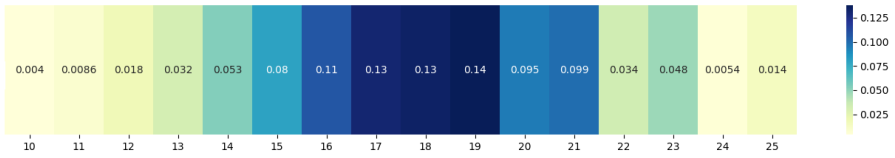


Figura 3: Heatmap de freqüència de aparició per número de Atoms en la base de dades QM9.

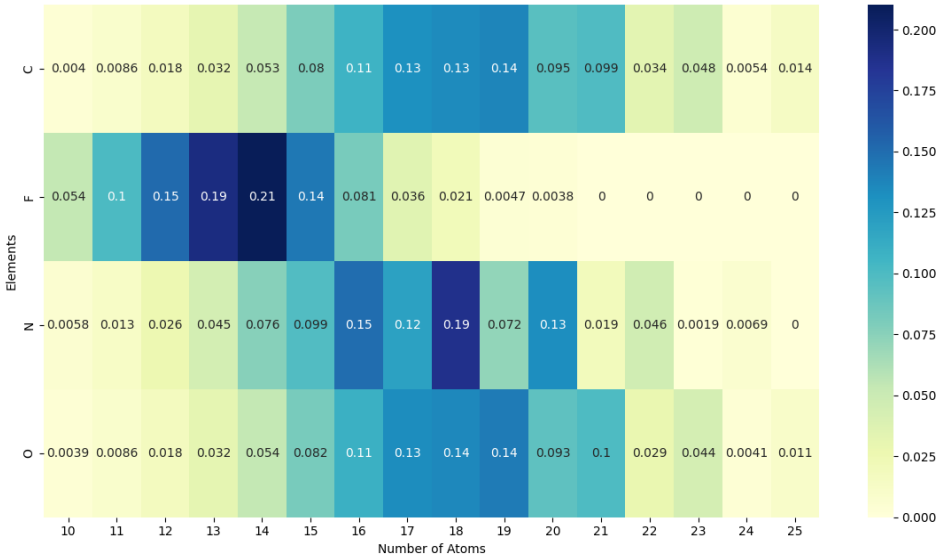


Figura 4: Heatmap de freqüència de aparició per número de Atoms per cada element en la base de dades QM9.

En conclusió, es pot esperar que en el cas que el model no pugui generalitzar be, es trobaran resultats no equilibrats, on les molècules menys representades en la base de dades no es tindran en compte durant l'entrenament, cosa que acabarà resultant en prediccions pobres per aquestes molècules.

Experiments:

En aquest apartat es discutiran la sèrie de experiments duts a terme per a la realització de les conclusions. Tots els resultats exposats son la mitja de 8 iteracions diferents. Els apartats

Impacte del numero de àtoms de la molècula en el validation loss del model:

Aquests testos estan realitzats utilitzant data frames tant per l'entrenament com per la validació. Els data frames utilitzats durant la validació son constants per tots els resultats en l'apartat, consisteixen de 15 diferents data frames amb molècules de 10 fins a 25 àtoms.

Figura 5: Model entrenat amb la base de dades sencera. Confirma la hipòtesis establerta en l'apartat anterior, on les dades amb menys representació resulten en validacions mes pobres.

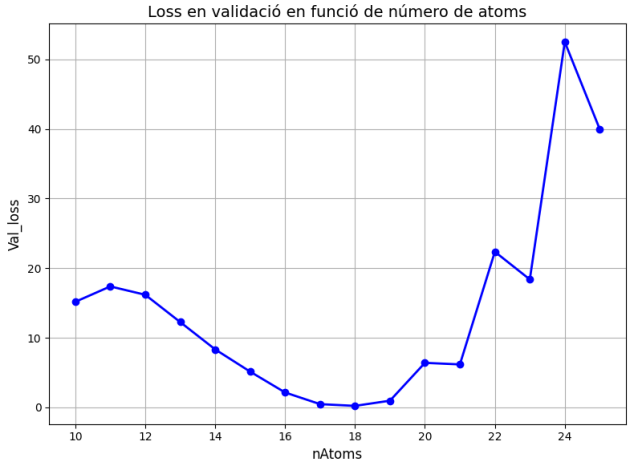


Figura 6: Model entrenat amb un data frame amb les molècules amb 17 àtoms o menys.

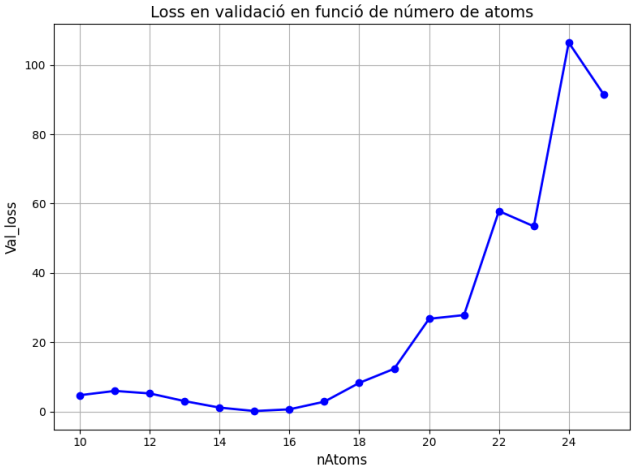
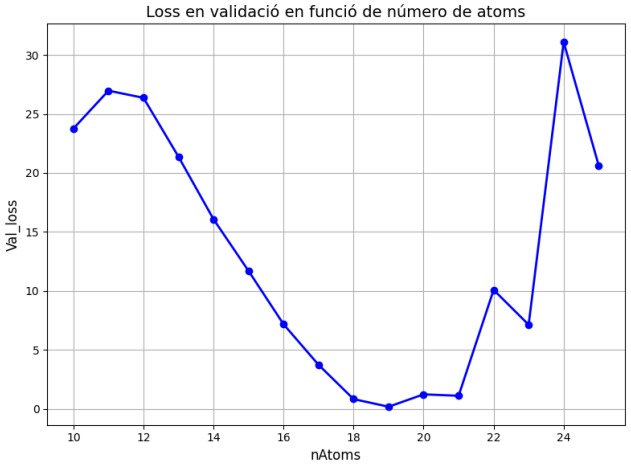


Figura 7: Model entrenat amb un data frame amb les molècules amb àtoms 19 o menys.



Impacte dels Elements de la molècula en el validation loss del model:

Aquests testos estan realitzats utilitzant data frames tant per l'entrenament com per la validació. Els data frames utilitzats durant la validació son constants per tots els resultats en l'apartat, consisteixen de 15 diferents data frames amb molècules de 10 fins a 25 àtoms.