UNIVERSIDAD NACIONAL DE GENERAL SAN MARTÍN
COMISIÓN NACIONAL DE ENERGÍA ATÓMICA
INSTITUTO DE TECNOLOGÍA
"Prof. Jorge A. Sabato"

Desarrollo de un modelo de Machine Learning para el estudio de la adsorción de gases en nanopartículas metálicas[(*)]

Por Valentino Sanguinetti

Director del Trabajo

Dr. Julien Lam

Tutor: Dr. Ruben Oscar Weht

[(*)] Trabajo de Seminario – Ingeniería en Materiales

República Argentina

2024

Índice de abreviaturas

DFT - Density Functional Theory
MD - Molecular Dynamics
MTN - Multiply-twinned Nanoparticles
VASP - Vienna ab initio Simulation Package
ACSF - Atom-centered symmetry functions
SOAP - Smooth Overlap of Atomic Positions
MSE - Mean squared error
MAE - Mean absolute error
OLS - Ordinary least squares
PE - Prediction error
LARS - Least-angle regression
Bi-LL - Bi-Linear Lasso-Lars
Bi-Rd - Bi-Linear Ridge

Contexto

Este trabajo fue realizado en la Unité Materiaux et Transformations (UMET), ubicada en Villneuve d'Ascq, Francia y asociada a la Universidad de Lille, École Nationale Supérieure de Chimie de Lille y Centre National de la Recherche Scientifique (CNRS). La UMET es una organización que se dedica al estudio de los materiales tanto para su aplicación a nivel industrial, como al entendimiento de los fenómenos físicos subyacentes. El trabajo se realizó bajo la supervisión del Dr. Julien Lam, quien se especializa en la utilización de potenciales de Machine Learning y cálculos de energía libre para el estudio de la formación de nanocristales. Durante el trabajo se contó con el apoyo económico tanto del CNRS de Francia como de la CNEA de Argentina, en cuanto a los recursos computacionales la supercomputadora de la univerisdad de Lille fue empleada.

Índice

1. Resumen

En tiempos recientes, las nanopartículas han atraído mucha atención como materiales para la aplicación en catálisis heterogénea. Junto con el progreso tecnológico que ha habido para la producción y caracterización de las mismas, es necesario desarrollar nuevas técnicas numéricas para el estudio de los mecanismos químicos en los procesos catalíticos y para guiar al desarrollo experimental. Durante la primera parte de este trabajo, se realizaron distintas pruebas con el objetivo de desarrollar una metodología que permita la obtención de un modelo de Machine Learning (ML), capaz de describir con precisión las interacciones entre una nanopartícula metálica y una molécula gaseosa. Durante estos ensayos, diferentes técnicas para generar conjuntos de entrenamiento, descriptores del entorno atómico y modelos de ML fueron utilizados. En las bases de datos empleadas se utilizaron en primera instancia potenciales interatómicos computacionalmente económicos pero imprecisos, y en una segunda instancia, las bases de datos fueron refinadas mediante el empleo de cálculos DFT, más precisos y costosos . Finalmente, se logró obtener una metodología robusta y precisa para la obtención de modelos de ML capaces de predecir la energía de adsorción entre moléculas gaseosas y nanopartículas metálicas.

Palabras clave: Machine Learning (ML), Catálisis, Nanopartícula, Adsorción, Embeded Atom Model (EAM), Density Functional Theory (DFT), Descriptor, Regresión lineal, regresión Lasso, regresión Ridge.

2. Abstract

In recent times metallic nanoparticles have gained a lot of attention for their use in heterogeneous catalysis, along with the progress in the production and characterization techniques, it is necessary to develop new numerical tools to understand the chemical mechanisms of the catalysis processes, and to guide future experiments. During the first part of this work, different test were conducted in order to develop a methodology for obtaining a Machine Learning (ML) model capable of accurately reproducing the interaction between a metallic nanoparticle and a gas molecule. During this period of testing, different techniques for generating training sets, descriptors of the atomic neighbourhood and ML models were employed, both with inexpensive and imprecise classical interatomic potentials, and precise DFT calculations. Finally, a robust methodology for the obtention of accurate ML models for the prediction of the adsorption energy of gas molecules on nanoparticles was obtained.

Key Words: Machine Learning (ML), Catalysis, Nanoparticle, Adsorption, Embeded Atom Model (EAM), Density Functional Theory (DFT), Descriptor, Linear regression, Lasso regression, Ridge regression.

3. Introducción

3.1. Motivación

En la catálisis heterogénea un sólido es agregado a una reacción entre especies en estado gaseoso, la presencia de una superficie permite que la reacción suceda mediante una ruta alternativa, lo que a su vez genera una mejoría en el ritmo de la misma. Los catalizadores sólidos tienen un rol fundamental en la industria química, dado que no son consumidos durante la reacción y permiten un incremento en la cinética de los procesos. En las ultimas décadas, debido al surgimiento de nuevas tecnologías relacionadas con la química verde, hay un gran incentivo por desarrollar nuevos materiales para ser utilizados como catalizadores en áreas como la síntesis verde, la reducción de emisiones, la limpieza del aire, etc. [1].

La utilización de metales de transición, como níquel, plata, o paladio están ampliamente establecida para la catálisis de reacciones químicas en estado gaseoso. Pero en tiempos recientes, debido al progreso en las técnicas de producción y caracterización de nanopartículas, un nuevo campo en el estudio de los catalizadores metálicos ha surgido [2]. Las nanopartículas son ideales para ser utilizadas en catálisis debido a sus propiedades únicas, una de las principales ventajas de la utilización de nanopartículas es la elevada proporción entre superficie y volumen que estas presentan. Dado que la catálisis se ve positivamente afectada por el incremento del área, el uso de nanopartículas permite la utilización de materiales muy caros, como oro o plata, sin comprometer el área total expuesta.

Además, las nanopartículas poseen propiedades físicas y químicas muy distintas de los metales macroscópicos (o bulk), estas propiedades dependen del tamaño y geometría de las mismas, por lo cual, ajustando los parámetros de producción es posible optimizar las propiedades catalíticas de las nanopartículas producidas. En resumen, hay un gran abanico de variables a investigar para encontrar las nanopartículas óptimas para su utilización en catálisis, y desde el punto de vista de las simulaciones físicas, existe un interés por desarrollar metodologías que sean potentes y flexibles, para poder analizar grandes conjuntos de nanopartículas y predecir su rendimiento catalítico.

Para poder predecir con alta precisión la actividad catalítica de una nanopartículas es necesario realizar simulaciones de química cuántica, las cuales son extremadamente costosas desde el punto de vista computacional. En particular, las nanopartículas presentan múltiples sitios de adsorción posibles y si se buscara hacer una descripción precisa de la actividad catalítica, cada sitio de adsorción y ruta de reacción posible debería ser investigada. Adicionalmente, si tenemos en cuenta el alto costo computacional de simular sistemas grandes (más de 100 átomos) con modelos cuánticos, como la Teoría del Funcional de la Densidad (DFT, por sus siglas en inglés), podemos concluir que este enfoque es inviable para nanopartículas de más de algunas decenas de átomos.

En consecuencia, en este trabajo nuestro objetivo será desarrollar un modelo capaz de evaluar el rendimiento catalítico de las nanopartículas mediante la predicción de la energía de adsorción. Esta decisión se basa en numerosos estudios que han identificado una relación del tipo 'volcán' entre la energía de adsorción de las especies gaseosas involucradas en cierta

reacción química y el ritmo de síntesis de la misma. Es decir, que la reacción es lenta si las especies son atraídas muy fuerte o débilmente a la superficie y el ritmo máximo de reacción se da para cierta energía óptima de adsorción intermedia [3]. Esto se puede explicar considerando que los materiales que poseen baja energía de adsorción interactúan débilmente con las moléculas gaseosas, y consecuentemente, no afectan significativamente la cinética de reacción. Por el contrario, si la interacción es muy fuerte, la mayoría de los sitios de reacción se encontrarán permanentemente ocupados y el catalizador no podrá interactuar con los reactivos. Muchos trabajos han logrado derivar las relaciones tipo 'volcán' para distintas reacciones químicas, mediante la utilización de modelos de micro-cinética y la relación de Brønsted–Evans–Polanyi [4].

Para calcular la energía de adsorción, en este trabajo se buscará desarrollar un modelo de Machine Learning (ML) capaz de predecir la energía de interacción (o binding), entre una única molécula gaseosa y una partícula metálica (la cuál, por motivos computacionales tendrá una geometría fija). El objetivo es que el modelo sea capaz de predecir la energía de binding para cualquier posible posicionamiento de la molécula con respecto a la nanopartícula, de este modo, se podrán identificar las configuraciones geométricas para las cuales la energía de interacción es mínima (mayor estabilidad), y en consecuencia es más probable que suceda la adsorción.

Para desarrollar este modelo de ML, un conjunto de entrenamiento deberá ser generado y el modelo tendrá que ser ajustado y testeado para poder asegurar que su funcionamiento sea consistente. Para testear distintos parámetros relacionados al modelo de ML y para desarrollar una metodología para la generación de los conjuntos de entrenamiento y de prueba, durante la primera parte de este trabajo, el modelo de ML será utilizado con bases de datos que contengan energías de interacción calculadas mediante potenciales interatómicos clásicos, lo cual nos permitirá evaluar la energía con un bajo esfuerzo computacional pero a la vez con baja precisión. Una vez que se hayan sentado las bases con los cálculos clásicos, se emprenderá el trabajo más laboriosos de obtener bases de datos más precisas, cuya energía sea calculada mediante el Método del Funcional de la Densidad (DFT, por sus siglas en inglés).

De este modo, nuestro objetivo será desarrollar una metodología que nos permita obtener un modelo capaz de predecir las propiedades catalíticas de una nanopartícula, sin importar su geometría, composición o reacción química para la cual se pretenda emplearla. Nuestra aspiración es que dicho modelo sea capaz de contribuir al avance del estudio de la catálisis heterogénea, proveyendo una herramienta que requiera de mínimo esfuerzo computacional y que pueda servir como punto de partida para análisis más complejos.

3.2. Geometría de las nanopartículas

Para estudiar las propiedades catalíticas que presentan las nanopartículas es fundamental conocer la posibles configuraciones geométricas que estas pueden presentar. Las nanopartículas poseen geometrías mucho más complejas que los metales macroscópicos, estas pueden exhibir distintas superficies cristalinas (facetas) ([111], [110], [100], etc.), y otras características

especiales, como los bordes donde dos superficies se intersectan o las aristas donde se encuentran los bordes. Cada una de estas características geométricas da lugar a distintos sitios de adsorción, en cada caso el entorno atómico es distinto y en consecuencia cada sitio presenta distinta energía de adsorción.

Adicionalmente, los tipos y cantidades de características geométricas presentes están determinadas por la forma y tamaño de la nanopartícula. En consecuencia, para estudiar con rigor la propiedades catalíticas, un amplio conjunto de configuraciones geométricas debe ser considerado [5].

Nanopartículas obtenidas por el criterio de Wulff

El criterio de Wulff, o construcción de Wulff, es una formulación desarrollada a principios del siglo 20, que fue crucial para entender y predecir la geometría de cristales macroscópicos. Más adelante, se encontró que el criterio de Wulff también puede ser aplicado a nanopartículas [4,6]. La construcción de Wulff se basa en el principio de Gibbs, según el cuál, la configuración estable de una cantidad de materia es la que minimiza la energía superficial.

Para el caso de sólidos cristalinos, los planos superficiales están caracterizados por sus índices de Miller (hkl), estos índices también caracterizan la energía superficial por unidad de área de dicho plano. Wulff propuso que para un cierto volumen, la geometría óptima puede obtenerse del siguiente modo:

En un conjunto de ejes cartesianos, desde el origen O, se debe dibujar un plano normal al vector $[hkl]$, a una distancia de O igual a d_{hkl}. Con:

$$d_{hkl} = \lambda * \gamma_{hkl} \tag{1}$$

Dónde, γ_{hkl} es la energía superficial del plano con índices de Miller (hkl) y λ es un factor de escala utilizado para definir el volumen del cristal. Finalmente, la geometría estable es la encerrada por la intersección de los planos, como se puede ver en la **Fig. 1**.

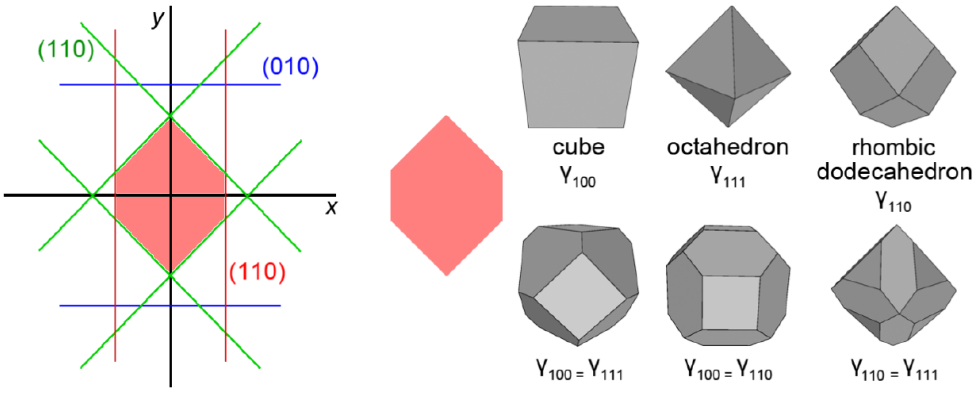


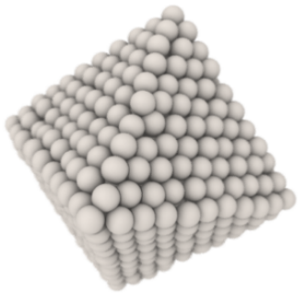
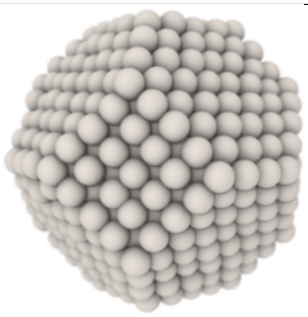
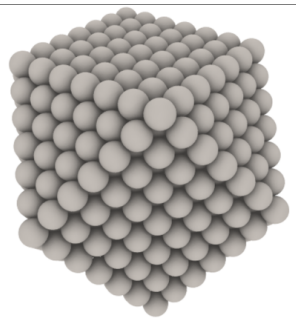
Figura 1: Construcciones de Wulff bi y tridimensionales.

Para el caso de los metales de los grupos XI y XII (como el oro y la plata), $\gamma_{111} < \gamma_{100} < \gamma_{110}$ [7], dependiendo del cociente $\gamma_{100}/\gamma_{110}$ octaedros (Oh) o sus truncaciones son las geometrías

más favorables. Los casos más predominantes de octaedros truncados son el octaedro truncado regular (To), que presenta facetas regulares hexagonales (111) y cuadradas (100), y el cubo-octaedro (Co), el cual exhibe una alternación de facetas triangulares (111) y cuadradas (100).

Estas nanopartículas son presentadas en la **Tabla 1**.

Tabla 1: Nanopartículas prominentes obtenidas por el criterio de Wullf.

Nanopartículas obtenidas por el criterio de Wullf			
Nombre	Abreviación	Descripción	Figura
Octaedro	Oh	Presenta facetas triangulares (111)	
Octaedro Truncado	To	Presenta facetas hexagonales (111) y facetas cuadradas (100)	
Cubo-Octaedro	Co	Presenta facetas triangulares (111) y facetas cuadradas (100)	

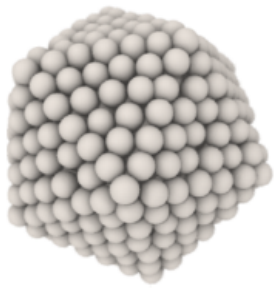
Nanopartículas múltiplemente macladas

Aunque es una herramienta muy útil, la validez de la construcción de Wulff está limitada debido a múltiples motivos, entre los que podemos destacar, el cambio en la energía superficial de las facetas como consecuencia de la interacción con una fase gaseosa o un sustrato, el

efecto de la energía de deformación de la red, efectos de tamaño finito, o la estabilización de ciertas geometrías debido al llenado de orbitales moleculares [8].

Se ha encontrado experimentalmente que las nanopartículas pequeñas pueden presentar lo que se denomina 'cyclic twinning', en estos casos las partículas poseen ejes de simetría rotacional de grado 5, formando un un patrón de planos gemelos (twin) intercalados. Dentro de la clase de las nanopartículas múltiplemente macladas, o como se denominan en inglés, Multiply-twinned Nanoparticles (MTN), el caso más prominente es el icosaedro (Ih), en la cual 20 tetraedros se encuentran en el centro de la partícula, generando una morfología donde ejes de simetría de grado 5 pasan a través de cada punto de intersección entre 5 tetraedros. A diferencia del caso de las geometrías obtenidas según el criterio de Wulff, donde solo existe una estructura continua fcc, los icosahedros poseen secciones con distintas orientaciones cristalinas y distinta estructuras (fcc y hcp). Para dar lugar a esta geometría, los tetraedros deben presentar cierto grado de deformación, pero fue demostrado por Marks [9] que la presencia de los bordes gemelos hace que la partícula sea estable o metaestable, dependiendo de su tamaño, explicando por qué estas partículas son tan prominentes. Otras partículas de la familia menos relevantes, como el ino-decahedro, o Marks-decaedro no serán consideradas en este trabajo.

Tabla 2: Multiply-Twinned Nanoparticles utilizadas.

Multiply-Twinned nanoparticles (MTN)			
Nombre	Abreviación	Descripción	Figura
Icosaedro	Ih	Presenta 20 tetraedros deformados, facetas (111) triangulares en todas las superficies y ejes de simetría de rotación de orden 5 que se encuentran en el centro.	

4. Metodología

4.1. Métodos para el cálculo de Energía

4.1.1. Potenciales interatómicos clásicos

En nuestro modelo se trabajará únicamente con sistemas estáticos. En el caso de los potenciales interatómicos clásicos, se considerará a cada átomo en el sistema como un punto cuyo comportamiento puede describirse según la mecánica clásica. La energía resultante de la interacción entre los átomos se calculará mediante un potencial interatómico, el cual se obtendrá a partir del modelo EAM (embedded atom model).

En el EAM, propuesto por Daw y Baskes en los ochentas [10], la energía que cada átomo posee está descripta por dos términos, el primero se corresponde a la energía que surge de insertar un átomo en una región con una densidad de carga debida a la presencia de los átomos circundantes, mientras que el segundo término surge de la interacción electrostática entre los núcleos. Esto se expresa del siguiente modo,

$$e_i = F_{\alpha_i}\left(\sum_{j\neq i}\rho_{\beta_j}(r_{ij})\right) + \frac{1}{2}\sum_{j\neq i}\phi_{\alpha_i\beta_j}(r_{ij}) \tag{2}$$

En consecuencia, la energía total del sistema es,

$$E = \sum_i e_i \tag{3}$$

donde, α_i y β_j se refieren al átomo i de la especia α y el átomo j de la especie β, respectivamente, mientras que ρ representa la densidad electrónica y ϕ a la interacción entre pares de núcleos.

Para un cierto átomo i, su respectiva energía e_i se calcula tomando en cuenta las interacciones con todos los átomos j que se hallan dentro de una esfera definida por un radio de corte. Las funciones $F\,,\rho\,,\phi$ y el radio de corte fueron extraídos de un trabajo de Foiles, Daw y Baskes [11]. En este, se ajustaron predicciones obtenidas utilizando el EAM con cálculos $ab-initio$ y propiedades medidas para distintos materiales.

Una limitación del empleo del EAM es que este fue descripto con la intención de ser representativo de las interacciones en el bulk de metales fcc, por lo cual solo los potenciales para el Cu, Ag, Au, Ni, Pd y Pt, pudieron obtenerse. Por ende, se decidió reemplazar al C por Ni y el O por Cu a la hora de realizar los cálculos, en la practica, definiéndose moléculas de $NiCu_2$ y NiCu. Por ende, aunque se espera que los valores de energía obtenidos sean imprecisos, debido al bajo coste computacional en la evaluación de la energía mediante el EAM, se optó por utilizar este potencial para obtener resultados cualitativos que permitan testear al modelo de ML.

Para los cálculos de mecánica clásicos, LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator, por sus siglas en inglés) [12], fue utilizado. LAMMPS emplea los potenciales definidos por Foiles, Daw y Baskes para interacción entre átomos de la misma especie y para obtener los potenciales mixtos realiza un promedio. Se utilizó LAMMPS para calcular la distancia del enlace Ni-Cu, d_0, y la energía de la molécula en el vacío, E_m, tanto para el NiCu como el $NiCu_2$ (**Tabla 3**). En el resto del texto, cuando se hable de las moléculas de CO_2 y CO, en el contexto de los cálculos mediante LAMMPS, se estará de hecho haciendo referencia a las moléculas cuya composición es Cu y Ni.

Tabla 3: Propiedades de equilibrio con el potencial EAM

Molécula	d_0 $[\mathring{A}]$	E_m $[eV]$
CO_2 ($NiCu_2$)	2.1917	-4.873
CO (NiCu)	1.1435*	30.195**

* Para el NiCu, se decidió utilizar la longitud de enlace correspondiente al CO obtenida mediante DFT, dado que el objetivo en este caso es preparar a las geometrías para realizar posteriormente los cálculos mediante DFT.
** Dada la diferencia entre la d_0 que se empleó y el modelo utilizado para calcular la energía, esta resulta positiva, vale aclarar que esto no es sensato desde el punto de vista físico.

Para obtener la energía de interacción (ligadura) en un cierto sistema, primero se debe calcular la energía potencial de la nanopartícula aislada E_{np} y la molécula aislada E_m, luego se calcula la energía del sistema nanopartícula-molécula E_s. Finalmente, se obtiene la energía de interacción o ligadura como,

$$\Delta E = E_s - (E_{np} + E_m) \tag{4}$$

4.1.2. DFT

La Teoría del funcional de la Densidad (DFT) es una teoría basada en la mecánica cuántica. Es un enfoque que es formalmente exacto pero en la práctica empírico, que busca resolver mediante primeros principios el problema fermiónico de múltiples electrones presente en los sistemas físicos [13, 14]. El modelo de DFT es muy empleado debido a que ofrece un buen compromiso entre esfuerzo de cálculo y precisión, en comparación con otros métodos basados en la mecánica cuántica. El objetivo de los modelos cuánticos es el de resolver la ecuación de Schrödinger,

$$\widehat{H}\Psi(r_1, r_2, ..., r_N) = E\Psi(r_1, r_2, ..., r_N) \tag{5}$$

Debido a que los núcleos tienen una masa mucho mayor que los electrones, estos pueden ser considerados como estáticos respecto a los electrones, en consecuencia, la función de onda

puede ser dividida en dos términos independientes, uno correspondiente a los electrones y otro para los núcleos, esto es conocido como la aproximación de Born-Oppenheimer. Luego, se dirige el esfuerzo en resolver la función de onda electrónica, dado que de esta pueden obtenerse la mayoría de las propiedades del sistema. En consecuencia, el foco se pone en el operador Hamiltoniano electrónico $\widehat{H}_e$, este cuenta con tres términos; la energía cinética, la interacción con los potenciales externos (V_{ext}) y las interacciones electrón-electrón (V_{ee}),

$$\widehat{H}_e = -\frac{1}{2}\sum_{i}^{N} \nabla_i^2 + \widehat{V}_{ext} + \sum_{i<j}^{N} \frac{1}{|r_i - r_j|} \tag{6}$$

Para el estudio de los materiales la aproximación de Hartree-Fock puede ser empleada. Según esta la función de onda de los electrones del sistema $\Psi_e(r_1, r_2, ..., r_N)$ puede ser obtenida a partir de un determinante de funciones de onda individuales para cada electrón $\phi_i(r_i)$. Esto lleva a las ecuaciones de Hartree-Fock (SCF), y en consecuencia a las condiciones matemáticas necesarias para resolver la función de onda electrónica del sistema. Los métodos de mecánica cuántica semi-empíricos (SQM) se basan en el formalismo de Hartree-Fock y proveen resultados muy precisos, sin embargo, estos son prohibitivamente costosos de emplear en sistemas con muchos electrones, dado que se basan en obtener la función de onda Ψ_e de dimensión $3N$.

Afortunadamente, no es necesario resolver las ecuaciones y despejar la función de onda Ψ_e para obtener información importante sobre los sistemas cuánticos, es suficiente conocer la densidad electrónica $\rho(r)$ y a partir de esta muchas propiedades del material pueden ser estimadas mediante distintos cálculos. Observando la ecuación de Schrödinger, se concluye que la energía total del sistema puede expresarse mediante la siguiente forma funcional,

$$E[\rho] = T[\rho] + V_{ext}[\rho] + V_{ee}[\rho] \tag{7}$$

Para obtener la energía, la teoría de Kohn y Sham es impuesta. Según esta la densidad electrónica puede describirse como la suma de las funciones de onda de N electrones no interactuantes,

$$\rho(r) = \sum_{i}^{N} |\phi_i|^2 \tag{8}$$

De ahí, la energía se obtiene como,

$$E[\rho] = T_s[\rho] + V_{ext}[\rho] + V_H[\rho] + E_{xc} \tag{9}$$

El cálculo de los primeros tres términos (energía cinética, energía de interacción con el núcleo y energía de interacción electrón-electrón) puede ser realizado con tanto exactitud como se quiera. Pero el último, el funcional de correlación e intercambio (exchange-correlation functional) (E_{xc}) no tiene una forma exacta. Los distintos métodos de DFT proveen distintas aproximaciones de E_{xc}, las cuales difieren en precisión, coste computacional y aplicabilidad a distintos sistemas físicos. En nuestro caso, para los cálculos de DFT, el VASP [15] (Vienna

ab initio Simulation Package, por sus siglas en inglés) será utilizado. Las propiedades de equilibrio para el CO obtenidas mediante VASP pueden apreciarse en la **Tabla 4**.

Tabla 4: Propiedades de equilibrio del CO con cálculos DFT

Molécula	d_0 $[\mathring{A}]$	E_m $[eV]$
CO	1.1435	-14.7915

4.2. Modelos de ML para el cálculo de la energía

Como se expresó anteriormente, el objetivo de este trabajo es desarrollar un algoritmo de ML capaz de calcular la energía de interacción entre una molécula y una nanopartícula metálica. La ventaja de la utilización de un modelo de ML sobre otros métodos de cálculo, es la posibilidad de obtener un buen compromiso entre precisión y esfuerzo computacional.

La premisa fundamental detrás de la implementación de un potencial interatómico de ML es la suposición de que la energía del sistema puede ser descrita como la suma de las energías de cada uno de los átomos que lo componen, esto se puede expresar como,

$$E = \sum_i E_i \tag{10}$$

Para calcular la contribución de cada átomo a la energía total regresiones lineales serán implementadas. Se escogieron regresiones lineales debido a que son la clase más simple de modelos de ML, en consecuencia, se espera que de este modo la transferabilidad del modelo será mayor que la que podría obtenerse si se optara por descripciones más complejas, como por ejemplo redes neuronales, es decir que se espera una mayor precisión para configuraciones que sean disímiles a las presentes en el conjunto de entrenamiento. Por otro lado, debido a la menor complejidad, también es de esperar que la precisión de los modelos lineales sea menor, por esto un gran esfuerzo deberá ponerse en optimizar el modelo para disminuir el error lo máximo posible.

Para desarrollar al modelo de ML, primero debemos obtener una base de datos, sobre esta el modelo será entrenado y testeado. La base de datos, denominada $\mathcal{D}$, puede definirse como,

$$\mathcal{D} = \{(\mathbf{X^i},\ y_i)\ , i = 1, 2, ..., n\} \tag{11}$$

donde, $\mathbf{X^i}$ son los inputs del sistema (descriptores, como se explica en la sección 2.3.1) e y_i son los outputs, energías de interacción en este caso, las cuales pueden obtenerse mediante LAMMPS o VASP, por último, n es el tamaño de la base de datos.

Una vez que el modelo haya sido entrenado, este será capaz de realizar predicciones de la energía de interacción para cualquier arreglo geométrico posible, es decir, para cualquier $\mathbf{X^k}$ que esté incluido en $\mathcal{D}$ o no.

4.2.1. Selección de los descriptores

Cuando se utiliza un algoritmo de ML para realizar cálculos físicos, no es recomendable utilizar las coordenadas cartesianas de los átomos como input del modelo. Esto se debe a que el output del algoritmo, en este caso la energía de interacción, es una función del valor absoluto del input. Como consecuencia, por ejemplo si se realizara un cambio en el origen de coordenadas del sistema, entonces el modelo produciría una energía de interacción distinta, por lo cual el modelo sería físicamente incorrecto.

Para solucionar este inconveniente funciones de las coordenadas cartesianas de los átomos son utilizadas como inputs, estas funciones que son llamadas descriptores, son elegidas de forma conveniente para que sean capaces de describir el entono de cada átomo mientras que al mismo tiempo posean las propiedades matemáticas necesarias para que el modelo tenga sentido físico: invariancia respecto a traslaciones y/o rotaciones, y respecto al intercambio de átomos de la misma especie (intercambio de índices). Adicionalmente, por motivos computacionales estas deben ser continuas, fácilmente derivables y por último, deben decaer a cero. Dos clases distintas de descriptores serán utilizadas este trabajo, la funciones de simetría centradas en el átomo (Atom-centered symmetry functions o ACSF, por sus siglas en inglés) [16] y los descriptores suaves de la superposición de las posiciones atómicas (Smooth Overlap of Atomic Positions descriptors o SOAP, por sus siglas en inglés) [17]. Ambas serán testeadas para determinar que funciones son más apropiadas para nuestro modelo de ML.

Funciones de simetría centradas en los átomos

El formalismo de las funciones de simetría centradas en los átomos (Atom-centered symmetry functions, o por sus siglas en inglés ACSF), fue descripto por Belher, del siguiente modo. Primero, se calcula a la función de corte (cutoff) como,

$$f(\mathbf{R_{ij}}) = \begin{cases} 0{,}5\,[\,cos(\frac{\pi \cdot R_{ij}}{R_C} + 1)] & \text{if } R_{ij} < R_C \\ 0 & \text{en otro caso} \end{cases} \tag{12}$$

Donde, $\mathbf{R_{ij}}$ es la distancia entre dos átomos y R_c el radio de corte. Esta función decae suavemente a cero y es utilizada en la construcción del resto de funciones.

Las funciones de simetría pueden agruparse en dos categorías, las funciones radiales que están compuestas de términos que involucran a dos cuerpos, y las funciones angulares, formadas por términos de tres cuerpos. Para un cierto átomo i usaremos dos funciones radiales:

$$G_i^1 = \sum_j f_c(\mathbf{R_{ij}}) \tag{13}$$

$$G_i^2 = \sum_j exp(-\eta(R_{ij} - R_s)^2) \cdot f_c(\mathbf{R_{ij}}) \tag{14}$$

donde G^1 representa una suma de funciones de corte de un átomo respecto al resto, y G^2 una suma de funciones gaussianas multiplicadas por la función de corte. Mediante el ajuste

de los parámetros η y R_S se definen distintas esferas alrededor del átomo central.

Mientras tanto, como función angular utilizamos a G^4:

$$\theta_{ijk} = arcos(\frac{\mathbf{R_{ij}} \cdot \mathbf{R_{ik}}}{R_{ij}\, R_{ik}}) \tag{15}$$

$$G_i^4 = 2^{1-\xi} \sum_{j,k\neq i}^{all} (1 + \lambda cos(\theta_{ijk}))^{\xi} exp(-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2).f_c(\mathbf{R_{ij}}).f_c(\mathbf{R_{ik}}).f_c(\mathbf{R_{jk}}) \tag{16}$$

donde, $\lambda = \pm 1$, corriendo el máximo del coseno entre $\theta_{ijk} = 0°$ y $\theta_{ijk} = 180°$ respectivamente. En cambio, el parámetro ξ describe la resolución angular, al incrementar ξ, entonces G^4 decae a cero más velozmente.

Para obtener el vector descriptor, $\mathbf{X^i}$, las funciones G^1, G^2 y G^4 son evaluadas utilizando distintas combinaciones de los parámetros (μ, R_s) y (μ, η, λ), luego los valores resultantes son concatenados en un único vector, así nos aseguramos de describir con precisión los sistemas.

Es importante aclarar que el número de descriptores es idéntico para todos los sistemas molécula-nanopartícula, siempre que el número de especies químicas presentes sea el mismo. Como consecuencia, es posible entrenar un modelo en una nanopartícula y posteriormente utilizarlo para realizar predicciones para otras nanopartículas de distintos tamaños o geometrías. Sin embargo, se deberá explorar si los modelos tienen una transferabilidad lo suficientemente buena como para que este tipo de predicciones sean precisas.

Descriptores suaves de la superposición de las posiciones atómicas

Los descriptores suaves de la superposición de las posiciones atómicas (Smooth Overlap of Atomic Positions descriptors, o por sus siglas en inglés SOAP) también fueron probados como una alternativa a las ACSF. Este tipo de funciones busca representar el entorno de los átomos mediante la utilización de funciones gaussianas que describen la densidad atómica, harmónicos esféricos y funciones radiales.

Primero, la estructura atómica es transformada en un campo de densidad atómica. Para cada especie (Z) presente en el sistema,

$$\rho^Z(\mathbf{r}) = \sum_{i}^{toda\ especie\ Z} exp(-\frac{1}{2\sigma^2} \cdot \|\mathbf{r} - \mathbf{R_i}\|^2) \tag{17}$$

Luego se procede a calcular los coeficientes c_{nlm}^Z, mediante la integración del producto entre la densidad atómica, una función esférica y una radial,

$$c_{nlm}^Z = \iiint_{\mathcal{R}^3} \rho^Z(\mathbf{r})\, Y_{lm}(\theta, \phi)\, g_{nl}(r) dV \tag{18}$$

donde, Y_{nl} son los armónicos esféricos ortonormalizados, y g_{nl} son funciones radiales (en nuestro caso utilizaremos las funciones estándar, gaussianos esféricos). Estas funciones se centran en el o los puntos de interés del sistema, en nuestro caso los átomos de C y O.

Entonces, utilizaremos funciones g_{nl} de la forma,

$$g_{nl} = \sum_{n'}^{n_{max}} \beta_{nn'l}\phi_{n'l}(r) \text{ con,}$$
$$\phi_{n'l}(r) = r^l exp(-\alpha_{nl} r^2) \tag{19}$$

donde $\beta_{nn'l}$ se obtiene de una condición de normalización y α_{nl} de imponer que la función decaiga a cero dentro de cierto radio de corte R_{cut}.

Por último, para construir el vector descriptor $\mathbf{p}$, los coeficientes $p_{nn'l}^{Z_i,Z_j}$ son calculados como,

$$p_{nn'l}^{Z_i,Z_j} = \pi\sqrt{\frac{8}{2l+1}} \sum_m c_{nlm}^{Z_i} \cdot c_{n'lm}^{Z_j} \tag{20}$$

Para construir $\mathbf{p}$, $p_{nn'l}^{Z_i,Z_j}$ es calculado para cada combinación de las especies atómicas i y j, todos los pares de n y n' hasta n_{max}, y los valores de l hasta l_{max}.

En conclusión, para definir al vector descriptor, que será llamado $\mathbf{X^i}$ por consistencia, deberemos elegir n_{max} y l_{max}, cuanto mayor sea su valor más precisa será la descripción del entorno atómico, pero a la vez el vector descriptor será más largo y más costoso de calcular. A la vez, también deben decidirse los valores de σ y R_{cut}, que determinan el rango de las interacciones y el decaimiento de la densidad atomica.

En la siguiente tabla, un resumen de los parámetros con los que cuenta cada familia de descriptores es presentado.

Tabla 5: Parámetros para el cálculo de los descriptores

Modelo	Parámetros
Atom-centered symmetry functions (ACSF)	R_{cut} : Global μ, R_S para G^2 μ, η t λ para G^4
Smooth Overlap of Atomic Positions descriptors (SOAP)	n_{max} y l_{max} para las funciones esféricas σ y R_{cut} para la precisión

Luego de calcular los descriptores de cierto sistema, mediante SOAP o ACSF, se aplica una normalización. Esto se debe a que para ciertos modelos de ML es recomendable contar con un input normalizado. Entonces, el vector descriptor resultante para cada sistema molécula-nanopartícula (i),se puede expresar finalmente como,

$$\mathbf{X^i} = (x_1^i,\ x_2^i,\ ...,\ x_\gamma^i)^t \tag{21}$$

donde, $\|\mathbf{X^i}\|^2 = 1$, y $\gamma = N_{descriptors}$ es un valor fijo dependiente del modelo y parámetros empleados.

4.2.2. Modelos de Machine Learning lineales

Con el objetivo de desarrollar un modelo que sea fácil de interpretar y pueda presentar cierto grado de transferabilidad se optó por utilizar regresiones lineales. Para determinar que tipo de modelo se desempeña mejor en los sistemas molécula-nanopartícula utilizados, distintas clases de regresiones lineales fueron probadas. [18]

Una regresión lineal es un método para obtener una estimación de un output (energías de binding y_i en nuestro caso) a partir de un input (vector de descriptores del sistema $\mathbf{X^i} = (x_1^i, x_2^i, ..., x_\gamma^i)^t$), suponiendo una relación lineal entre las dos variables. Esto se puede expresar del siguiente modo,

$$\widehat{y}^i = \mu(\mathbf{X^i}) + e^i = w_0 + \sum_{j=1}^{\gamma} w_j \, x_j^i + e^i \tag{22}$$

donde, μ es la función que asocia a inputs y outputs; $w_0, w_1, w_2, ..., w_\gamma$ son parámetros de ajuste desconocidos y e^i es una variable aleatoria inobservable (el componente de error).

Partiendo de una base de datos $\mathcal{D}$, podemos obtener la mejor estimación de μ, $\widehat{\mu}$, de acuerdo a cierto criterio. La condición más común para obtener a $\widehat{\mu}$ es la minimización del error cuadrático medio (MSE, por sus siglas en inglés). Definimos a $\mathcal{L}$, el conjunto de entrenamiento o de aprendizaje , $\mathcal{L} \subseteq \mathcal{D}$, como,

$$\mathcal{L} = \{(\mathbf{X_i}, \, y_i) \, , i = 1, 2, ..., r, r \leq n\} \tag{23}$$

y definimos al estimador ordinario de cuadrados mínimos (OLS, por sus siglas en inglés), como el que minimiza el MSE para el conjunto de aprendizaje,

$$\min_w \|\widehat{\mathbf{y}}_\mathbf{L} - \mathbf{y}\|_\mathbf{2}^\mathbf{2} = \min_w \|\mathcal{X} \cdot \mathbf{w} - \mathbf{y}\|_\mathbf{2}^\mathbf{2} \tag{24}$$

donde $\mathbf{w} = (w_0, w_1, ..., w_\gamma)^t \, , \mathbf{y} = (y_1, y_2, ..., y_r)^t$, y $\mathcal{X} = \left(\begin{array}{c|c|c|c} 1 & 1 & \cdots & 1 \\ \mathbf{X^1} & \mathbf{X^2} & \cdots & \mathbf{X^r} \end{array}\right)$

De esto, se puede obtener analíticamente nuestro estimador, el cual se llamará estimador lineal clásico,

$$\widehat{y}_L = \tilde{w_0} + \sum_{j=1}^{\gamma} \tilde{w_j} \, x_j^i \tag{25}$$

Este nos permite realizar predicciones para inputs, $\mathbf{X^{NEW}}$, que no se encontraban originalmente en $\mathcal{D}$ y cuya energía de interacción y^{NEW} es desconocida,

$$\widehat{y}^{NEW} = \widehat{y}_L(\mathbf{X^{NEW}}) = \tilde{w_0} + \sum_{j=1}^{\gamma} \tilde{w_j} \, x_j^{NEW} \tag{26}$$

Una vez que la regresión fue ajustada, es deseable contar con una estimación del error cometido al predecir las energías de binding. La forma ideal de cuantificar el error en las

predicciones es mediante la creación de un subconjunto de la base de datos, el conjunto de prueba, $\mathcal{T} \subseteq \mathcal{D}, \mathcal{T} \cap \mathcal{L} = \emptyset$. A partir de este conjunto, el error en la predicción (PE) puede ser estimado, por ejemplo mediante el cálculo del error medio absoluto (MAE, por sus siglas en inglés) para las predicciones de las energías de $\mathcal{T}$,

$$\widehat{PE}(\widehat{\mu}, \mathcal{T}) = MAE = \frac{1}{t} \cdot \sum_{(\mathbf{X^i}, y^i) \in \mathcal{T}} |\widehat{y}_L(\mathbf{X^i}) - y^i| \tag{27}$$

donde t es el tamaño del conjunto de prueba.

Es importante remarcar que esta metodología para estimar el PE, aunque recomendada, es solo posible en los casos donde $\mathcal{D}$ es lo suficientemente grande como para que se pueda extraer una fracción de los puntos para obtener $\mathcal{T}$ sin que esto afecte la calidad de las predicciones. Para los casos donde $\mathcal{D}$ es muy pequeño otras metodologías pueden ser empleadas.

Regresión Ridge

Siguiendo con los modelos lineales, en la regresión Ridge [18], la relación entre descriptores y energía se supone como lineal, pero el objetivo es minimizar una función que contiene una penalización por el valor absoluto de la suma de los cuadrados de los coeficientes de ajuste (norma l2),

$$\min_w \ \left[\|\widehat{\mathbf{y}}_{\mathbf{Rd}} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|_2^2 \right] \tag{28}$$

donde α es un parámetro que determina el peso de la penalización.

La ventaja de la regresión Ridge es que promueve la reducción del valor de los coeficientes que están asociados a descriptores que son poco influyentes, como consecuencia, la regresión Ridge es más robusta, y menos propensa al overfitting o a la divergencia en los casos en donde la relación entre inputs y outputs es difícil de modelizar. Además, la mayor cuantificación del efecto que los descriptores tienen en el valor de energía nos permite tener una mayor perspectiva del comportamiento físico de los sistemas.

Una diferencia importante entre la regresión Ridge y la de cuadrados mínimos es la presencia del parámetro de penalización, α, cuando $\alpha = 0$ la regresión de Ridge es equivalente a la OLS, pero a medida que se incrementa el valor de α el modelo esta menos incentivado a ajustar los valores de energía y por el contrario la influencia del valor de la norma de los coeficientes aumenta. Por este motivo es fundamental definir un valor apropiado de α para la precisión del modelo sea satisfactoria. La forma óptima de definir el valor de α (hiperparámetro del modelo) es ajustar las predicciones utilizando un subconjunto de $\mathcal{D}$ que se denomina conjunto de validación , $\mathcal{V}$. Donde $\mathcal{V}, \mathcal{L}$ y $\mathcal{T}$ son disjuntos y $\mathcal{D} = \mathcal{L} \cup \mathcal{T} \cup \mathcal{V}$. De nuevo, si el tamaño de la base de datos no es lo suficientemente grande otra metodología debe emplearse para seleccionar el valor del hiperparámetro.

Regresión Lasso

De forma similar a la regresión Ridge, la regresión Lasso [18] considera un término de penalización por el valor de los coeficientes, la diferencia es que en este caso el término cuenta

con la suma del valor absoluto de los coeficientes (norma l1),

$$\min_w \left[\frac{1}{2\mathrm{r}} \|\widehat{\mathbf{y}}_{\mathbf{LL}} - \mathbf{y}\|^2 + \alpha \|\mathbf{w}\|_1 \right] \tag{29}$$

Aunque la definicion de las regresiones Ridge y Lasso es similar, estas poseen dos grandes diferencias. La primera es que debido a la forma en que la regresión Lasso es definida no existe un método analítico para despejar los coeficientes, en consecuencia, se debe implementar un algoritmo para obtenerlos. En nuestro caso, la regresión de ángulo mínimo (LARS, por sus siglas en inglés) es empleada, la cual es altamente eficiente y escalable. Debido a la utilización del algoritmo LARS, en lo remanente del texto este modelo será llamado regresión Lasso-Lars.

La segunda diferencia entre el modelo Ridge y el Lasso-Lars es que en el segundo no solo se disminuye el valor de los coeficientes asociados a descriptores de baja relevancia, sino que el modelo esta incentivado a asignar el valor de 0 en los casos donde la influencia de los descriptores sea insignificante. Esto es muy útil, debido a que si se utilizan demasiados descriptores es más probable incurrir en overfitting, por el contrario, si demasiados pocos inputs son utilizados en las predicciones el modelo será incapaz de representar la forma en que inputs y outputs están correlacionados. De nuevo, una gran atención debe ponerse en elegir un valor de α adecuado.

Regresión Elastic Net

Por último, la regresión Elastic Net [18] también será considerada. En este caso, la regularización cuenta con una mezcla de los términos utilizados para la regresión Lasso y Ridge.

$$\min_w \left[\frac{1}{2\mathrm{r}} \|\widehat{\mathbf{y}}_{\mathbf{EN}} - \mathbf{y}\| \; + \; \mathrm{l}_{\mathrm{ratio}} \cdot \alpha \|\mathbf{w}\|_1 \; + \; (1 - \mathrm{l}_{\mathrm{ratio}}) \cdot \alpha \|\mathbf{w}\|_2^2 \right] \tag{30}$$

donde, l_{ratio} tiene un valor entre 0 y 1, y controla cuanto el modelo se asemeja a Lasso ($l_{ratio} = 1$) o Ridge ($l_{ratio} = 0$). En nuestro caso se utilizara un valor intermedio, con $l_{ratio} = 0{,}5$.

Las diferencias y similitudes entre los cuatro tipos de regresiones exploradas pueden ser observadas geométricamente. En este sentido, el primer término de la minimización define un elipsoide (en rigor hiperelipsoide) centrado en el punto definido por la regresión OLS (en el espacio de los coeficientes), mientras que las diferentes regresiones, Ridge, Lasso y Elastic Net, arrojarán como resultado un vector de coeficientes, $\mathbf{w}$, dado por la intersección de la hipersuperficie que es definida por la noma que cada regresión emplea (l1, l2, l1 + l2), con el elipsoide anteriormente mencionado. Esto puede observarse para un caso simple bidimensional en la **Fig. 3**.

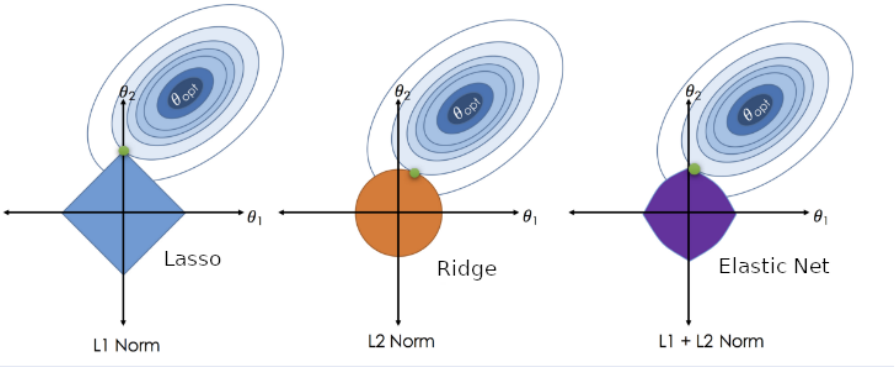


Figura 2: Diagrama del comportamiento de cada una de las regresiones empleadas.

Regresiones Bi-lineales

Dada la simplicidad de los modelos lineales, en algunos casos estos son algebraicamente incapaces de reproducir la complejidad de las interacciones presentes en los sistemas físicos. Una forma posible de incrementar artificialmente la complejidad de los modelos, conservando la linealidad de los mismos, es introduciendo términos cuadráticos como inputs.

La longitud del vector descriptor es γ, si quisiéramos introducir términos cuadráticos al input, entonces, podemos multiplicar cada elemento x_i por otro x_j, de esto obtenemos $\frac{\gamma(\gamma+1)}{2}$ nuevos descriptores (solo una de las posibles permutaciones entre i y j es considerada). Al agregar estos nuevos términos al vector original, obtenemos un nuevo vector de longitud $\nu = \gamma + \frac{\gamma(\gamma+1)}{2}$ $(\nu \gg \gamma)$.

Para evitar incurrir en costos computacionales excesivamente altos durante la evaluación de los modelos Bi-lineales, se optó por aplicar un proceso de filtrado a los descriptores que serán utilizados en la generación del conjunto extendido. Por este motivo, solo los descriptores asociados a coeficientes que sean distintos de cero luego de un ajuste mediante la regresión Lasso-Lars serán empleados.

En resumen, partiendo de un vector descriptor $\mathbf{X^i}$, definimos a $\mathbf{K^i}$ donde,

$$\mathbf{K^i} = (x_1^i,\ x_2^i,\ ,\ ...,x_\omega^i)^t \ \text{, con}\ \ x_l^i \mid w_l^{LL} \neq 0 \,\forall\, l\ ,\ \omega \leq \gamma$$

(donde w_l^{LL} son los coeficientes para la regresión Lasso-Lars), luego definimos,

$$\mathbf{K^{2,i}} = (x_1^i \cdot x_1^i,\ x_2^i \cdot x_1^i,\ \dots\ ,\ x_\omega^i \cdot x_1^i\ ,\ x_2^i \cdot x_2^i\ ,\dots\ ,x_\omega^i \cdot x_\omega^i)^t \tag{31}$$

por último, se define al conjunto extendido como,

$$\mathbf{K^{BL,i}} = \mathbf{K^i} \cup \mathbf{K^{2,i}}$$

Con este conjunto extendido como input cualquiera de las regresiones anteriormente descriptas puede utilizarse, dando lugar a predicciones de la forma,

$$\begin{gathered}\widehat{y}^i_{BL} = \mathbf{K^{BL,i}}.\mathbf{w} + w_0 \\ \widehat{y}^i_{BL} = w_0 + w_1 x^i_1 + ... + w_\omega x^i_\omega + w_{\omega+1}(x^i_1 \cdot x^i_1) + ... + w_\upsilon (x^i_\omega \cdot x^i_\omega) \\ \text{con } \upsilon = \omega + \frac{\omega(\omega+1)}{2}\end{gathered} \tag{32}$$

4.3. Posicionamiento de la molécula gaseosa

Para obtener información respecto a como la nanopartícula interactúa con las moléculas, deberemos generar una multitud de sistemas en los que la molécula se sitúe en distintas ubicaciones con respecto a la nanopartícula. Para una molécula lineal de dos o más átomos (por ejemplo, CO o CO_2), el máximo número de grados de libertad que el sistema puede tener es cinco, tres correspondientes al desplazamiento del centro de la molécula (elegido como el centro del átomo de C) desde el origen de coordenadas (que fue elegido como el centro de la nanopartícula), y dos ángulos de rotación del átomo/s de O con respecto al átomo de C.

Entonces, el sistema formado por una nanopartícula y una molécula puede ser descripto por $Df = (x_C, y_C, z_C, \theta_m, \phi_m)$, donde θ_m y ϕ_m, corresponden a las coordenadas esféricas según la convención mostrada en la **Fig. 2**, y describen la rotación de la molécula con respecto a su centro, elegido como el átomo de C. Análogamente, Df pude definirse como $Df = (r_C, \theta_C, \phi_C, \theta_m, \phi_m)$, donde r_C, θ_C y ϕ_C describe la posición del átomo de C en la convención para las coordenadas esféricas de la **Fig. 2**.

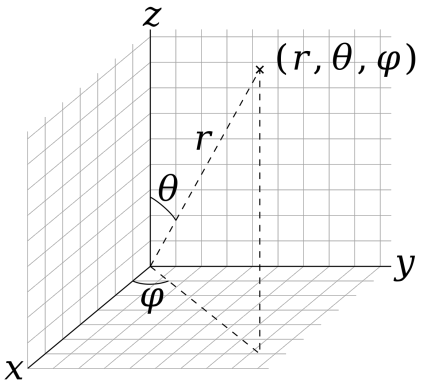


Figura 3: Descripción geométrica de las coordenadas esféricas.

Para poder obtener un conjunto de entrenamiento para el modelo de ML, se deberá desarrollar un algoritmo para obtener conjuntos de Df que contengan suficiente información respecto a las posibles interacciones físicas entre la nanopartícula y la molécula. Al mismo tiempo, estos conjuntos deberán ser lo suficientemente chicos como para que sean computacionalmente obtenibles y se evite el "overfitting", caso en el cual el modelo de ML está sobre-ajustado a los datos en el conjunto de entrenamiento, por lo cual comienza a capturar peculiaridades del conjunto en vez de capturar el comportamiento general y en consecuencia

pierde la capacidad de predicción para datos con los cuales el modelo no fue entrenado. Adicionalmente, cuando se posicione la molécula gaseosa se debe procurar no generar sistemas que sean "no físicos", en consecuencia, deberá evitarse la superposición atómica.

Pueden definirse muchos algoritmos para cumplir con estos objetivos, en este trabajo se probaran dos enfoques distintos, luego su rendimiento será comparado para poder así elegir al que sea más apto y utilizarlo para la obtención de los conjuntos de entrenamiento que contengan cálculos de DFT (como se puede ver en la sección 4.2). En lo sucesivo ambos enfoques son descriptos.

4.3.1. Posicionamiento aleatorio

El modo más simple de generar múltiples desplazamientos y rotaciones de la molécula gaseosa es mediante una asignación aleatoria de Df.

Para determinar los ángulos de rotación , θ_m/θ_C y ϕ_m/ϕ_C, utilizaremos un algoritmo que garantice una distribución uniforme en coordenadas esféricas. La deducción de dicho algoritmo se puede ver en [19]. Allí se concluye que una distribución homogénea se obtiene mediante las siguientes fórmulas,

$$\phi = 2 \cdot \pi \cdot x_1 \tag{33}$$

$$\theta = arcos(2 \cdot x_2 - 1) \tag{34}$$

donde x_1 y x_2 son variables aleatorias uniformes entre 0 y 1.

Para la definición de la posición radial del átomo de C, debemos considerar que nuestro modelo debe ser adaptable a múltiples nanopartículas y que estas pueden presentar complejas geometrías no esféricas. Para lidiar con este desafio, se decidió determinar las posiciones del siguiente modo:

Primero, el centro de la nanopartícula, $\overline{x}$, es determinado promediando la posición de cada átomo en esta. Luego la distancia media de los átomos al centro, $\overline{r}$, es calculada.

A continuación, r_C se define como la suma de $\overline{r}$ y un desplazamiento aleatorio, δ, que es obtenido mediante una distribución aleatoria uniforme, cuyo limite inferior es elegido para evitar generar excesivas configuraciones con superposiciones atómicas y su limite superior lo suficientemente grande como para que la energía de interacción sea despreciable. Una posible mejoría sería utilizar una distribución no uniforme, que presentase mayor densidad de probabilidad en las regiones de energía más negativa, donde más interés tenemos en capturar con precisión el comportamiento de las interacciones físicas.

Por último, para evitar la superposición atómica se define un radio de superposición, r_{cut}. Al generar un sistema se procede a comprobar si algún átomo de la nanopartícula se encuentra a una distancia menor a r_{cut} de algún átomo de la molécula, en caso de que esto suceda este sistema es considerado 'no físico' y se procede a generar otro.

4.3.2. Malla más Muestreo Informado

Dado que nuestro objetivo es lograr una mayor precisión del modelo de ML para las configuraciones en el entorno de los sitios de adsorción, donde la energía potencial presenta mínimos locales, y debido a la naturaleza asimétrica de las nanopartículas, resulta complejo describir una distribución aleatoria que presente mayor densidad de probabilidad en las regiones de mayor interés para cada combinación de θ_C y ϕ_C. Como consecuencia, se deberá optar por un enfoque alternativo [20].

La principal desventaja de la utilización de una distribución aleatoria uniforme para la generación de conjuntos de entrenamiento es que, si el intervalo radial muestreado es muy angosto entonces el modelo será incapaz de reproducir el comportamiento físico a largas distancias (la energía de interacción debería tender a cero), por el contrario, si el intervalo muestreado es demasiado amplio la calidad del entrenamiento disminuye, dado que solo una pequeña fracción de los puntos en el conjunto de entrenamiento se encuentra en la región de baja energía.

Con el objetivo de sobrellevar ambas dificultades, se optó por definir una metodología mixta. En este enfoque, en primera instancia, mediante la generación de una malla de puntos se obtiene una descripción general de cómo varía la energía a corto y largo alcance. La malla está compuesta de puntos equiespaciados a lo largo de ciertas direcciones, de los puntos que están lejos de la superficie se procura obtener la información necesaria para poder entrenar al modelo para describir las interacciones a largo rango. Al mismo tiempo, de los puntos cerca de la superficie, la posición aproximada del mínimo de energía para esa dirección puede ser obtenida. Luego, se puede obtener una superficie aproximada de mínima energía, interpolando entre los puntos de mínima energía obtenidos para cada dirección. Por último, se realiza un muestreo informado, mediante la generación de puntos aleatorios en la cercanía de esta superficie. El algoritmo se puede describir del siguiente modo:

Primero, se genera la malla de puntos, esta está compuesta de M lineas, cada cual posee una combinación distinta de los ángulos θ_C y ϕ_C. A su vez, cada línea cuenta con N puntos equiespaciados en la dirección radial (r_C). Con respecto a la rotación de la molécula, θ_m y ϕ_m, deben ser fijados según algún criterio para disminuir a 3 los grados de libertad de la malla, esta decisión tendrá un efecto en los valores de energía obtenidos, pero de otro modo, si se tuvieran en cuenta rotaciones en la malla el numero de puntos a evaluar resultaría demasiado alto. Las posiciones en la malla pueden escribirse del siguiente modo,

$\{Df_M\} = \{Df_1^{\alpha_1}, Df_2^{\alpha_1}, ..., Df_N^{\alpha_1}, Df_1^{\alpha_2}, Df_2^{\alpha_2}, ..., Df_N^{\alpha_2}, Df_1^{\alpha_M}, ..., Df_N^{\alpha_M}\}$

donde, α_i corresponde a una combinación única de los ángulos θ_C, ϕ_C y el subíndice está asociado a los distintos valores de r_C a lo largo de la linea.

Se tomó la decisión de definir θ_C y ϕ_C de modo que el átomo de C siempre estuviese orientado hacia la superficie de la nanopartícula. Esta elección se fundamenta en ciertos ensayos, en los cuales se halló que esta configuración presenta en general baja energía y en consecuencia es más relevante. En total hay $N \times M$ puntos en la malla.

En segundo lugar, para cada línea se halla el radio correspondiente al mínimo valor de energía en esa dirección (r_C^{*k}), así se obtiene el conjunto de M puntos, $\{(\theta_C^1, \phi_C^1, r_C^{*1})$, $(\theta_C^2, \phi_C^2, r_C^2), ..(\theta_C^M, \phi_C^M, r_C^{*M})\}$.

En tercer lugar, se interpola entre los puntos $(\theta_C^k, \phi_C^k, r_C^{*k})$ para obtener la superficie que aproxima r_C^* para cada combinación de ángulos θ_C y ϕ_C.

Finalmente, para realizar un muestreo informado se generan combinaciones aleatorias de los grados de libertad angulares $\theta_C, \phi_C, \theta_m$ y ϕ_m, mediante la formula para distribuciones uniformes (sección 2.1.1), para el caso de r_C, se calcula radio correspondiente a θ_C, ϕ_C $(r_C^*(\theta_C, \phi_C))$, y se suma un desplazamiento aleatorio mediante una distribución normal con media $\mu = 0$ y una desviación estándar σ_s apropiada.

$$\phi_m,\ \phi_C\ =\ 2 \cdot \pi \cdot x_1 \tag{35}$$

$$\theta_m,\ \theta_C\ =\ arcos(2 \cdot x_2 - 1) \tag{36}$$

$$r_C\ = r_C^{min}(\theta_C, \phi_C) + x_3 \tag{37}$$

donde x_1 y x_2 son variable aleatorias uniformes entre 0 y 1. Y x_3 es una variable aleatoria normal con media $\mu = 0$ y dev. std. σ_s.

De la ejecución de este algoritmo, un conjunto $\{Df_S\}$ es obtenido (muestreo), compuesto de K muestras. Luego, el conjunto resultante (entrenamiento) es $\{Df_T\} = \{Df_M\} \cup \{Df_S\}$ con $N \times M + K$ puntos.

5. Resultados

[1]Durante la primera parte del trabajo se realizó un proceso extensivo de prueba del modelo de ML utilizando LAMMPS para calcular las energías. Durante esta etapa, los distintos modelos fueron probados con sistemas con uno, tres y cinco grados de libertad. Además, se ajustó el valor del parámetro de penalización α y se evaluó la mejor forma de generar conjuntos de entrenamiento. El objetivo de esta multitud de pruebas fue el de validar la metodología de trabajo, antes de aplicarla para generar una costosa base de datos de energías obtenidas mediante DFT.

Se optó por emplear sistemas nanopartícula-molécula para los cuales se tuviera acceso a experiencia previa de cálculos mediante DFT, por este motivo, se decidió utilizar nanopartículas de oro puro y moléculas de CO. El entrenamiento y testeo del modelo de ML se realizó utilizando un octaedro truncado de 79 átomos, cuyo radio es de aproximadamente 7,2 $\mathring{A}$ (medido en la dirección x), este se puede observar en la **Fig 4**.

[1]Este proyecto está basado en el trabajo realizado por Jona Nägerl durante su pasantía M1 realizada bajo la supervisión de Julien Lam, en la UMET en 2023 [21]. Parte del código desarrollado por él fue tomado como base y modificado a lo largo largo del trabajo, en particular los scripts necesarios para crear los archivos tipo POSCAR que se utilizan como input de los cálculos por VASP, este código utiliza el modulo ovito para python [22]. A partir de sus resultados pudo validarse el correcto funcionamiento de un modelo de ML lineal para la reproducción de las interacciones modelizadas mediante potenciales interatómicos clásicos, a su vez, se pudo observar que el posicionamiento aleatorio posee ciertas limitaciones a la hora de generar conjuntos de entrenamiento.

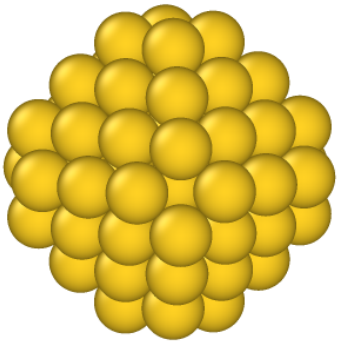


Figura 4: Nanopartícula empleada para el entrenamiento y el testeo del modelo de ML.

Tanto para la nanopartícula, como para la molécula de CO, la energía en el vacío y longitud de enlace de equilibrio, d_{C-O}, fueron calculadas tanto con DFT como con LAMMPS. Dado que el objetivo de la primera parte del trabajo era ensayar la metodología que luego sería utilizada para los cálculos mediante VASP, se optó por emplear para los cálculos por LAMMPS a la longitud de enlace obtenida por DFT. En la **Tabla 6** una comparación de las propiedades en el vacío obtenidas mediante LAMMPS y VASP es presentada.

Tabla 6: Parámetros de equilibrio obtenidos mediante LAMMPS y VASP.

Modelo	$d_{C-O}[\mathring{A}]$	$E_{CO}[eV]$	$E_{TO79}[eV]$
LAMMPS	1,143*	30,19*	-268,66
VASP	1,143	-14,7915	-223,2

* Ver la Tabla 3 para aclaraciones.

5.1. Ensayos con energías obtenidas mediante LAMMPS

5.1.1. Sets de entrenamientos obtenidos mediante muestreo aleatorio y con descriptores ACSF

Pruebas en sistemas unidimensionales

En primer instancia, para establecer el valor que el hiperparámetro α debe tener y para validar el funcionamiento de los modelos de ML, se generaron conjuntos de aprendizaje de distintos tamaños y con un único grado de libertad, movimiento a lo largo del eje x.

Para determinar el valor óptimo de α, los modelos lineales fueron entrenados sobre un conjunto de entrenamiento compuesto de 50 puntos obtenidos mediante muestreo aleatorio. r_C se obtuvo mediante una distribución aleatoria uniforme entre $\bar{r} + 1{,}8\mathring{A}$ y $\bar{r} + 9\mathring{A}$, donde $\bar{r}$ es el radio medio de la nanopartícula, mientras que, el resto de grados de libertad se fijó

como cero. Para el conjunto de testeo, se utilizaron 100 puntos equiespaciados a lo largo de x, con el resto de Df fijado como 0. Para los ajustes lineales, se utilizaron valores de α entre 10^{-18} y 10^{-2}, con incrementos de 100 entre cada valor.

Se observó que para $\alpha = 10^{-8}$ o menor los modelos presentas predicciones muy similares. Se decidió tomar $\alpha = 10^{-10}$ para el resto de los cálculos, dado que para este valor se obtuvo buena precisión y se evitó el overfitting.

En la siguiente figura, se muestra una comparación del MAE que presenta cada uno de los modelos lineales como función del valor de α empleado.

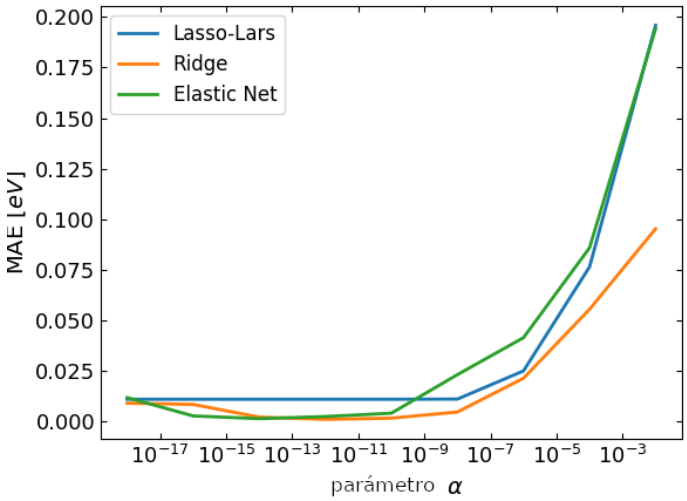


Figura 5: Variación del MAE como función del valor de α para las regresiones Lasso-Lars, Ridge y Elastic Net lineales.

Pruebas en sistemas de cinco dimensiones

Una vez completados los ensayos preliminares, se prosiguió con ensayos sobre sistemas que contaran con los cinco grados de libertad posibles, usando siempre $\alpha = 10^{-10}$. De este modo, no es necesario definir un conjunto de validación $\mathcal{V}$, y en consecuencia, una mayor fracción de $\mathcal{D}$ puede utilizarse para el entrenamiento y testeo del modelo de ML.

Para evaluar el comportamiento de cada una de las regresiones presentadas en la sección 4.3.2, se generaron conjuntos de entrenamiento mediante asignación aleatoria de los elementos en Df. Para esto, r_C se obtuvo de la suma del radio medio de la nanopartícula y un desplazamiento aleatorio uniforme entre 1,44 y 9 $\mathring{A}$. Mientras que, ambos pares de ángulos $(\theta_C, \phi_C$ y $\theta_m, \phi_m)$ se generaron utilizando el algoritmo que asegura distribuciones angulares uniformes (sección 2.1.1.). Para el radio de superposición atómica, r_{cut}, se empleó un valor de 1,80 $\mathring{A}$, ya que es una buena estimación del radio del Au más el del C u O.

Para estudiar la convergencia de los modelos al ser entrenados bajo estas condiciones, se utilizaron conjuntos de entrenamiento de distintos tamaños. También se compararon las

predicciones para distintos conjuntos del mismo tamaño, de este modo puede evaluarse la reproducibilidad al entrenar mediante muestreo aleatorio.

Para el conjunto de testeo, $\mathcal{T}$, se decidió utilizar distintas líneas, con ángulos de rotación fijos y r_C equiespaciado. Por consistencia con los resultados presentados en la secciones anteriores, se decidió orientar al O en la dirección $+z$ y se empleó la línea correspondiente al eje $+x$, pero también se utilizaron otras líneas para disminuir la parcialidad en el testeo del modelo. En la **Tabla 7** se presenta la definición de cada línea empleada en $\mathcal{T}$.

Tabla 7: Definición de las líneas presentes en $\mathcal{T}$

Nombre	θ_C [°]	ϕ_C [°]	[°] θ_m	ϕ_m [°]	# de puntos	r_c^{min} [$\mathring{A}$]
X	90	0	0	0	100	7.6
20 °	90	20	0	0	100	8.0
45-20 °	45	20	0	0	100	6.2

En las siguientes figuras se presenta el comportamiento de cada uno de los modelos, al ser entrenados con cinco conjuntos independientes. Por claridad en las comparaciones, solo se presenta la dirección X, aunque se observa un comportamiento similar para cada una de las líneas en $\mathcal{T}$.

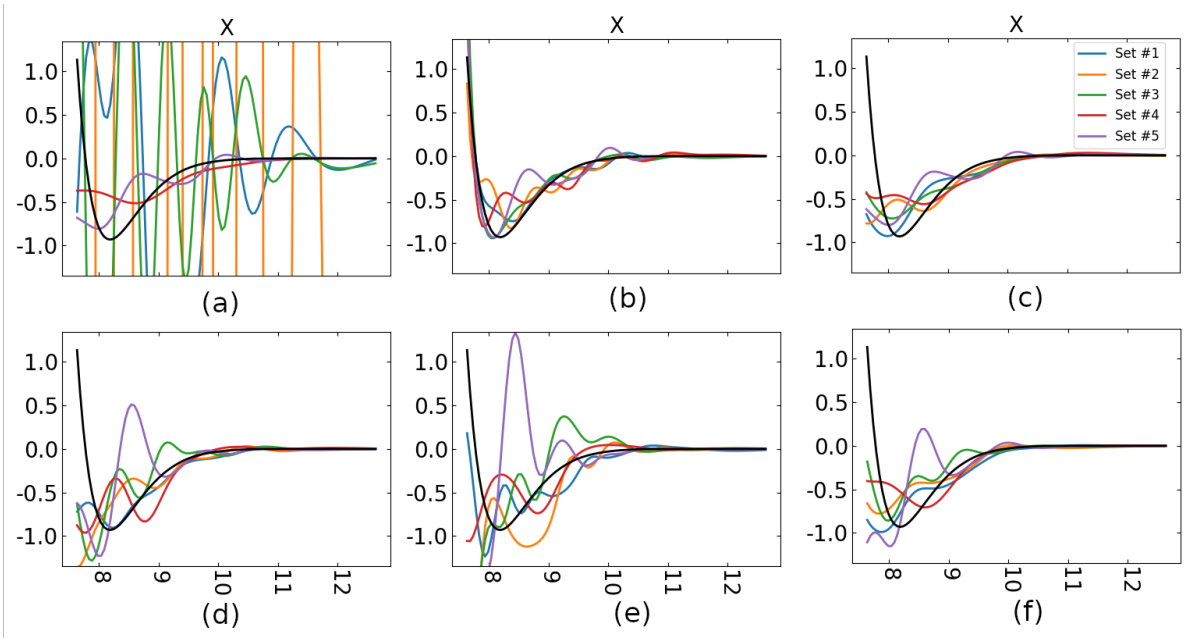


Figura 6: Predicciones de los modelos de ML entrenados con 5 conjuntos independientes generados por muestreo aleatorio de 500 puntos. Para (a) Lasso-Lars, (b) Ridge, (c) Elastic Net, (d) B-L Lasso-Lars, (e) B-L Ridge, (f) B-L Elastic Net. En el eje x se muestra el radio (r_C) en $\mathring{A}$, en el eje y la energía de binding en eV.

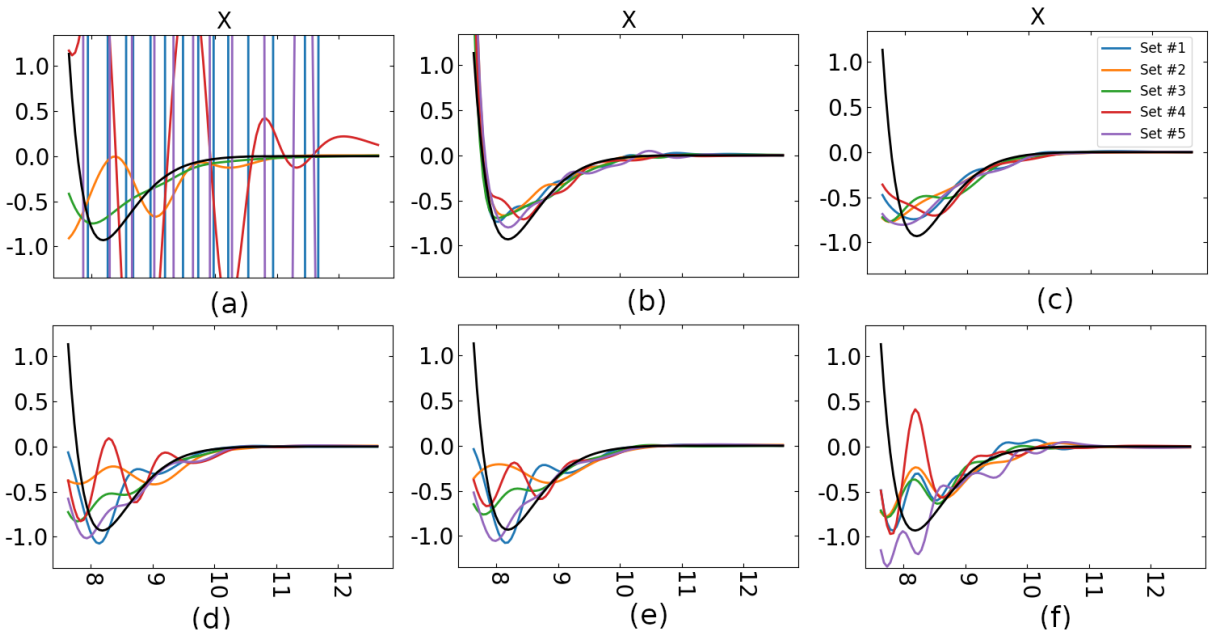


Figura 7: Predicciones de los modelos de ML entrenados con 5 conjuntos independientes generados por muestreo aleatorio de 1.000 puntos.

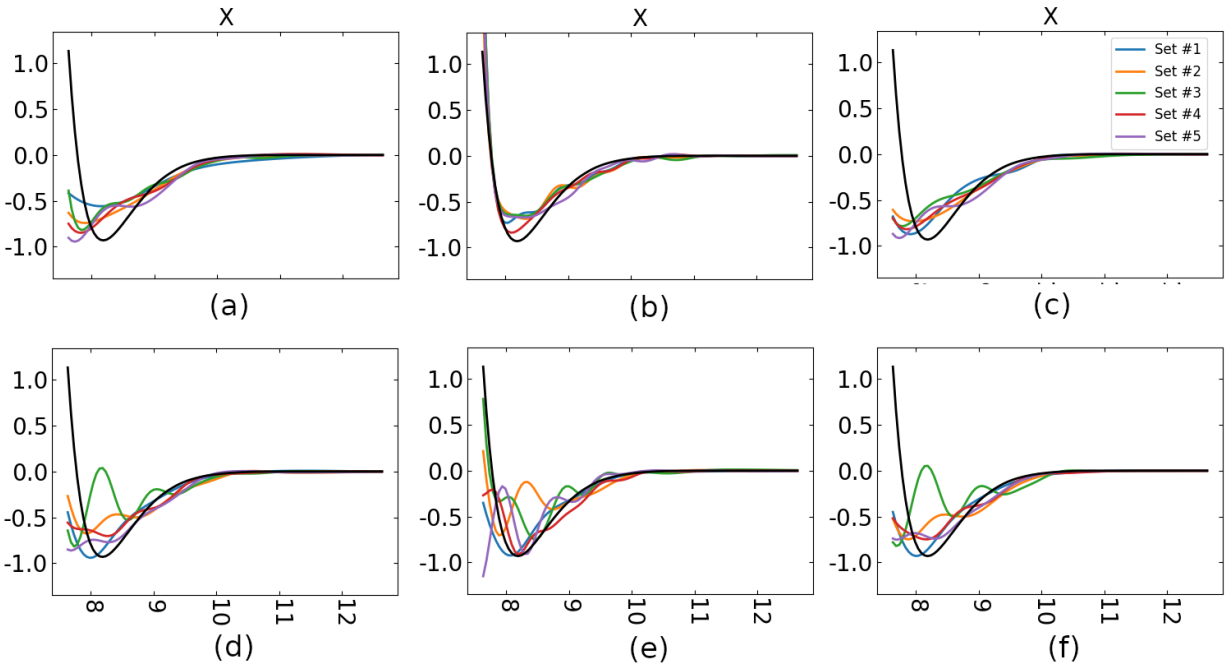


Figura 8: Predicciones de los modelos de ML entrenados con 5 conjuntos independientes generados por muestreo aleatorio de 1.500 puntos.

En la **Fig. 6** se puede observar como ninguno de los modelos genera predicciones razonables,

esto indica que 500 puntos en el conjunto de entrenamiento no es suficiente. A partir de la **Fig. 7** y la **Fig. 8** se observa que modelo Lasso-Lars lineal exhibe divergencia cuando es entrenado sobre 1.000 puntos o menos. Adicionalmente, sin importar el tamaño del conjunto de entrenamiento se puede ver como los modelos Bi-lineales se desempeñan peor que los lineales. Por último, en todos los casos el modelo que mejor reproduce a la curva de referencia es la regresión Ridge lineal. En el resto de esta sección el énfasis se pondrá en este modelo, ya que en general es el más preciso.

5.1.2. Sets de entrenamientos generados mediante mallado más muestreo informado y con descriptores ACSF

Una vez que los modelos fueron evaluados sobre conjuntos de entrenamiento generados por muestreo aleatorio, se decidió ensayar la metodología alternativa para la generación de conjuntos de entrenamiento (**sección 3.1.2**). Las simetrías de la nanopartícula fueron utilizadas para definir la menor región angular que debía ser considerada para obtener una descripción completa del entorno energético. Esta región corresponde a $\theta_C \in [\,0°, 90°]$ y $\phi_C \in [\,0°, 45°]$, como se observa en la **Fig. 9**.

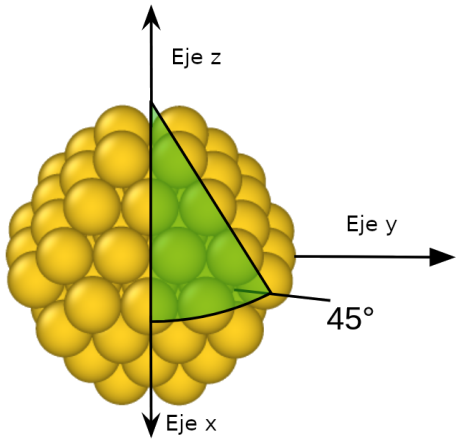


Figura 9: Área utilizada para el muestreo informado.

Para la generación del conjunto correspondiente al mallado, $\{Df_M\}$, se emplearon los ángulos presentados en la **Tabla 8**. Para r_c el primer punto utilizado en cada línea fue más cercano a la superficie para el cual no se observó superposición atómica, luego r_c fue incrementado mediante pasos de 0,2 $\mathring{A}$, hasta obtener 25 puntos en cada línea.

Tabla 8: Ángulos utilizados en la malla

Polar (θ_C , θ_m) [°]	Azimutal (ϕ_C , ϕ_m) [°]	# de puntos
0	0	25
22,5	0; 15; 30; 45	100
45	0; 15; 30; 45	100
67,5	0; 15; 30; 45	100
90	0; 15; 30; 45	100

El resultado es una malla con 425 puntos, correspondientes a 17 direcciones distintas y 25 posiciones radiales por línea. Una vez que se calcula la energía de interacción para cada configuración en la malla, se extrae el radio correspondiente a la energía mínima para cada línea. A partir de esto, se procede a definir la superficie aproximada de mínima energía, interpolando linealmente entre los puntos obtenidos en el espacio (θ_C, ϕ_C, r_C) y forzando las condiciones de contorno necesarias dado que se está trabajando en coordenadas esféricas. En la **Fig. 10.a** se presentan las energías obtenidas para la malla, mientras que en la **Fig. 10.b** se muestra el resultado de la interpolación.

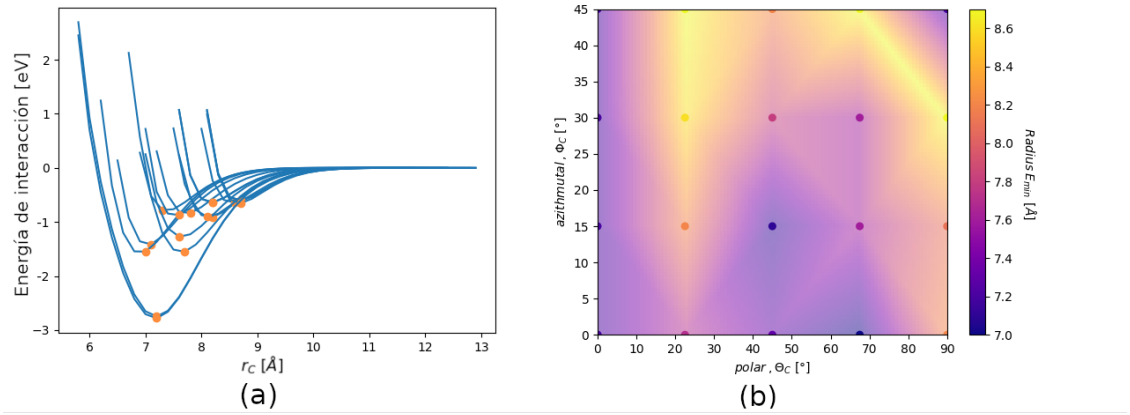


Figura 10: Malla obtenida mediante el potencial interatómico EAM. En (a) se grafica la energía vs. radio para cada una de las lineas en la malla, en (b) los puntos correspondientes a los mínimos en (a) son utilizado para realizar una interpolación y obtener la superficie aproximada de mínima energía, $r_C^{min}(\theta_C, \phi_C)$.

Al entrenar los modelos de ML sobre los 425 puntos de la malla, las predicciones son bastamente imprecisas, esto es razonable debido a la baja cantidad de puntos que se utilizan. Sin embargo, es importante remarcar como los modelos reproducen satisfactoriamente las curvas de referencia, tanto en la cercanía de la superficie como a grandes distancias. En consecuencia, se puede concluir que la malla contiene información suficiente como para que el modelo pueda predecir que la energía debe ir a 0 si la molécula esta muy lejos de la superficie y a infinito si esta muy cerca. En la siguiente figura se presentan los resultados para la regresión Ridge.

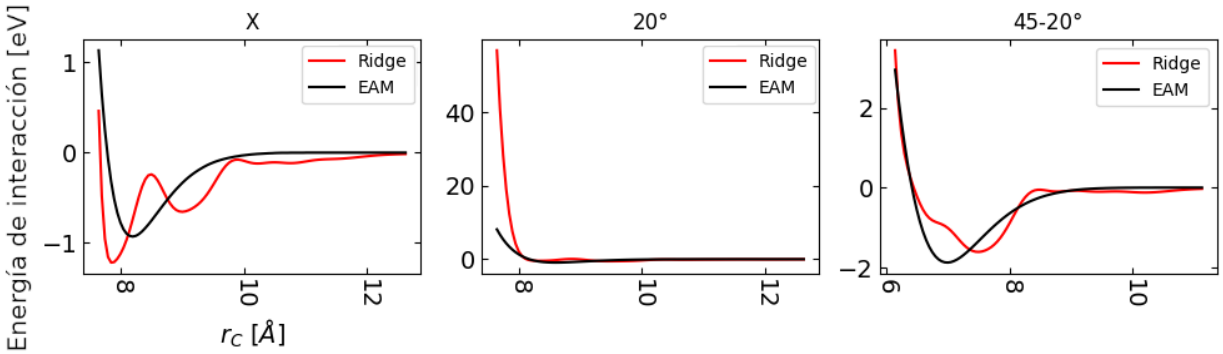


Figura 11: Predicción para la regresión Ridge lineal entrenada con la malla.

Tras encontrar que no es suficiente entrenar al modelo con los puntos de la malla para obtener predicciones adecuadas se procedió a investigar el efecto de incluir en los conjuntos $\mathcal{L}$ cantidades cada vez mayores de puntos generados por muestreo informado.

Para todos los modelos se continuó utilizando $\alpha = 10^{-10}$, r_C fue generado mediante un desplazamiento aleatorio de la superficie aproximada de mínima energía, este se obtuvo mediante una distribución normal. Se investigaron distintos valores de desviación estándar (σ_s) entre 0.01 y 2, y se encontró que la calidad de las predicciones no variaba significativamente, finalmente se optó por utilizar $\sigma_s = 1$.

En las siguientes figuras, se presenta una comparación de las predicciones obtenidas mediante la regresión Ridge, al ser entrenada utilizando conjuntos de tamaño similar y generados mediante ambas técnicas.

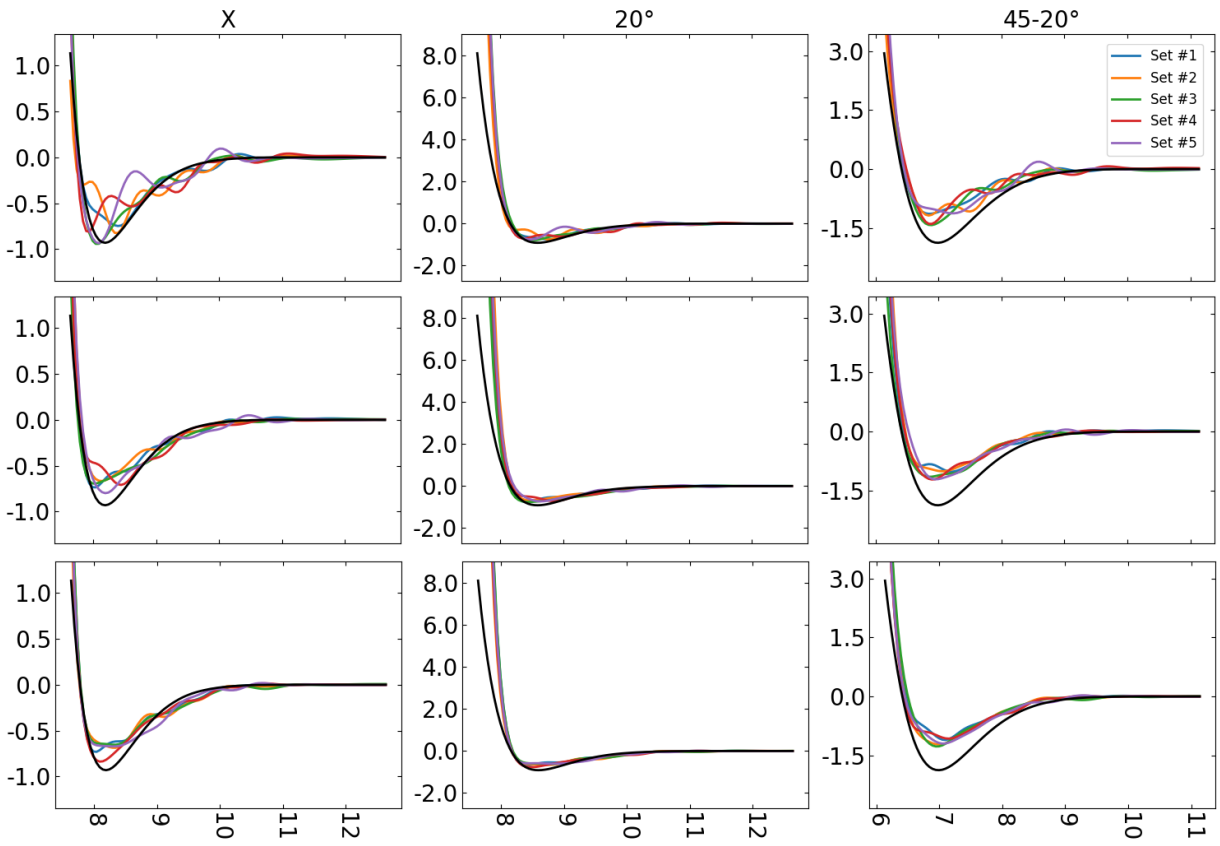


Figura 12: Predicciones para el modelo Ridge entrenado con 5 conjuntos independientes obtenidos mediante muestreo aleatorio, conteniendo de arriba hacia abajo 500, 1.000 y 1.500 puntos. En el eje x se muestra el radio (r_C) en $\mathring{A}$, en el eje y la energía de binding en eV.

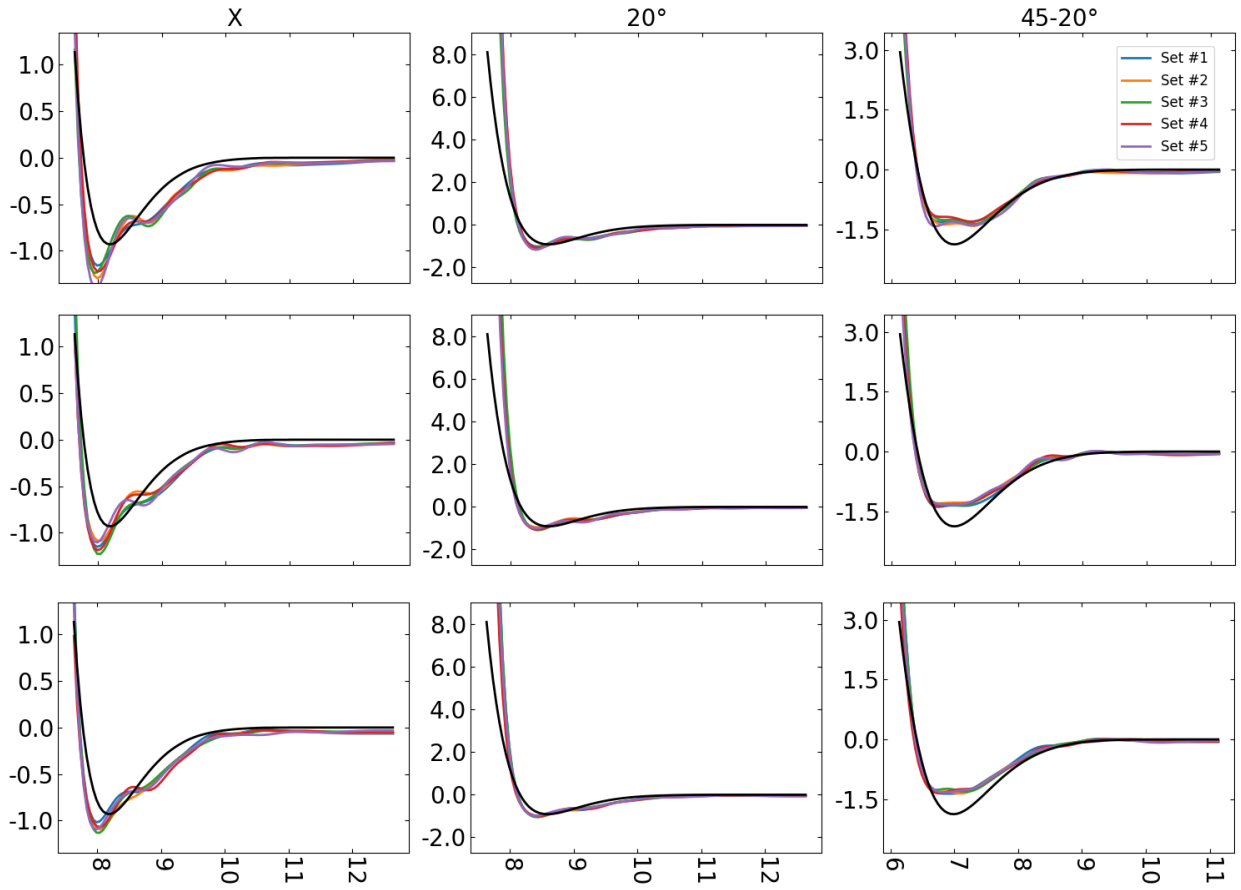


Figura 13: Predicciones para el modelo Ridge entrenado con 5 conjuntos independientes obtenidos mediante mallado más muestreo informado, conteniendo de arriba hacia, la malla más 250, 500 y 750 puntos. El número total de puntos es 675, 925 y 1.175 respectivamente.

Se puede observar que los modelos entrenados con conjuntos obtenidos mediante la técnica de la malla más muestreo informado producen predicciones mucho más consistentes, dado que las curvas obtenidas son más cercanas entres sí. Además, los modelos convergen más velozmente a las curvas de referencia, obtenidas a partir de los cálculos mediante LAMMPS. Sin embargo, debe remarcarse que en algunas direcciones, como por ejemplo la $45-20°$, todos los modelos presentan un mayor error en las predicciones, esto puede deberse a la complejidad del entorno energético en la cercanía de ciertas características (features) de la nanopartícula.

Para observar mejor la precisión de los modelos en la región de mayor interés (energía negativa), se presentan las predicciones únicamente en el intervalos radiales donde la energía calculada mediante LAMMPS es negativa. En la **Fig. 14** se muestra los resultados del modelo con mejor desempeño, Ridge entrenado con la malla más 5.000 puntos. Una comparación completa de las predicciones arrojadas por Ridge para conjuntos $\mathcal{L}$ compuestos por la malla más 250, 500, 1.000, 5.000 y 10.000 puntos se puede observar en la **Fig. A** en el apéndice.

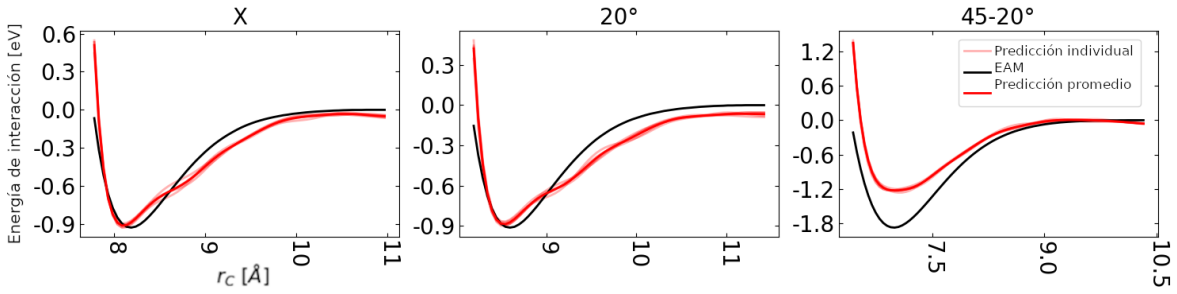


Figura 14: Predicciones para la regresión Ridge entrenada con cinco conjuntos independientes formados por la malla más 5.000 puntos obtenidos por más muestreo informado

Se puede observar que entre 5.000 y 10.000 puntos el modelo logra la mayor consistencia. Por otro lado, la mayor precisión se logra para 1.000 puntos. También se ve, que el modelo es incapaz de reproducir completamente la forma de las curvas de referencia, especialmente para los ángulos $45 - 20°$. En esta dirección, se observa una clara discrepancia en el valor mínimo de energía, sin embargo, la posición radial del mínimo es similar en ambos casos.

Para obtener una estimación del error del modelo el MAE fue calculado. Sin embargo, esta no es una forma óptima de juzgar que tan buenas son las predicciones, debido a que lo que buscamos es mayor precisión en el entorno del mínimo de energía y una reproducción cualitativa del comportamiento que presenta el EAM a corto y largo rango. Además, cuando la molécula esta muy cerca de la superficie la energía toma valores muy elevados, por esto, existe una diferencia de orden de magnitud para los puntos en distintas regiones. Por este motivo, el error en los puntos de alta energía puede " esconder" al error en las regiones de interés.

Una forma de evitar estos errores en la cuantificación del error es calcular el MAE solo en los puntos de $\mathcal{T}$ con energía negativa, así podemos cuantificar el error solo en el área donde deseamos alta precisión. En la **Tabla 9** el MAE promedio obtenido para los puntos negativos en $\mathcal{T}$, todos los puntos en $\mathcal{T}$ y todos los puntos en $\mathcal{L}$ es presentado.

Tabla 9: Comparación del MAE para la regresión Ridge entrenada mediante la malla más muestreo informado y energías calculadas con el potencial interatómico EAM.

# Puntos muestreados	MAE: $\mathcal{T}$ $[eV]$ (solo negativo)	MAE: $\mathcal{T}$ $[eV]$	MAE: $\mathcal{L}$ $[eV]$
250	0,1421	0,5040	0,1775
500	0,1323	0,479	0,2015
750	0,1258	0,4184	0,2014
1.000	0,1194	0,5504	0,234
5.000	0,1350	1,189	0,2339
10.000	0,1673	1,673	0,2304

Las tendencias observadas son las siguientes, para $\mathcal{L}$ el MAE incrementa con el número de puntos muestreados, esto indica que a medida que el modelo es entrenado con más información

hay un compromiso mayor entre la precisión que cada punto adicional logra. Respecto a todos los puntos en $\mathcal{T}$, se observa que a medida que el tamaño de $\mathcal{L}$ incrementa también lo hace el MAE, esto puede deberse a un mayor error para las predicciones en el corto y/o largo rango. Por último y más importante, al considerar solo los puntos negativos en $\mathcal{T}$, se ve que el MAE mínimo es obtenido para 1.000 puntos generados por muestreo informado, esto probablemente se debe a que en la dirección $45 - 20°$ las predicciones comienzan a desviarse más de la curva de referencia a medida que aumenta el tamaño del conjunto de aprendizaje. De todos modos, este análisis es parcial para determinar en que caso se obtiene el mejor desempeño, dado que solo tiene en cuenta la precisión numérica y no el mejoramiento en la consistencia o suavidad de las curvas obtenidas con el tamaño de $\mathcal{L}$.

5.1.3. Sets de entrenamientos generados mediante mallado más muestreo informado y con descriptores SOAP

Una vez que se exploraron los límites en el desempeño de los modelos utilizando descriptores ACSF, se decidió investigar si era posible obtener una mejora en las predicciones, o una disminución en el número de puntos necesarios en $\mathcal{L}$ al utilizar descriptores SOAP.

Para evaluar el valor que debían tener los parámetros a utilizar en el cálculo de los descriptores, se decidió probar distintas combinaciones de n_{max} y l_{max}, para cada una de estas los modelos fueron entrenados utilizando múltiples combinaciones de σ y r_{cut}, mediante un script de python (listado completo en la **Tabla C** del apéndice). Así, n_{max} y l_{max} fueron secuencialmente aumentados hasta que se obtuvo una precisión suficientemente buena, los valores de parámetros que fueron seleccionados se presentan en la **Tabla 10**.

Tabla 10: Parámetros empleados en los descriptores SOAP para energías calculadas mediante LAMMPS.

n_{max}	l_{max}	σ	r_{cut}
6	4	5.5	0.1

Empleando estos parámetros, se calcularon los descriptores y se entrenó los modelos con la malla descripta en **Tabla 8**. En la siguiente figura pueden observarse las predicciones obtenidas para las regresiones lineales Lasso-Lars y Ridge.

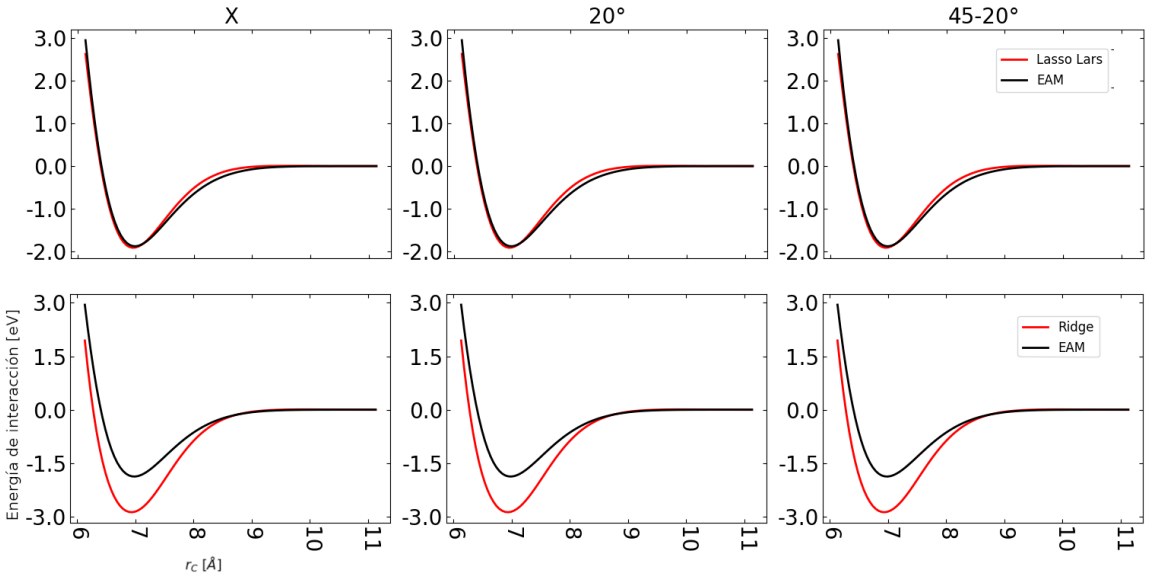


Figura 15: Predicciones de los modelos entrenados sobre una malla para el potencial interatómico EAM. Arriba Lasso-Lars, abajo Ridge.

Se observa que utilizando los descriptores SOAP los modelos son mucho más capaces de describir las interacciones. En particular, la regresión lineal Lasso-Lars produjo predicciones mucho más precisas que el mejor modelo que pudo ser obtenido con descriptores ACSF (malla + 1.000 puntos muestreados). En la siguiente figura, se muestran las predicciones al agregar 250 puntos mediante muestreo informado, para 5 conjuntos independientes.

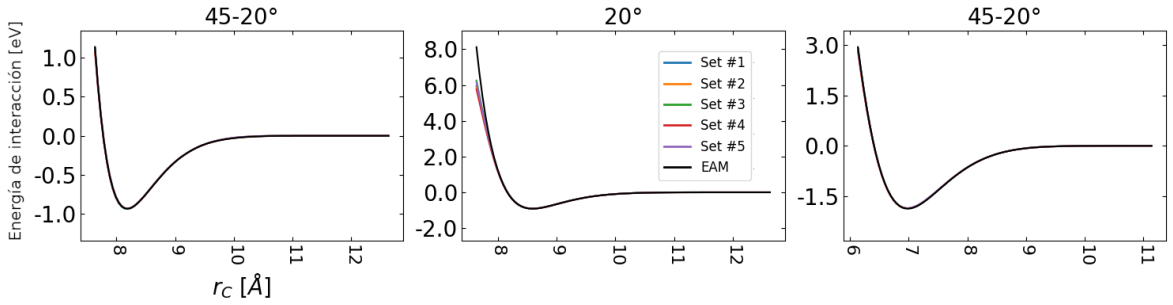


Figura 16: Predicciones para la regresión linear Lasso-Lars entrenada con 5 conjuntos independientes, compuestos por la malla + 250 puntos.

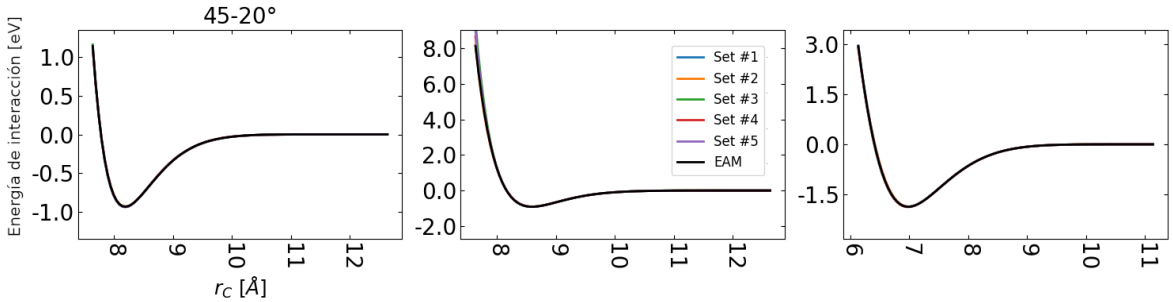


Figura 17: Predicciones para la regresión lineal Ridge entrenada con 5 conjuntos independientes, compuestos por la malla + 250 puntos.

A simple vista ambos modelos producen predicciones casi perfectas, por este motivo el MAE es una herramienta razonable para determinar cual de los dos tiene un mejor desempeño. En la siguiente tabla se presenta el MAE promedio para ambos modelos, tanto para $\mathcal{T}$ como $\mathcal{L}$.

Tabla 11: Comparación del MAE para descriptores SOAP y energías obtenidas por el EAM

# puntos muestreados	MAE: $\mathcal{T}$ $[eV]$ (solo negativo)	MAE: $\mathcal{T}$ $[eV]$	MAE: $\mathcal{L}$ $[eV]$
Ridge			
250	0,0030	0,0087	0,0017
500	0,0027	0,0067	0,0025
750	0,0024	0,0089	0,0030
Lasso-Lars			
250	0,0034	0,0226	0,0048
500	0,0033	0,0210	0,0056
750	0,0032	0,0205	0,0060

Se puede ver que la precisión no mejora significativamente al aumentar el tamaño del conjunto de entrenamiento, por lo cual contar con la malla más 250 puntos es suficiente. Al comparar ambos modelos, se ve que Ridge genera un MAE menor, especialmente para el caso de todos los puntos en $\mathcal{T}$, si solo se consideran los puntos negativos la diferencia entre la precisión de ambos modelos no es substancial.

5.2. Modelos entrenados con energías obtenidas por DFT

A partir de los ensayos realizados en la **sección 6.1**, se estableció que es posible entrenar un modelo lineal de ML capaz de reproducir con precisión el entorno energético que se obtiene al emplear el potencial interatómico EAM. Habiendo demostrado que los mejores resultados se obtienen al utilizar la metodología de la malla más muestreo informado, en conjunto con

descriptores SOAP, se procedió a utilizar estos métodos para entrenar modelos con bases de datos que contuvieran energías obtenidas por cálculos DFT. Para esto se empleó la misma nanopartícula de 79 átomos y molécula de CO.

Los cálculos de DFT fueron realizados utilizando el funcional de intercambio y correlación (exchange-correlation functional) GGA-PW91 y el método de proyector de ondas aumentadas implementado en VASP. Se realizaron distintas pruebas para elegir los parámetros a emplear para los cálculos en VASP, los valores resultantes se presentan en la **Tabla D** en el apéndice.

En primer lugar, utilizando los ángulos presentados en la **Tabla 8** se generó una malla. Para algunas direcciones se utilizaron menos de 25 puntos, dado que se observó que la convergencia de la energía de interacción a 0 ocurría en un rango de distancia menor que para el caso del potencial EAM. En total se obtuvieron 343 puntos, con estos se procedió a obtener la superficie interpolada de radio de mínima energía y posteriormente, se obtuvieron más puntos mediante muestreo informado, de acuerdo a la metodología presentada en la **sección 2.1.2**. En las siguientes figuras se presentan los resultados para la malla.

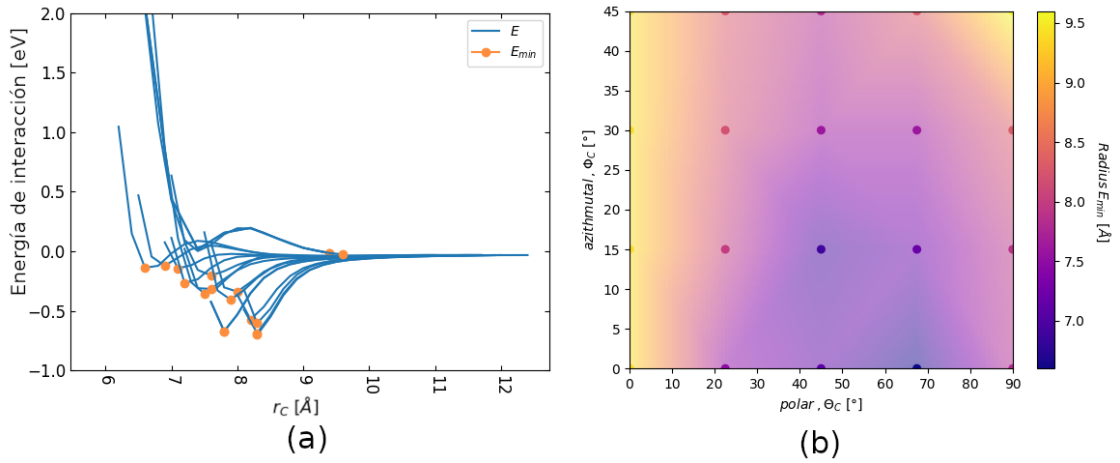


Figura 18: Malla obtenida mediante cálculos de DFT. En (a) cada linea es presentada en un gráfico de energía de interacción vs. radio (r_C), en (b) los puntos correspondientes a los mínimos de (a) se usan para obtener $r_C^{min}(\theta_C, \phi_C)$ mediante una interpolación.

Si comparamos los valores de energía de interacción obtenidos con el método de EAM y DFT, se observa que para DFT en algunas direcciones (θ_m=0°, ϕ_m=0° y θ_m=90° , ϕ_m=45°) el mínimo de la curva es el valor asintótico de 0 a largo rango (en estos casos r_C^* es tomado como el punto más lejano en esta dirección), este comportamiento no fue observado para el potencial EAM. Además, en general las curvas presentan un comportamiento más complejo, donde dependiendo de la dirección puede observarse un pico positivo entre E_{min} and $E = 0$. Por último, para DFT la energía de interacción está en el rango de -0.75 eV o mayor, mientras que para el modelo EAM algunos puntos cuentan con energías de interacción cercanas a -3 eV, en consecuencia, hay una diferencia de orden de magnitud para la energía de ligadura en ambos casos.

A continuación, se obtuvo un conjunto de entrenamiento, $\mathcal{L}_{DFT}$, consistente en los 343

puntos de la malla, 600 puntos generados por muestreo informado y 33 puntos que procedían de ensayos realizados con anterioridad y que se decidió agregar a la base de datos. Los puntos obtenidos se presentan en la siguiente figura,

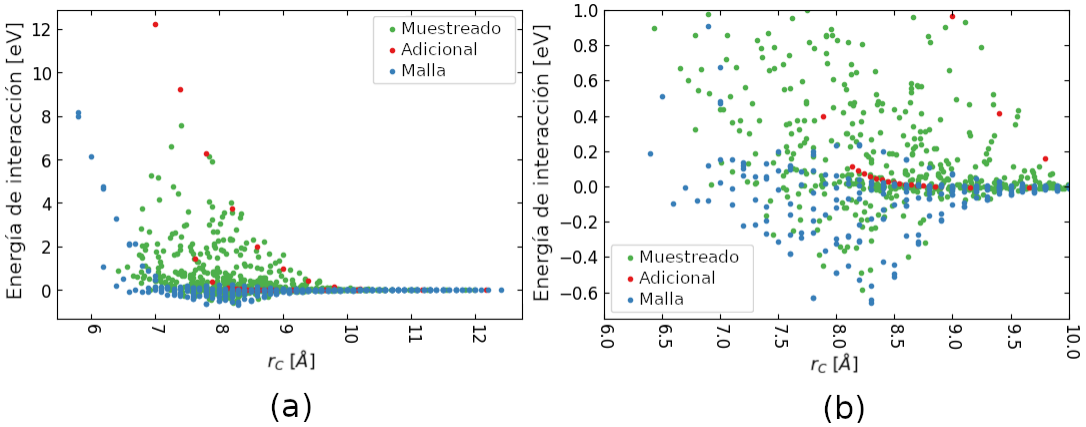


Figura 19: Energía de interacción vs. r_C para los puntos en $\mathcal{L}_{\mathcal{DFT}}$. (a) todos los puntos, (b) detalle de las configuraciones más relevantes.

Se puede observar que la mayoría de puntos de baja energía presentes en $\mathcal{L}_{DFT}$ se corresponden con los que proceden de la malla, esto valida que es más probable que ocurra la adsorción cuando el C está orientado hacia la superficie de la nanopartícula. En consecuencia, la forma en que la malla fue definida es apropiada, dado que de este modo se garantiza que $\mathcal{L}_{DFT}$ contenga configuraciones de alta relevancia.

Una vez que se obtuvo $\mathcal{L}_{DFT}$, se generó un test de prueba $\mathcal{T}_{DFT}$. Para esto se decidió utilizar 3 lineas con 30 puntos en cada una, se escogieron ángulos que no estuvieran presentes en la malla, para asegurarse de que no hay intersección entre los puntos en $\mathcal{T}_{DFT}$ y $\mathcal{L}_{DFT}$. En la siguiente tabla se detallan los ángulos utilizados para cada línea.

Tabla 12: Definición de las líneas presentes en $\mathcal{T}_2$

Nombre	θ_C [°]	ϕ_C [°]	θ_m [°]	ϕ_m [°]	# de puntos	r_c^{min} [Å]
20 °	90	20	180	20	30	8.0
45-20 °	45	20	135	20	30	6.2
20-20 °	20	20	110	20	30	7.9

Luego, empleando $n_{max} = 6$ y $l_{max} = 4$, se probaron distintos valores de σ y r_{cut}. Se decidió utilizar la combinación de parámetros que minimizara el MAE para $\mathcal{T}_{DFT}$ para el modelo que mejor se desempeñó, el Ridge Bi-Lineal en este caso. En la siguiente tabla se presentan los parámetros resultantes.

Tabla 13: Parámetros empleados en los descriptores SOAP para energías calculadas mediante VASP.

n_{max}	l_{max}	σ	r_{cut}
6	4	7.0	0.4

En el siguiente gráfico, las predicciones obtenidas mediante los tres modelos que presentaron un menor MAE para todos los puntos en $\mathcal{T}_{DFT}$ son presentadas, Lasso-Lars lineal, Lasso-Lars Bi-lineal y Ridge Bi-lineal. Luego, en la **Tabla 14** se muestra el MAE de cada modelo.

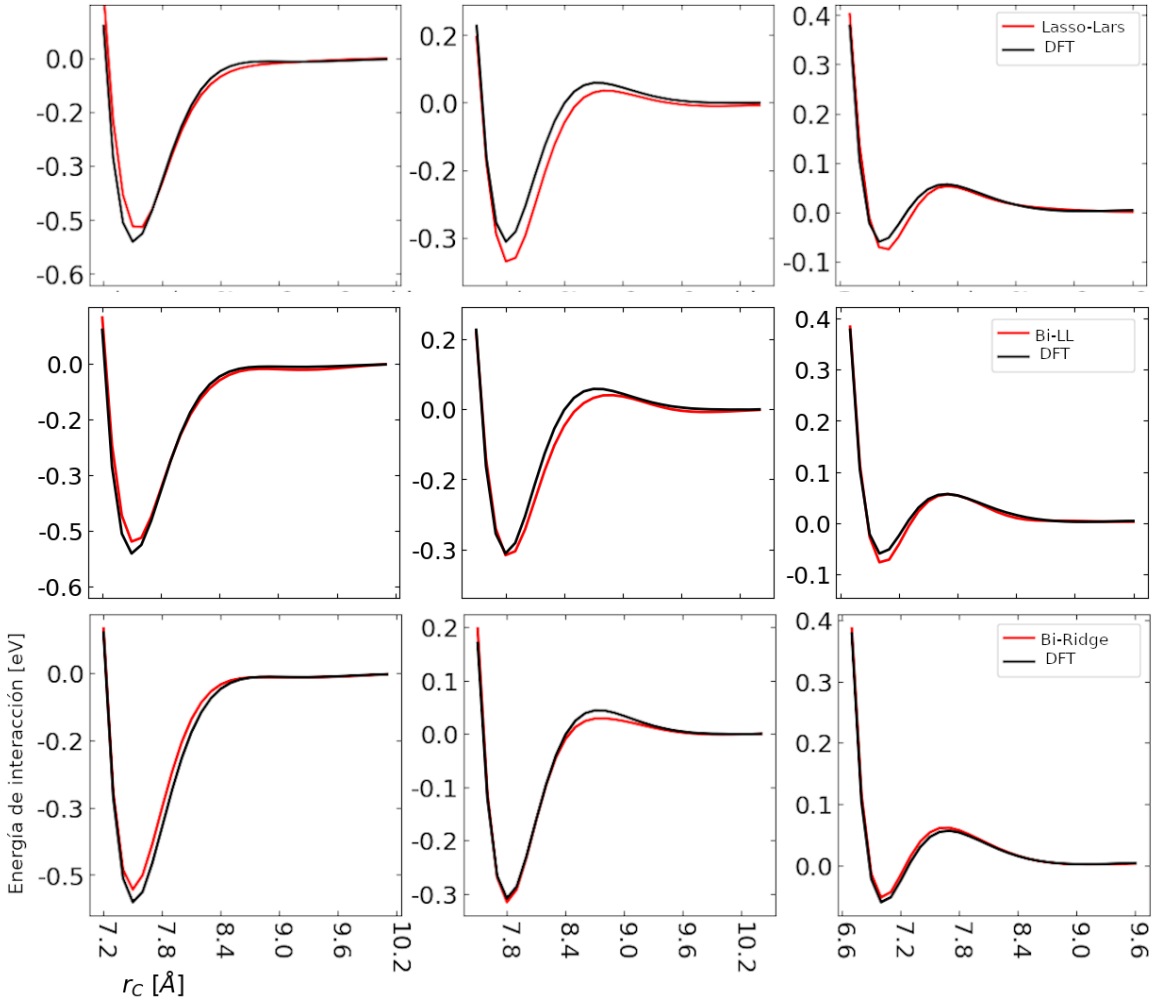


Figura 20: Predicciones obtenidas a partir de $\mathcal{L}_{DFT}$. De arriba hacia abajo, Lasso-Lars lineal, Lasso Lars Bi-lineal y Ridge Bi-lineal. De izquierda a derecha las lineas presentadas e la **tabla 12**.

Tabla 14: Comparación de MAE para las predicciones de energías calculadas mediante DFT

Modelo	MAE: $\mathcal{T}_{\mathcal{DFT}}$ $[eV]$ (solo negativo)	MAE: $\mathcal{T}_{\mathcal{DFT}}$ $[eV]$	MAE: $\mathcal{L}_{\mathcal{DFT}}$ $[eV]$
Lasso-Lars	0,0205	0,0158	0,0191
Lasso-Lars Bi-lineal	0,0097	0,0130	0,0831
Ridge Bi-lineal	0,0071	0,0093	0,0001

Se observa que la regresión Ridge Bi-lineal es la que más certeramente describe las interacciones, el MAE que este modelo presenta para los puntos en $\mathcal{L}_{DFT}$ es especialmente bajo, esto podría indicar que todavía hay lugar para mejorar la precisión de las predicciones mediante el agregado de más puntos a $\mathcal{L}_{DFT}$. De todos modos, para el número de puntos que fueron muestreados se juzga que se logra un buen compromiso entre precisión y esfuerzo computacional requerido para obtener la base de datos.

Para evaluar el efecto de reducir el número de puntos presentes en $\mathcal{L}_{DFT}$ se decidió retirar progresivamente puntos obtenidos por muestreo informado, y a medida que se extraían puntos se calculaba el MAE en $\mathcal{T}_{DFT}$. Los resultados pueden observarse en los siguiente gráficos,

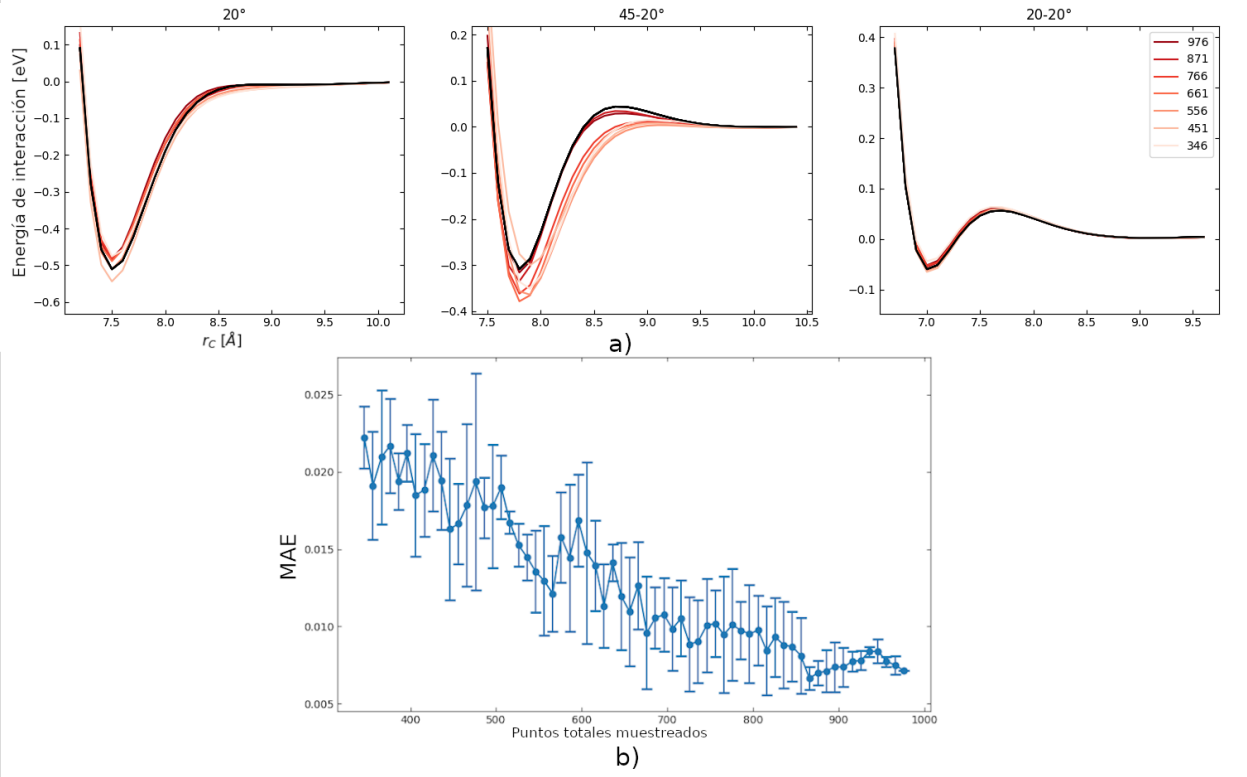


Figura 21: Para el modelo Ridge Bi-lineal, a) predicciones y b) MAE, al extraer puntos en $\mathcal{L}_{DFT}$ obtenidos por muestreo informado.

Se puede ver que el MAE está negativamente correlacionado con el número de puntos muestreados. En la **Figura 21.b** se observa que a medida que se incrementa el número de puntos en el conjunto de entrenamiento hay una disminución tanto del valor medio como desviación del MAE. Cuando hay alrededor de 900 puntos en $\mathcal{L}_{DFT}$ se alcanza cierto grado de convergencia para el MAE, con un valor de aproximadamente 0,0070 eV, consistente con el valor presentado en la **Tabla 14**. De las curvas presentadas en la **Figura 21.a** se puede concluir que al eliminar puntos de $\mathcal{L}_{DFT}$, si bien la precisión disminuye, especialmente para la dirección 45-20 °, la forma de la curva de referencia se sigue reproduciendo de forma cualitativa.

6. Conclusiones y perspectiva

Durante este proyecto las capacidades y limitaciones del empleo de modelos lineales de ML para el estudio de las interacciones entre nanopartículas metálicas y moléculas gaseosas fueron exploradas en profundidad. En la primera parte del trabajo, utilizando métodos clásicos para el cálculo de la energía de interacción, fue posible desarrollar y testear una metodología para la generación de conjuntos de entrenamiento que demostró ser superior al muestreo aleatorio. Esta metodología, que fue denominada el método de la malla más muestreo informado, permitió obtener predicciones más precisas y consistentes que la alternativa. Además, esta metodología no estándar puede ser utilizada para generar conjuntos de entrenamiento sin importar el tamaño o forma de las nanopartículas utilizadas. Por este motivo, se espera que esta herramienta pueda poseer aplicación más allá de los límites de este trabajo.

Durante la segunda parte del proyecto, los métodos que generaron mejores resultados para el potencial interatómico EAM fueron aplicados para desarrollar un modelo de ML basado en cálculos mediante DFT. Se demostró que obtener dicho modelo es posible, con precisión y coste computacional razonables, en particular se encontró que el modelo Ridge Bi-lineal es el más adecuado. Debe ser considerado, que al emplear modelos Bi-lineales los cálculos de energía de interacción requieren de un esfuerzo computacional mayor que el caso de modelos lineales, de todos modos, el tiempo de cómputo es sustancialmente menor que el requerido por VASP. Para el caso del modelo de ML el tiempo de cálculo está en el orden de los microsegundos, mientras que para VASP un cálculo de minimización electrónica toma horas, en consecuencia, el modelo de ML puede ser aplicado fácilmente al estudio de las características catalíticas de la nanopartículas metálicas.

Con esto, se sentaron los cimientos para desarrollar una metodología innovadora para el estudio de las propiedades catalíticas de las nanopartículas. Para continuar con el desarrollo de este método pruebas adicionales deberán ser conducidas, la transferabilidad actual de modelo deberá ser evaluada mediante la aplicación a otras nanopartículas y/o moléculas para poder definir si es posible predecir la energía de interacción para nanopartículas en las cuales el modelo no haya sido entrenado. Nuestra intención es que una vez que la transferabilidad haya sido validada adecuadamente, el modelo pueda convertirse en una herramienta capaz de avanzar el conocimiento en el campo de la catálisis.

7. Apéndice

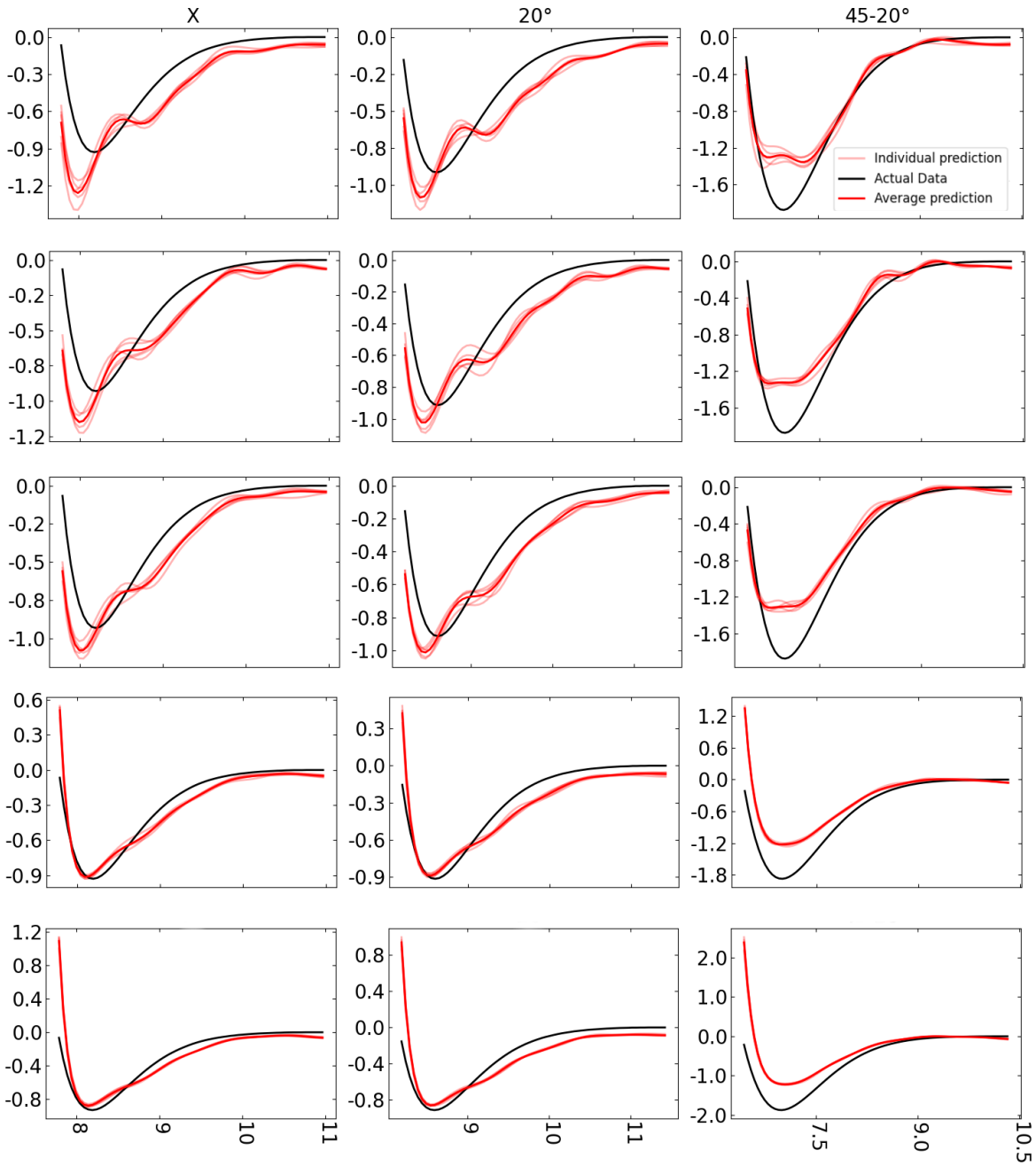


Figura A: Predicciones para el modelo Bi-LL Ridge entrenado con 5 conjuntos independientes que contienen la malla más 5 (de arriba hacia abajo) 250, 500, 750, 5.000 y 10.000 configuraciones obtenidas por muestreo informado. En el eje x se muestra el radio (r_C) en $\mathring{A}$, para el eje y la energía de interacción en eV.

Tabla A: Parámetros utilizados para el cálculo de los descriptores ACSF

Funciones de dos cuerpos (G^2)					
η	R_S	η	R_S	η	R_S
1,0	0	2,0	4,0	2,5	1,0
2,0	0	3,0	1,0	2,5	2,0
3,0	0	3,0	2,0	2,0	2,5
4,0	0	3,0	3,0	1,0	3,5
0,5	0	3,0	4,0	2,5	2,5
1,0	1,0	4,0	3,0	1,5	3,5
1,0	0,5	4,0	4,0	1,5	3,2
1,0	2,0	4,5	3,0	4,0	1,0
1,0	4,0	4,5	4,0	2,0	4,5
2,0	1,0	1,5	3,0		
2,0	2,0	1,5	4,0		
2,0	3,0	2,5	3,0		

Tabla B: Parámetros utilizados para el cálculo de los descriptores ACSF

Funciones de tres cuerpos (G^4)		
η	ζ	λ
1,0	1,0	1,0
1,0	2,0	2,0
1,0	2,0	1,0
2,0	2,0	2,0
1,0	4,0	4,0
3,0	2,0	2,0
4,0	1,0	1,0
3,0	4,0	2,0

Tabla C: Valores de parámetros ensayados para el cálculo de los descriptores SOAP. (todas las combinaciones fueron consideradas)

σ	r_{cut}
0.1	2.0
0.2	3.0
0.3	3.5
0.4	4.0
0.5	4.5
0.6	5.0
0.7	5.5
	6.0
	7.0

Tabla D: Parámetros utilizados en los cálculos mediante VASP

Parámetro	Valor	Comentario [15]
PREC	NORMAL	
ALGO	FAST	
EDIFF	1e-6	
ENCUT	415	
ISPIN	2	cálculos con spin polarizado
NUPDOWN	0	
ISMEAR	0	Gaussian smearing
SIGMA	0.2	ancho del smearing
Box size	30	caja cúbica

Referencias

[1] Zhen Ma and Francisco Zaera. Heterogeneous catalysis by metals. pages 1–16, 2014.

[2] Didier Astruc. Introduction: Nanoparticles in catalysis. *Chemical Reviews*, 120(2):461–463, 2020. PMID: 31964144.

[3] A Logadottir, T.H Rod, J.K Nørskov, B Hammer, S Dahl, and C.J.H Jacobsen. The brønsted–evans–polanyi relation and the volcano plot for ammonia synthesis over transition metal catalysts. *Journal of Catalysis*, 197(2):229–231, January 2001.

[4] Kai S. Exner. Paradigm change in hydrogen electrocatalysis: The volcano's apex is located at weak bonding of the reaction intermediate. *International Journal of Hydrogen Energy*, 45(51):27221–27229, October 2020.

[5] K Rossi. Multiscale modelling of metallic nanoparticles structural and catalytic properties. 1 Jun 2019.

[6] Georgios D Barmparis, Zbigniew Lodziana, Nuria Lopez, and Ioannis N Remediakis. Nanoparticle shapes by using wulff constructions and first-principles calculations. *Beilstein Journal of Nanotechnology*, 6:361–368, February 2015.

[7] Abhirup Patra, Jefferson E Bates, Jianwei Sun, and John P Perdew. Properties of real metallic surfaces: Effects of density functional semilocality and van der waals nonlocality. *Proc. Natl. Acad. Sci. U. S. A.*, 114(44):E9188–E9196, October 2017.

[8] L. Delgado Callico. Computational study of the thermal and chemical stability of metallic nanoparticles, Nov 2022.

[9] L D Marks. Experimental studies of small particle structures. *Reports on Progress in Physics*, 57(6):603–649, June 1994.

[10] Murray S. Daw, Stephen M. Foiles, and Michael I. Baskes. The embedded-atom method: a review of theory and applications. *Materials Science Reports*, 9(7–8):251–310, March 1993.

[11] S. M. Foiles, M. I. Baskes, and M. S. Daw. Embedded-atom-method functions for the fcc metals cu, ag, au, ni, pd, pt, and their alloys. *Physical Review B*, 33(12):7983–7991, June 1986.

[12] Lammps documentation. `https://docs.lammps.org/Manual.html`. Accessed: 07.07.2024.

[13] Markus Bursch, Jan-Michael Mewes, Andreas Hansen, and Stefan Grimme. Best-practice dft protocols for basic molecular computational chemistry**. *Angewandte Chemie*, 134(42), September 2022.

[14] NM Harrison. An introduction to density functional theory. *Nato Science Series Sub Series III Computer and Systems Sciences*, 187:45–70, 2003.

[15] Vasp documentation. `https://www.vasp.at/`. Accessed: 04.06.2024.

[16] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7), February 2011.

[17] Lauri Himanen, Marc O J Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.*, 247(106949):106949, February 2020.

[18] Alan Julian Izenman. *Modern multivariate statistical techniques*. Springer texts in statistics. Springer, New York, NY, December 2008.

[19] Uniform distribution on a sphere. `https://www.bogotobogo.com/Algorithms/uniform_distribution_sphere.php`. Accessed: 24.07.2023.

[20] Pavlo O Dral, Alec Owens, Sergei N Yurchenko, and Walter Thiel. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.*, 146(24):244108, June 2017.

[21] Jona Nägerl. Machine learning approaches to model interactions between carbon dioxide molecules and catalytic nanoparticles, 2023.

[22] Lammps documentation. `https://www.ovito.org/manual/python/modules/ovito.html`. Accessed: 10.09.2024.