# ANALISIS “STATE OF THE ART”:

# Introducció:

Abans de profunditzar en els diversos treballs seleccionats, és essencial establir alguns

conceptes fonamentals. Tot i que en els treballs estudiats s’hi tracten molts tipus de simulacions de dinàmica molecular (DM a partir d’ara), en aquest apartat es contemplaran majoritàriament 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, es a dir l'estat d'energia més baixa que pot tenir el sistema. Aquesta energia es dona en electrons-volts **(eV)** o be 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 numero atòmic **(Z)** dels àtoms del sistema. Aquestes condicions inicials es 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 les interaccions del sistema.

Les posicions **(R)** son comunament representades per coordenades cartesianes, que són les més senzilles, 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 obtenir 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. Petant no es realista dur a terme simulacions a gran escala i queden limitats a molècules petites. Dit això, els mètodes que utilitzen Machine Learning s'han convertit en una alternativa prometedora per al càlcul de l'energia de l'estat fonamental de molècules més grans i complexes.

Recentment han aparegut nous paquets que fan servir tècniques de Machine Learning per a accelerar les simulacions, com SchNetPack i TorchMD. Aquests paquets permeten reduir dràsticament el temps de càlcul, cosa que els fa més atractius per als investigadors que realitzen simulacions a gran escala. En la següent secció es presentaran diversos treballs que han fet servir aquests paquets de programari per a realitzar les seves simulacions de dinàmica molecular.

# SchNetPack2.0:

SchNetPack 2.0 és una “toolbox” dissenyada per al desenvolupament i desplegament de xarxes neuronals (XN) per a simulacions de DM. 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 son: una “data pipeline” flexible, modularitat a l’hora de construir els models de XN, implementació de PyTorch per a DM, una interfície de comandos basada en Hydra per a simplificar-ne l’ús i integració de PyTorch Lightning que permet gestionar i realitzar entrenaments fàcilment.

### **Data Pipeline:**

La “data pipeline” es un component crucial per al funcionament de SchNetPack 2.0, forma el framework que permet processar i preparat les dades per els models de XN. Esta compost per 2 components principals en forma de classe: **ASEAtomsData, AtomsLoader**.

**ASEAtomsData** proporciona una interfície per carregar i manipular les dades. Es un afegit a la interfície de la llibreria Atomic Simulation Environment (ASE)[10], proveïda per PyTorch això permet a l’usuari establir una sèrie de “preprocessing transforms” que s’apliquen a les dades individualment previ a que siguin agrupades i enviades als models. Aquesta funcionalitat és particularment útil per calcular llistes de veïns, eliminar offsets o assignar propietats.

Tabla

Descripción generada automáticamente con confianza baja

Taula 1: Transforms/Operacions de pre- i postprocesament disponibles.

Aquestes operacions són importants per garantir que les dades estiguin en el format correcte i que puguin ser processades eficientment pel model. Algunes de les operacions més comuns inclouen el càlcul de llistes de veïns, la eliminació d'offsets i el casting de propietats. Per exemple, les llistes de veïns s'utilitzen per identificar quins àtoms estan prou propers per interactuar entre ells. Les operacions no nomes s’apliquen abans d’enviar al model sinó també en la sortida. És important destacar que aquestes operacions només s'habiliten per a la predicció i no durant el entrenament o l'avaluació del model, donat que la funció utilitzada per avaluar el model es independent de les operacions de postprocesament.

**AtomsLoader** agrupa grans quantitats de dades per processar-les amb models. Concretament, AtomsLoader és una funcionalitat Pytorch per carrega de dades amb una funció collate personalitzada que permet agrupar les dades de manera personalitzada per adaptar-se a les necessitats i aplicacions específiques. Permet carregar aquests lots en paral·lel així assolint un processament ràpid i eficient de grans conjunts de dades.

Diagrama

Descripción generada automáticamente

Figura 1: Diagrama del funcionament de la carrega, preprocesament i agrupa- ment de les dades amb l’ús dels mòduls ASEAtomsData i AtomsLoader. A la dreta exemple dels diccionaris generats durant el procés, (sobre) diccionaris després del preprocesament, (sota) després de l’agrupament. Tots els valors son Tensors.

Figura extreta del paper original[10]

La “data pipeline” representada en la Figura 1 (esquerra), explica els passos per els quals passen les dades per a poder ser processades per els models. S’inicia el procés llegint les dades guardades en la base de dades ASE mitjançant la interfície proporcionada per ASEAtomsData, seguidament s’apliquen les operacions de preprocesament necessàries (llistat de totes les operacions disponibles en la Taula 1). Una vegada les dades han estat pre-processades, es passen a AtomsLoader, que carrega les dades que les agrupa per passar-les al model. La funció collate en AtomsLoader està personalitzada per al agrupament correcte de dades i garanteix que els tensors índex com `idx\_i`, `idx\_j` i `idx\_m` estiguin correctament desplaçats per referir-se a la posició correcta en l’agrupació.

En el segment de codi de la Figura 1 (dreta), s’exposa un exemple de la representació d’una col·lecció de 3 àtoms en la base de dades ASE[11], un diccionari de propietats on la clau es una llista del numero atòmic **Z** i les posicions **r** dels àtoms, i les dades son les propietats com ara l’**energia** en la configuració actual o **cell** utilitzada per representar les dimensions i forma de l’espai on es troben els àtoms. En verd el resultat del càlcul de les llistes veïns `idx\_i`, `idx\_j` i de les nullificacions dels offsets. A sota el resultat de la agrupació amb una altre col·lecció d’àtoms i en vermell la generació del tensor “idx\_m” que indexa les entrades amb el seu la seva col·lecció original.

### **Models:**

SchNetPack te com a objectiu tant subministrar models amb estructures ja definides com permetre el desenvolupament i entrenament de models nous, la classe **AtomisticModel** es la base d’aquesta idea.

La classe AtomisticModelés una classe que hereta de la classe nn.Module[12] de PyTorch, el que significa que hereta tota la funcionalitat de nn.Module i es pot integrar fàcilment en fluxos de treball existents de PyTorch. En la pràctica, això significa que els usuaris poden utilitzar AtomisticModel per definir les seves pròpies arquitectures de xarxa neuronal utilitzant una combinació de capes predefinides (com ara capes convolucionals o capes denses) i capes personalitzades (com ara Atomwise).

Dit això SchNetPack utilitza majoritàriament la classe NeuralNetworkPotential

### Bases de dades:

SchNetPack disposa de diverses classes que permeten accelerar el procés d’entrenament i validació dels models, aquestes classes descarreguen i estructuren

les dades en una base de dades SQLite amb l’ajuda del paquet ASE[8]. Actualment pot processar les següents bases de dades:

* QM9: 133 885 molècules orgàniques, amb fins a 9 àtoms pesats de C, O, N i F.
* ANI1: 20 milions de conformacions per a 57 454 molècules orgàniques de C, O i N.
* ISO17: 129 isòmers de C7O2H10 amb 5000 geometries inestables.
* MD17: 8 molècules orgàniques amb 500 mil entrades (de mitja) per cada una.
* MaterialsProject: dades respecte la cristal·lització de àtoms fins al numero atòmic 94.

Diagrama

Descripción generada automáticamenteTambé proporciona classes per a alimentar les dades al model durant l’entrenament i per a la realització d’aquest. Controla l’entrenament, validació, copies de seguretat entre altres funcionalitats.

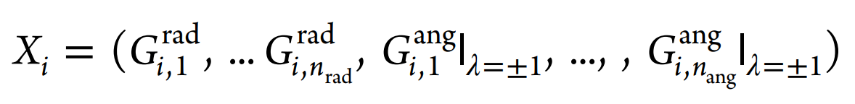
*Figura 2: Estructura de les classes per gestionar bases de dades i entrenament dels models.*

### Models:

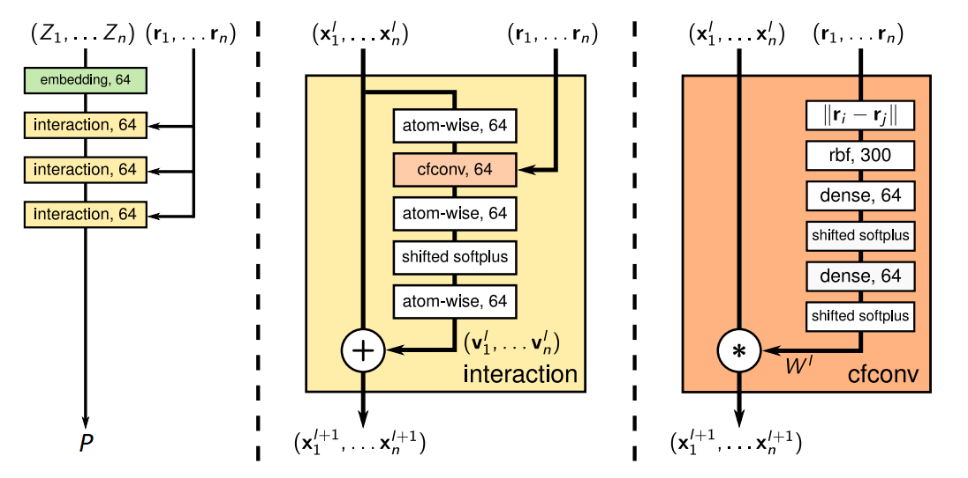
Diagrama

Descripción generada automáticamenteEls models que proporciona SchNetPack es divideixen en dos etapes, la representació i la predicció. La primera etapa consisteix de dos possibles mètodes de representació dels sistemes de molècules, wACSF i SchNet. Aquests sistemes estan compostos per n àtoms que son descrits per el seu numero atòmic Z i la seva posició R.

*Figura 1: Estructura bàsica dels blocs de representació i predicció*

(w)ACSF: o “Weighted Atom-Centered Symmetry Functions”, es una representació “descriptor-based” es a dir una representació que es calcula prèvia a l’entrenament del model això les fa especialment bones en casos on les dades disponibles per l’entrenament no son suficients per a entrenar una representació “end-to-end”. Esta basat en la combinació de RSFs o “Radial Symmetry Functions” i ASFs o “Angular Symmetry Functions” . Les RSFs es calculen a partir del sumatori de les distancies entre l’àtom que es busca representar i els àtoms que entren dintre de la distancia màxima que es considera rellevant. Aquestes distancies se’ls assigna un pes seguint una funció determinada. Les ASFs com en les RSFs tenen en compte els àtoms veïns del tractat però en el cas de les ASFs el sumatori considera l’angle entre els àtoms i no les distancies.

*Formula de (w)ACSF on Grad es el resultat de RSFs i Gang el de ASFs.*

SchNet: es una representació “end-to-end”, això comporta que aprèn a representar el sistema molecular mitjançant l‘entrenament d’una “deep neural network” o DNN. La estructura d’aquesta es la següent (*Figura 3*):

*Figura 3: Estructura de la DNN utilitzada per la representació SchNet*

* Embedding: Aquesta capa es responsable de generar un “feature vector” Xi a partir de cada tipus de àtom Zi, la dimensió d’aquest vector ve donada per F. Aquesta capa s’inicialitza aleatòriament i es va adaptant durant l’entrenament.
* Dense: Aquesta capa és la més senzilla de totes i simplement realitza una multiplicació de matrius seguida d'una funció d'activació no lineal. Aquesta capa és útil per a la combinació d'informació provinent de les capes anteriors.
* Imagen que contiene objeto, reloj

  Descripción generada automáticamenteAtom-wise: Aquestes capes són similars a les capes densas, però amb la diferència que cada àtom té els seus propis pesos (W) i biaixos (b). Això significa que la informació que es processa en aquestes capes és específica per a cada àtom i permet modelar propietats atomístiques com ara la seva energia o la seva força.
* Shifted Softplus: Aquesta capa és una funció d'activació no lineal que redueix el risc d'explosió de gradients que pot sorgir amb altres funcions. Aquesta funció és definida per la fórmula , que és similar a la funció ReLU però amb un pendent suau.



* Imagen que contiene Logotipo

  Descripción generada automáticamenteContinuous-filter Convolution: Aquesta capa realitza una convolució per a generar informació d'interacció entre àtoms. La convolució es fa amb un filtre continu en lloc d'un filtre discret, el que permet processar informació en espais no regulars, com ara l'espai tridimensional on es troben els àtoms. Això resulta en una millor adaptabilitat a estructures complexes i irregulars.
* Interaction: Aquesta capa combina totes les capes anteriors per a obtenir una representació global de la molècula. Aquesta capa és crucial per a la predicció de propietats químiques com ara l'energia o la densitat electrònica. En aquesta capa també s'inclouen tècniques com ara el dropout o la normalització de capa per a millorar la generalització del model.

# TorchMD:

TorchMD és un marc de treball de deep learning per a simulacions de MD. Està construït sobre PyTorch i ofereix una sèrie d'eines i models per simular eficientment el comportament de les molècules en entorns complexos.

TrochMD parteix de la representació implementada en SchNet.  A partir d’aquí fa una predicció pròpia usant eines especifiques implementades en Pytorch (paral·lelisme, capes, GPU, ...). Per a l’entrenament usa el framework “Pytorch Lightining” ].Per acabar aquest treball destaca en les aplicacions i resultats que presenta, du a terme 4 experiments:

* Comparació dels resultats amb ACEMD3
* Entrenament i avaluació en la base de dades QM9
* Demostració de la utilitat del model end-to-end (SchNet)
* Coarse-Graning (baix nivell de detall) en sistemes de “proteine folding”

# DeePMD-kit:

Es un paquet de programari de codi obert que utilitza DNN per a la construcció de potencials d'interacció atòmica es a dir és capaç de predir les forces i les energies que actuen sobre els àtoms d'un sistema en funció de la seva configuració espacial. Aquest treball esta enfocat a la simulació de TiN el qual una estructura cristal·lina cúbica (que per tant no s’adhereix gaire a l’objectiu del treball).

Diagrama

Descripción generada automáticamenteSegueix un funcionament similar a SchNet en el que es defineixen uns descriptors, a partir d’una funció que te en compte tant les coordenades com altres factors s’obtenen uns valors els quals passen per una capa de embedding NN que s’ocupa d’acabar de filtrar-los i per acabar es tornen a combinar amb les coordenades per constituir els descriptors que s’utilitzaran per la predicció.

El treball també entra en detall de el software usat per a generar les dades d’entrenament el qual es CALPHAD, Thermo-Calc 2021b i TCTI3 (en aquest no explica el funcionament intern d’aquests softwares però pot ser útil poder consultar-los en cas de necessitar generar dades)

Per acabar fa una sèrie de tests on compara el model NNP amb el software MEAM (el mètode establert actualment per fer aquest tipus de simulacions) i observa que NNP obté resultats molt similars a MEAM amb un cost i una eficiència molt menor per tant fent el mètode un èxit en aquest camp.

Figures opcionals:

Texto

Descripción generada automáticamente

Figura 2: Fragment extret del codi de la classe ASEAtomsData que pren com a base la classe BaseAtomsData () original en el [Github del Projecte](https://github.com/atomistic-machine-learning/schnetpack/blob/master/src/schnetpack/data/atoms.py). Es poden veure els atributs de propietats i transforms mencionats anteriorment. En l’apartat d’Arguments del codi s’explica aquests i altres no mencionats com ara les unitats.

1. **Bibliografia:**

[1] - Frenkel, Daan, and Berend Smit. **Understanding Molecular Simulation from Algorithms to Applications.** *Academic Press*, 2002.

[2] - Braun, Efrem, et al. **Best Practices for Foundations in Molecular Simulations [Article v1.0]. Living** *Journal of Computational Molecular Science*, vol. 1, no. 1, 2019

[3] - M. P. Allen and D. J. Tildesley. **Computer Simulations of Liquids.** *Oxford University Press*, 2002

[4] - J. Chem. Theory Comput. 2022, 18, 4, 2479–2493, Publication Date:March 8, 2022

[5] - 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.

[6] - 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.

[7] - 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.

[8] - Larsen, A. H.; Mortensen, J. J.; Blomqvist, et. al, **The atomic simulation environmenta Python library for working with atoms.** *Journal Phys: Condens. Matter*, 2017, 29, 273002.

[9] - [**The Schrödinger Equation for a Molecule**](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Thermodynamics_and_Chemical_Equilibrium_(Ellgen)/18%3A_Quantum_Mechanics_and_Molecular_Energy_Levels/18.04%3A_The_Schrodinger_Equation_for_a_Molecule), (consultat: 10/5)

[10] - Kristof T. Schütt , Stefaan S. P. Hessmann, Niklas W. A. Gebauer, et. al, **SchNetPack 2.0: A neural network toolbox for atomistic machine learning** , *J. Chem. Phys.* 158, 144801 (2023)

[11] - [**ASE representation for Atoms**](https://wiki.fysik.dtu.dk/ase/ase/atoms.html), (consultat: 16/5)

[12] - [**Documentacio nn.Module**](https://pytorch.org/docs/stable/generated/torch.nn.Module.html)**,** (consultat: 16/5)