# Informe Inicial

# INFORMACIO PRELIMINAR / ANTECEDENTS

La simulació de dinàmica molecular és una tècnica computacional per estudiar el comportament de sistemes moleculars. Això proporciona informació sobre l'estructura, energia, dinàmica i altres propietats del sistema.

La simulació de dinàmica molecular té un ampli ventall d'aplicacions en diferents àmbits. En la indústria farmacèutica, s'utilitza per dissenyar i optimitzar fàrmacs a partir de la comprensió de les interaccions moleculars entre els components del medicament i el receptor del pacient. En el camp de la biologia, la simulació de dinàmica molecular s'utilitza per estudiar les proteïnes i altres biomolècules, analitzant les seves propietats estructurals i la seva interacció amb altres molècules. En l'enginyeria de materials, la simulació de dinàmica molecular s'utilitza per dissenyar materials amb propietats específiques, com ara resistència mecànica o conductivitat tèrmica. A més, aquesta tècnica també s'utilitza en la física per estudiar les propietats dels sòlids i dels líquids a nivell molecular, i en la química per investigar les reaccions químiques i les propietats de les molècules

Per resoldre les equacions de moviment en la simulació de dinàmica molecular, s'utilitzen mètodes numèrics com l'algorisme de Verlet[1][2] o l'algorisme de Gear[3]. Aquests mètodes es basen en la discretització del temps i en la resolució iterativa de les equacions de moviment per obtenir l'evolució temporal de les posicions i velocitats de les partícules. Així mateix, per calcular les forces d'interacció entre les partícules s'utilitzen diferents mètodes, com ara el mètode “boundary element”[4]. L'elecció del mètode numèric depèn de la naturalesa del sistema i de la precisió i eficiència que es requereixen en la simulació.

A mesura que augmenta la complexitat del sistema, augmenta també el nombre de molècules i la quantitat de dades que cal processar. Això, a la seva vegada, augmenta el temps necessari per a completar cada pas de temps. Així, per exemple, per a simular la dinàmica molecular d'una proteïna amb milions d'àtoms, són necessàries simulacions que poden requerir desenes de milers de passos de temps i que poden trigar dies o setmanes per a ser completades en un ordinador de propòsit general. Per això, els investigadors de la dinàmica molecular han hagut de recórrer a tècniques d'optimització i paral·lelització per a reduir el temps necessari per a completar les simulacions. Això inclou l'ús de supercomputadors i la programació en paral·lel, que permeten repartir el treball en diferents nodes d'un sistema computacional per a accelerar el procés.

En resum, els mètodes numèrics utilitzats en la dinàmica molecular són computacionalment intensius i poden requerir molt de temps per a completar simulacions de sistemes complexos. La reducció del temps necessari per a aquestes simulacions continua sent un repte important per a la investigació en aquesta àrea.

Recentment, s'han desenvolupat noves tècniques basades en l'aprenentatge computacional per a fer front als problemes de complexitat de la dinàmica molecular[5][6][7]. Aquestes tècniques inclouen els mètodes basats en xarxes neuronals, que permeten entrenar models a partir de dades experimentals o de simulacions prèvies. Aquest Treball proposa estudiar els treballs publicats i fer una adaptació/millora de les metodologies aplicades en aquests.

# OBJECTIU DEL TFG

En aquesta secció, s'estableixen els objectius del projecte i se'ls assigna una prioritat segons la seva importància i grau d'impacte en el resultat final. Per a això, es proposa la utilització d'una taula que defineixi clarament cada objectiu i la seva categoria de prioritat, que pot ser essencial, secundari o opcional. Així, es podrà establir una guia clara per al desenvolupament del projecte i es podrà assegurar que s'assoleixen els objectius més importants i necessaris en primer lloc.

|  |  |  |  |
| --- | --- | --- | --- |
| Objectiu | Descripció | | Prioritat |
| O1 | Desenvolupar un coneixement profund dels fonaments teòrics i pràctics de les **xarxes neuronals (NN)** i la **simulació de dinàmica molecular (MD)** | | Essencial |
| O2 | Desenvolupar una comprensió crítica dels avantatges i les limitacions de les NN en la simulació de MD, comparant-les amb altres **tècniques i abordatges existents**. | | Essencial |
| O3 | **Estudiar** els treballs mes recents del camp de la MD que apliquen NN. | | Essencial |
|  | O3.1 | Bases de dades utilitzades | |
|  | O3.2 | Mètodes de representació dels sistemes de molècules | |
|  | O3.3 | Arquitectures de xarxes neuronals utilitzades | |
|  | O3.4 | Avaluar el rendiments dels mètodes | |
| O4 | **Explorant noves estratègies** per millorar els mètodes usats en els treballs estudiats. | | Essencial |
|  | O4.1 | Proposar noves fonts de dades d’entrenament | |
|  | O4.2 | Dissenyar noves arquitectures de NN | |
|  | O4.3 | Explorar tècniques de transferència d'aprenentatge | |
|  | O4.4 | Comparar el rendiments amb els mètodes originals | |
| O5 | **Definir, entrenar i validar** una arquitectura de NN específicament dissenyada per a la simulació de MD per interaccions **entre molècules senzilles**. | | Secundari |
|  | O5.1 | Escollir bases de dades a utilitzar per l’entrenament i validació | |
|  | O5.2 | Definir mètodes de representació dels sistemes de molècules | |
|  | O5.3 | Explorar diferents tipus d'arquitectures de xarxes neuronals | |
|  | O5.4 | Comparar el rendiments amb els altres treballs | |

Taula 1 – Sumari dels objectius del treball

# METODOLOGIA

Es separarà el projecte en fases, en cada fase es poden establir cicles de treball curts per assegurar un seguiment adequat del progrés del projecte. Per a cada cicle, es predefiniran objectius específics a assolir i es generaran informes breus per verificar si s'han assolit tots els objectius desitjats i explicar les raons si no s'han assolit. També es permetrà canviar entre cicles de treball sempre que es justifiqui adequadament o no afecti negativament a altres tasques pendents. Les reunions amb el tutor del projecte es realitzaran de forma regular, segons el que es consideri més apropiat.

Taula 2 – Fases del Treball

**Fase de formació:** S'ha de desenvolupar un coneixement profund dels fonaments teòrics i pràctics de les xarxes neuronals i la simulació de dinàmica molecular. Aquesta fase es pot realitzar amb una combinació de lectures i tutories amb especialistes dels temes.

**Fase d'avaluació:** En aquesta fase, es realitzarà una avaluació crítica dels avantatges i les limitacions de les xarxes neuronals en la simulació de dinàmica molecular, comparant-les amb altres tècniques i abordatges existents. Es realitzaran estudis de casos específics en els quals s'utilitzin xarxes neuronals per a la simulació de dinàmica molecular, i es valorarà el seu rendiment.

**Fase d'exploració:** En aquesta fase, es proposaran noves estratègies per millorar els mètodes utilitzats en els treballs estudiats. Es poden proposar noves fonts de dades d'entrenament, dissenyar noves arquitectures de xarxes neuronals i explorar tècniques de transferència d'aprenentatge. S'ha de comparar el rendiment de les noves estratègies amb els mètodes originals.

**Fase d’innovació:** En aquesta fase, es definirà, entrenarà i validarà una arquitectura de xarxes neuronals específicament dissenyada per a la simulació de dinàmica molecular per interaccions entre molècules senzilles. Es poden escollir bases de dades per a l'entrenament i la validació, definir mètodes de representació dels sistemes de molècules i explorar diferents tipus d'arquitectures de xarxes neuronals. Es compararà el rendiment de la nova arquitectura amb altres treballs estudiats.

# PLANIFICACIO

En el marc d'un TFG, la planificació és un element clau per assegurar l'èxit del projecte. La taula de planificació presentada inclou una sèrie d'objectius específics que han de ser assolits en un període de temps determinat. A través d'aquests objectius, es pretén comprendre les limitacions i avantatges de les xarxes neuronals (NN) en medicina, recopilar informació dels treballs existents, identificar les tècniques NN més efectives i aplicar-les per crear noves arquitectures NN específiques per a la medicina. A més a més, la planificació també inclou l'elaboració de dos informes de progrés, un informe final i la preparació d'una presentació i un pòster per comunicar els resultats obtinguts.

|  |  |  |
| --- | --- | --- |
| **Setmana** | **Objectius** | **Assoliment** |
| **1** | **O1** | Comprensió general dels conceptes clau |
| **2** | **O1, O2** | Identificació de les limitacions de les NN en MD |
| **3** | **O2** | Comprensió dels avantatges de les NN en MD |
| **4** | **O2, O3** | Recopilació d'informació dels treballs existents |
| **Informe Inicial** | | |
| **5** | **O3.1, O3.2** | Conèixer les bases de dades disponibles |
| **6** | **O3.2, O3.3, O3.4** | Entendre les arquitectures de les NN aplicades |
| **7** | **O3.2, O3.3, O3.4** | Estimar el rendiment de les NN |
| **8** | **O3.3, O3.4** | Comprendre les tècniques NN més efectives |
| **9** | **O3.4, O4.1** | Identificar les tècniques NN més eficients |
| **Informe de Progrés (I)** | | |
| **10** | **O4.1** | Explorar noves fonts de dades |
| **11** | **O4.1, O4.2** | Creació de noves arquitectures de NN |
| **12** | **O4.1, O4.2 O4.3** | Creació de noves arquitectures de NN |
| **13** | **O4.2 O4.3, O4.4** | Aplicar tècniques de transferència d'aprenentatge |
| **14** | **O4.3 O4.4** | Comparar el rendiment amb els originals |
| **Informe de Progrés (I)** | | |
| **15** | **O5.1, O5.2** | Seleccionar dades per a l'entrenament |
| **16** | **O5.1, O5.2, O5.3** | Creació de noves arquitectures NN específiques |
| **17** | **O5.2, O5.1, O5.4** | Comparar rendiments amb altres treballs |
| **Informe final** | | |
| **18** | **-** | Preparació Presentació |
| **19** | **-** | Preparació Presentació |
| **Proposta de Presentació** | | |
| **20** | **-** | Realització del Pòster |
| **Pòster** | | |

1. **ANALISIS “State of the art”:**

**SchNetPack:**

Presenta una arquitectura per representar i entrenar un model per predicció de MD, La primera etapa consta de dos possibles una de cada un dels tipus mencionats anteriorment , wACSF (description based) i SchNet (end-to-end).

wACSF es una representació que es calcula prèvia al entrenament/predicció del model i esta basada en per àtom comprova per els àtoms adjacents les posicions (radial symetry functions) i els angles entre ells (angular symetry functions) i convina aquesta informació així generant uns valors per a poder representar la seva situació en l’espai.

SchNet es una DNN que aprèn a representar els sistemes a partir de un entrenament, aquesta busca seguir correctament les lleis físiques.

A l’hora de la predicció usa unes capes denses definides en el treball com a atomisme (també usades durant la representació SchNet).

Diagrama

Descripción generada automáticamente

Com a afegit també hi ha una breu explicació de com usar els models i entrenar-los a partir de bases de dades publiques amb l’ajuda de funcions (les quals no entra massa en com funcionen)

**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.

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.