

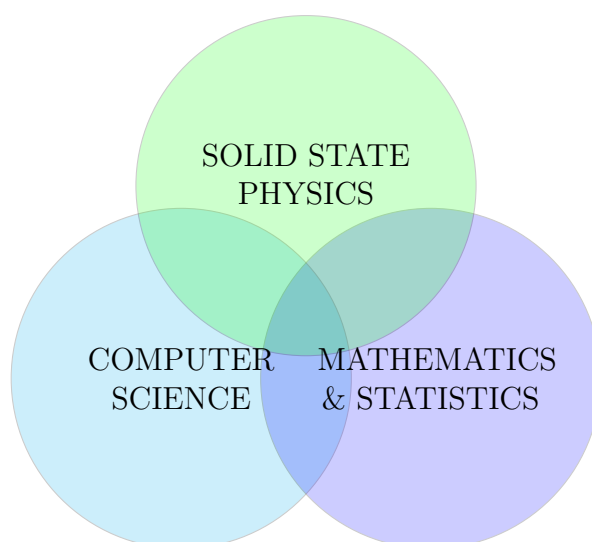


UNIVERSIDAD TÉCNICA
FEDERICO SANTA MARÍA

DEPARTAMENTO
DE FÍSICA

Trabajo de Fin de Grado

Licenciatura en Ciencias, mención Física



“MACHINE LEARNING PARA LA DETERMINACIÓN DE ESTRUCTURAS ESTABLES EN ALEACIONES $\text{Ti}_x\text{Ta}_{1-x}$ ”

PABLO ÁLVAREZ SÁNCHEZ

Profesores: [Eric Suárez](#) & [Jhon González](#)

5 de marzo de 2024

Resumen

El diseño de aleaciones ha sido crucial para los avances tecnológicos en nuestra sociedad. En particular las aleaciones de Titanio y Tántalo, que son las investigadas en este trabajo, han atraído la atención de la industria médica al presentarse como materiales de alta biocompatibilidad. El problema de encontrar la configuración más estable en una aleación, a medida que aumentamos el número de impurezas, se vuelve inviable de resolver mediante el enfoque de aplicar simulaciones de dinámica molecular a todo espacio de configuraciones, ya que este crece exponencialmente con el número de impurezas. Con el enfoque propuesto en este trabajo, el costo computacional se ve abruptamente reducido, manteniendo constante el requerimiento de simuladores de dinámica molecular cuando se aumenta el número de impurezas. Para lograr esto ocupamos técnicas de Machine Learning para investigar las relaciones que hay entre la geometría de una configuración de la aleación de Titanio-Tántalo y su energía, logrando la predicción de las configuraciones más estables en función del número de impurezas al encontrar los parámetros geométricos que minimizan la energía de sistema.

Palabras clave: *Material Science; Machine learning; Titanium Alloys; DFT.*

Índice general

1. Introducción	4
1.1. Importancia de las aleaciones de Titanio y Tántalo	5
1.1.1. Búsqueda de aleaciones de Titanio-Tántalo	6
1.2. Objetivos	8
2. Marco Teórico	10
2.1. Introducción a la la Física del Estado Sólido	10
2.1.1. Cristales	10
2.1.2. Simetrías	12
2.1.3. Índices de Miller	14
2.1.4. Red recíproca	15
2.1.5. Teorema de Bloch	16
2.2. Modelo teórico	18
2.2.1. Teoría del Funcional de Densidad	18
2.3. Softwares	22
2.3.1. Generación de estructura	22
2.3.2. Simulaciones de dinámica molecular <i>ab initio</i>	24
2.4. Aprendizaje Automático	26
2.4.1. Flujo de trabajo de Aprendizaje Automático en el diseño de aleaciones	26
2.4.2. Tipos	31
3. Metodología	33
3.1. Generación de datos	33
3.2. Preprocesamiento de datos	34

3.2.1. El descriptor <code>meandistance</code>	35
3.3. Exploración	38
3.3.1. El descriptor <code>cantmax</code> y SVM	42
3.4. El Modelo lineal: $E0_{n_{\text{impurezas}}} \propto \text{meandistance}$. Hipótesis y evaluación.	47
3.4.1. División de entrenamiento y de prueba	48
3.4.2. Construcción de modelos para números de impurezas desconocidos	49
3.4.3. Modelo para la pendiente y el intercepto	50
3.4.4. Evaluación de modelos	50
4. Resultados	52
4.1. Confirmación del modelo lineal en <code>meandistance</code> para la Energía (E_0) dado un número de impurezas	52
4.2. Un modelo para la Energía (E_0) construido con los datos reales de 1, 2, 3 y 4 impurezas	52
4.2.1. Modelo para la Energía (E_0) en términos de <code>meandistance</code> y <code>n_impurezas</code>	53
4.3. Descriptor <code>cantmax</code> y su importancia en la selección de configuraciones candidatas	53
4.3.1. Criterio de selección de la configuración de mínima energía	54
4.3.2. Sin muestras para un número de impurezas dado	54
4.3.3. Con muestras para un número de impurezas dado	59
4.4. Simetría de las configuraciones y Energía	59
4.4.1. Extremos de Energía	60
5. Conclusiones	62
5.1. Trabajos futuros	63
5.2. Limitaciones de los resultados	64
Anexos	65
A. Aprendizaje Automático	66
A.1. Definiciones	66
A.2. Aprendizaje supervisado	67
A.2.1. Regresión polinomial	68
A.3. SVM	68

B. Teoría del funcional de densidad	71
B.1. DFT	71
B.1.1. Aproximación de Hartree-Fock.	75
B.1.2. Segundo Teorema de Hohenberg-Kohn	76
B.1.3. Hamiltoniano molecular	77
B.1.4. Aproximación de Born-Oppenheimer	79
B.1.5. Teoremas Hohenberg-Kohn	80
B.1.6. Método Kohn-Sham	82
B.1.7. Funcionales de Correlación-Intercambio	85
C. Calculos por medio de VASP	89
C.0.1. input file	89
C.0.2. output file	90
C.0.3. E0 Efree	90
C.0.4. Energías de la iteración	90
D. Simetría de las configuraciones	94
D.1. Simetrías	94
D.1.1. Número de peraciones de simetría	94
D.1.2. Número de simetrías de rotación	96
D.1.3. Desviación estándar del descriptor <code>meandistance</code>	97
D.1.4. Tipo	98
D.1.5. Matriz de Correlación	99
Bibliografía	100

Capítulo 1

Introducción

El desarrollo de nuevos materiales juega un rol fundamental en el desarrollo de nuevas tecnologías, debido a la necesidad de materiales novedosos con propiedades ideales. Las aleaciones en particular han atraído la atención en diversos campos, ofreciendo materiales con ventajas como alta dureza, gran resistencia a la corrosión, alta conductividad eléctrica y biocompatibilidad. Por ejemplo, las aleaciones de aluminio presentan una gran resistencia a la corrosión y soldabilidad, las aleaciones de cobre una gran resistencia a la corrosión y alta conductividad eléctrica, las aleaciones de Titanio una alta resistencia mecánica y biocompatibilidad que es de gran interés en la industria médica.

Junto con el desarrollo de nuevos materiales, a medida que aumenta el tamaño de los datos, la ciencia basada en datos se propone como un cuarto paradigma científico [1] después de las tres eras de descubrimientos científicos impulsados por experimentos convencionales, métodos teóricos y simulación computacional. En la figura [Fig. 1.1](#) observamos estos cuatro paradigmas representados con flechas, destacando hitos importantes para la ciencia de materiales.

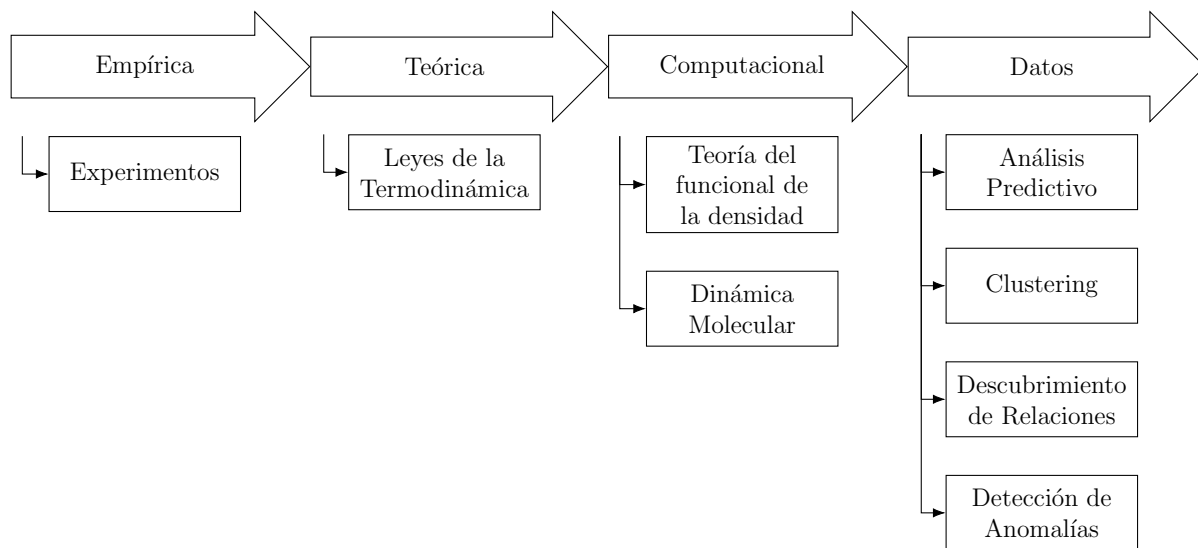


Figura 1.1: Los cuatro paradigmas de la ciencia en el transcurso de la historia con énfasis en los hitos de la ciencia de materiales.

En la última década, un subcampo de la ciencia basada en datos que evoluciona rápidamente es el *Aprendizaje Automático*, el cual es una disciplina cruzada entre el aprendizaje estadístico y la informática.

El Aprendizaje Automático realiza nuevas predicciones al comprender los patrones ocultos en los datos y así puede proporcionar orientación para el diseño de materiales. Actualmente, se han logrado notables avances en la ciencia de materiales usando técnicas de Aprendizaje Automático, como por ejemplo los modelos GNoME, desarrollados por *DeepMind*[4], que *mediante el uso de técnicas de deep learning han descubierto más de 2,2 millones de estructuras estables con respecto a trabajos anteriores* [16]. El Aprendizaje Automático nos permite descubrir nuevos materiales de una manera nueva y diferente al método experimental de “prueba y error”, el cual implica un gran costo debido a las propias limitaciones y dificultades que tiene la síntesis de estos.

1.1. Importancia de las aleaciones de Titanio y Tántalo

El titanio y sus aleaciones son cruciales en campos como la aeronáutica y la industria automotriz debido a su resistencia, ligereza. En la industria médica las aleaciones de titanio son de gran interés, ya que existen aleaciones que presentan una alta compatibilidad con los tejidos humanos. A pesar de sus ventajas, el titanio puro presenta limitaciones como un módulo elástico alto que puede causar problemas de integración en aplicaciones biomédicas. Para superar esto, se dopa al titanio con elementos como el

tántalo, que mejora sus propiedades mecánicas y reduce el módulo elástico [3], acercándolo más al de los huesos humanos y disminuyendo el efecto de blindaje de estrés que puede conducir a la reabsorción ósea.

1.1.1. Búsqueda de aleaciones de Titanio-Tántalo

Las aleaciones puras a base de Ti y Ta exhiben dos fases de equilibrio: a temperaturas más bajas tienen una estructura HPC ($\alpha-Ti$), mientras que a temperaturas más altas poseen una estructura BCC ($\beta-Ti$) [2]. Dado un porcentaje de impurezas, el objetivo es encontrar la configuración que minimice la energía del sistema, ya que esta será la configuración más estable y la que se formará en la aleación.

Para lograr esto se considera el enfoque computacional, donde por medio de ayuda de softwares buscamos la configuración más estable y para luego se estudian sus propiedades mediante el uso de simulaciones de dinámica molecular y así determinar si es una aleación viable o no.

Para la búsqueda de la configuración estable de la aleación de Titanio y Tántalo se pueden considerar distintos enfoques. A continuación se describe el enfoque más intuitivo, que consiste en un método exhaustivo.

Método exhaustivo: Complejidad y no viabilidad

Para encontrar la configuración de menor energía para un número de impurezas dado, el método exhaustivo consiste en calcular la energía de todas las configuraciones posibles, para luego escoger la de menor energía. Este método no resulta viable cuando empezamos el número de impurezas crece, ya que al generar computacionalmente las configuraciones posibles nos enfrentamos a un problema combinatorio: estamos buscando todas las configuraciones posibles. Luego de generar las configuraciones, calculamos la energía por medio de la simulación para cada una de ellas y así encontrar la más estable mediante una búsqueda entre todos esos pares **configuración-energía**.

En la figura Fig. 1.2 se muestra la cantidad de configuraciones posibles para nuestro problema, donde tenemos 32 posiciones y n sustituciones:

$$n_combinaciones = \binom{32}{n} = \frac{32!}{n!(32-n)!}$$

Es importante mencionar que estas cantidades de átomos muestran un caso en particular, que es el que se estudiará en este trabajo, pero el problema combinatorio se presenta para cualquier número y

tipo de átomos cuando usamos este enfoque.

Este primer enfoque mencionado no considera simetrías ni condiciones de borde periódicas, por lo tanto hay redundancia en las configuraciones. En la figura Fig. 1.2 se muestra la cantidad de configuraciones posibles usando SOD [23] el cuál es un software que considera simetrías y condiciones de borde periódicas, evitando la redundancia en las configuraciones, lo cual tiene como consecuencia una cantidad de configuraciones considerablemente menor. A pesar de que la cantidad de configuraciones sea menos, aún se observa el comportamiento exponencial.

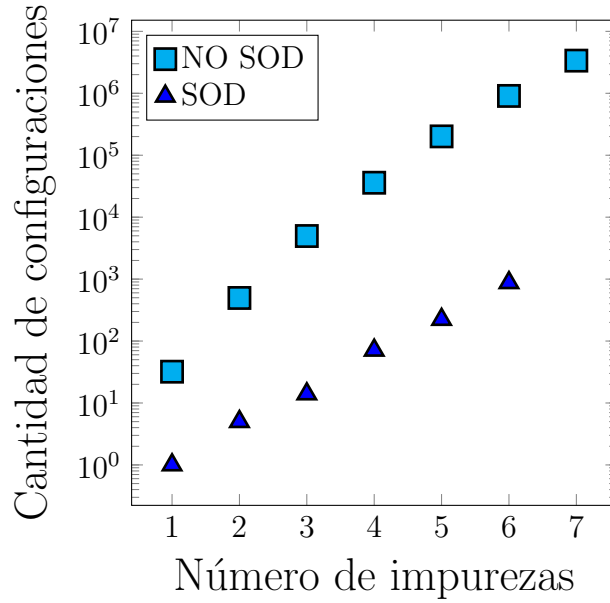


Figura 1.2: Número de configuraciones no equivalentes

El enfoque actual consiste en la generación de todas las configuraciones no equivalentes, donde se consideran equivalentes si hay una transformación isométrica que convierte una en la otra. Para generar esas configuraciones usamos el software *SOD* y luego calcular la energía de cada una de ellas haciendo uso del software de simulación de dinámica molecular *VASP* (Viena Ab-initio Simulation Package). En la figura Fig. 1.3 se muestra un diagrama que resume el flujo de trabajo actual y que este trabajo propone mejorar.

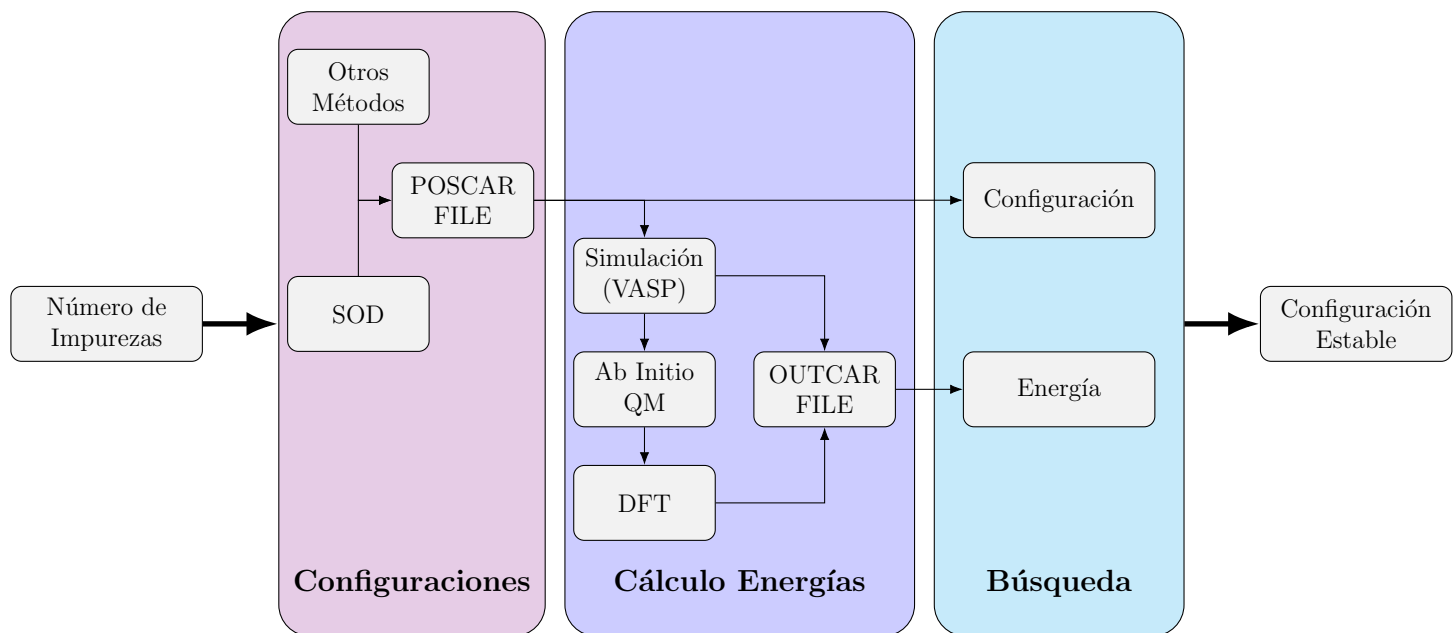


Figura 1.3: Flujo de trabajo actual.

Este enfoque tiene un alto costo computacional, el cual se debe principalmente a que el costo de la simulación para obtener la energía de una sola configuración es alto y el número de configuraciones crece de manera exponencial al aumentar el número de impurezas.

1.2. Objetivos

El objetivo de este trabajo es generar un algoritmo basado en Machine Learning que permita predecir la configuración o configuraciones más estables a partir de un número de sustituciones en un cristal. En específico se estudió un cristal de Titanio con sustituciones de Tántalo.

Como fue discutido en [Subsec. 1.1.1](#) cuando aumentamos el número de impurezas el número de configuraciones posibles crece de manera exponencial incluso usando SOD, en cual también tiene un costo alto costo computacional, por lo tanto, el método exhaustivo debe ser evitado debido a la no escalabilidad que tiene. Así es como surge el interés de encontrar descriptores que nos permitan al menos clasificar las configuraciones para así reducir el número de configuraciones a considerar en las simulaciones.

En [Fig. 1.4](#) se muestra un diagrama que resume nuestro objetivo: Encontrar un método que nos permita predecir la configuración o configuraciones más estables, reduciendo el universo de configuraciones a calcular por métodos de primeros principios para un número de impurezas (o porcentaje de impureza).

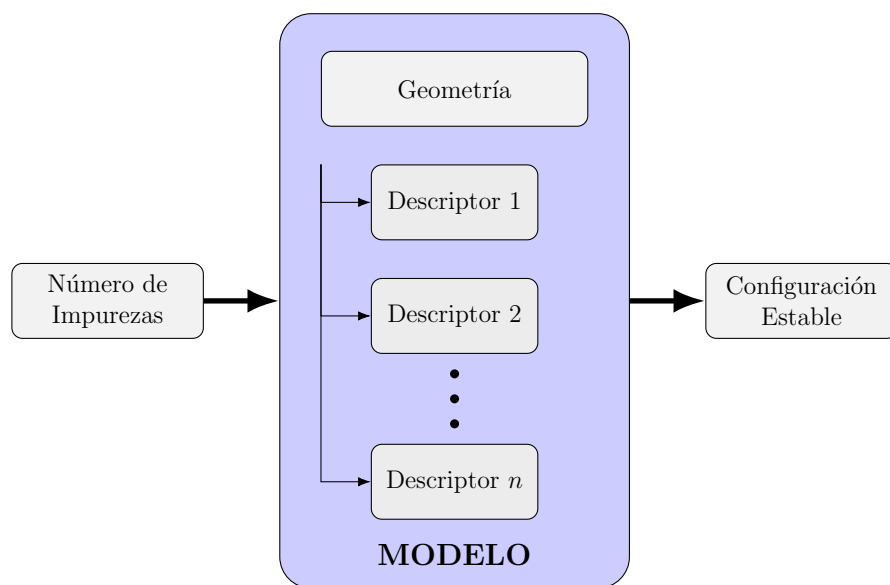


Figura 1.4: Nuestro objetivo: Encontrar un método que nos permita predecir la configuración o configuraciones más estables para un determinado número de impurezas.

Capítulo 2

Marco Teórico

2.1. Introducción a la la Física del Estado Sólido

Esta rama de la física estudia las propiedades físicas de los materiales sólidos, utilizando áreas de estudio de la física tales como la mecánica cuántica, la cristalografía, el electromagnetismo y la metalurgia física. Esta rama forma la base teórica de la ciencia de materiales y su desarrollo ha sido fundamental en el campo de las aplicaciones tecnológicas de microelectrónica, ya que por ejemplo a hecho posible el desarrollo de materiales semiconductores y transistores.

La física del estado sólido, como teoría general, se centra en los cristales.

2.1.1. Cristales

Un cristal puede describirse como un sólido cuya estructura se compone de átomos dispuestos en un patrón periódico y tridimensional. Los cristales están constituidos por una red periódica (lattice) y una base (las posiciones de los átomos dentro de la red). Existe una cantidad finita de redes de Bravais. A menudo, se abstrae la presencia de los átomos y se considera, en cambio, un conjunto de puntos imaginarios. Estos puntos mantienen una relación espacial constante con los átomos del cristal y sirven como una referencia para entender cómo se estructura el cristal.

Para formar este conjunto de puntos se divide el espacio con tres grupos de planos. Cada grupo contiene planos que son paralelos y equidistantes entre sí. Esta división resulta en una serie de celdas que son idénticas en forma, tamaño y orientación con respecto a sus celdas adyacentes. Cada celda tiene la forma de un paralelepípedo. Los planos que segmentan el espacio se cortan entre sí formando líneas,

y estas líneas se intersectan en los puntos mencionados antes. Se puede ver una representación de esto en la Fig. 2.1 y Fig. 2.3.

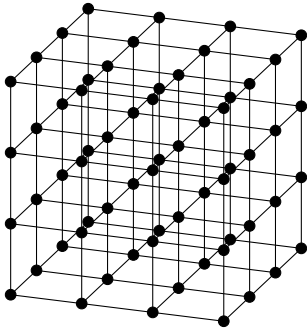


Figura 2.1: Red Cristalina

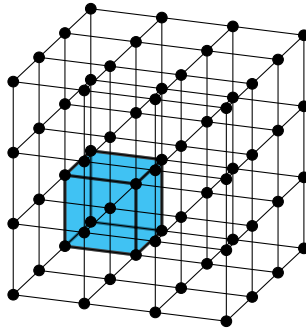


Figura 2.2: Red Cristalina

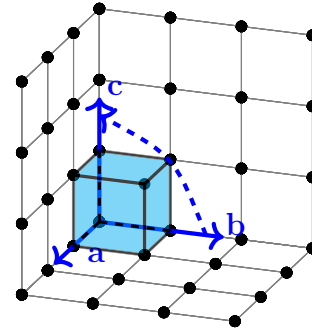


Figura 2.3: Celda Unitaria

Estos puntos tienen una característica distintiva: cada uno está rodeado por un entorno uniforme, lo cual significa que la configuración de puntos que se percibe desde cualquier punto dado, en una dirección particular, es idéntica a la vista desde cualquier otro punto de la red cristalina en esa misma dirección.

Considerando que todas las celdas en la red de puntos de la Fig. 2.1 son idénticas en forma y tamaño, cualquier celda puede seleccionarse como celda unidad. La geometría y dimensiones de esta celda unidad pueden describirse mediante tres vectores, $\vec{a}$, $\vec{b}$, y $\vec{c}$ que se originan en una esquina de la celda elegida como punto de partida (fig. 2.3).

Estos vectores son fundamentales, ya que definen la celda. Estos vectores son conocidos como los ejes cristalográficos de la celda y se caracterizan por sus magnitudes (a , b , c) y ángulos interiores (α , β , γ). A estas magnitudes y ángulos se les denomina parámetros de la red de la celda unidad.

La estructura completa de puntos emerge de la repetición sistemática de los vectores $\vec{a}$, $\vec{b}$, y $\vec{c}$, partiendo del punto de origen de la red. Así, la posición de cualquier punto en la red puede expresarse como $P\vec{a} + Q\vec{b} + R\vec{c}$, donde P , Q y R son números enteros.

La segmentación del espacio en tres grupos de planos da lugar a celdas unidad de distintas formas y tamaños, y consecuentemente, a varios tipos de redes cristalinas. Estas celdas varían según los valores específicos asignados a las magnitudes a , b y c y a los ángulos α , β y γ . Los diversos tipos de redes cristalinas que resultan de este proceso se clasifican en siete sistemas cristalinos, los cuales están enumerados en el Tabla 2.1.

Sistema	Longitudes axiales y ángulos	Celda de Bravais	Símbolo de celda
Cúbica	Tres ejes iguales en ángulos rectos $a = b = c$, $\alpha = \beta = \gamma = 90^\circ$	SC	P
		BCC	I
		FCC	F
Tetragonal	Tres ejes en ángulos rectos, dos ejes iguales $a = b \neq c$, $\alpha = \beta = \gamma = 90^\circ$	SC	P
		BCC	I
Ortorrómbica	Tres ejes desiguales en ángulos rectos $a \neq b \neq c$, $\alpha = \beta = \gamma = 90^\circ$	SC	P
		BCC	I
		Base C	C
		FCC	F
Trigonal ó Rom- bodrónica	Tres ejes iguales en ángulos iguales $a = b = c$, $\alpha = \beta = \gamma \neq 90^\circ$	SC	P
Hexagonal	Dos ejes coplanares iguales a 120° , tercer eje a ángulos rectos $a = b \neq c$, $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$	SC	P
Monoclínica	Tres ejes desiguales, un par no a ángulos rectos $a \neq b \neq c$, $\alpha = \gamma = 90^\circ \neq \beta$	SC	P
		Base C	C
Triclínica	Tres ejes desiguales, ángulos desiguales y ninguno recto $a \neq b \neq c$, $\alpha \neq \beta \neq \gamma \neq 90^\circ$	SC	P

Cuadro 2.1: Sistemas cristalinos y sus características.

2.1.2. Simetrías

La propiedad que define a un cristal es su simetría. A continuación se describen las operaciones de simetría más comunes en cristales.

Los cristales se clasifican principalmente por su simetría espacial, definida matemáticamente como el grupo de transformaciones que mantienen invariante su estructura geométrica. Estas incluyen tras-

laciones, rotaciones e inversiones, conocidas como transformaciones ortogonales: conservan ángulos y longitudes.

El orden de una operación es el número de veces que se debe aplicar para obtener la identidad, es decir, para que el objeto retorne a su estado original. Las transformaciones ortogonales forman un grupo denominado grupo ortogonal, cuyas herramientas teóricas simplifican el estudio de simetrías.

Operaciones de Simetría

Un objeto puede tener un conjunto de operaciones de simetría que lo dejan inalterado, formando su grupo de simetría: un subgrupo del grupo ortogonal. Por ejemplo, rotar un cubo 90 grados alrededor de una de sus caras lo deja aparentemente inalterado, lo que convierte a dicha rotación en una operación de simetría.

Simetría Periódica

Los cristales se definen por su simetría periódica, por lo tanto existen ciertos vectores de traslación no alteran su estructura. Se pueden escoger tres vectores independientes llamados vectores de red, que forman una subestructura paralelepípeda conocida como celda unidad. La celda unidad puede replicarse para recrear el cristal completo.

La elección de la celda unidad no es única y puede variar sin afectar la simetría subyacente. Típicamente, el origen se elige en la ubicación de un átomo o elemento de simetría.

Simetría de Grupo Puntual

Los cristales también pueden poseer simetría de grupo puntual. Las operaciones relevantes se representan mediante matrices ortogonales 3×3 . Si una matriz M tiene un determinante de 1, la operación es rotacional; si es -1, implica inversión, denominándose roto-inversión. Los grupos que contienen solo rotaciones, reflexiones e inversiones se llaman grupos puntuales.

Grupos de Simetría en Tres Dimensiones

Un grupo de simetría puede dividirse en su componente traslacional (subgrupo-t) y su componente roto-inversivo (grupo puntual o subgrupo-k). Solo 32 grupos puntuales son compatibles con una red en 3D, denominados grupos puntuales cristalográficos. Según el teorema de restricción cristalográfica, los

cristales 3D solo pueden tener simetría rotacional de 1, 2, 3, 4 o 6 veces [17]. Un grupo de simetría que incluye simetría de red 3D se llama grupo espacial y existen 230 grupos.

2.1.3. Índices de Miller

Cualquier plano en la red puede ser indexado por tres puntos no colineales. Estos tres puntos son las intersecciones del plano con los ejes de coordenadas. Para determinar fácilmente la orientación del cristal, usamos los índices de Miller hkl . Para obtener estos índices debemos tomar los recíprocos de las intersecciones y obtener un conjunto de los enteros más pequeños que proporcionen la misma proporción. Estos pequeños enteros se interpretan como los índices de Miller y son expresados generalmente como (hkl) . Por ejemplo, si las intersecciones fuesen $\{3, 2, \frac{2}{3}\}$, los valores recíprocos serían $(\frac{1}{3}, \frac{1}{2}, \frac{3}{2})$ y los números en esta proporción son $\frac{1}{6}(2 \ 3 \ 3)$. Por lo tanto, los índices de Miller son (233) .

Al tener los índices de Miller, podemos identificar el tipo de celda

Centrada en el cuerpo

$$h + k + l = 2n$$

Centrada en las caras

h, k, l son todos par o impar

Single face centering

$$k + l = 2n, h + l = 2n, h + k = 2n$$

Rombohédrica centrada

$$-h + k + l = 3n \text{ (obverse) y } h - k + l = 3n \text{ (reverse)}$$

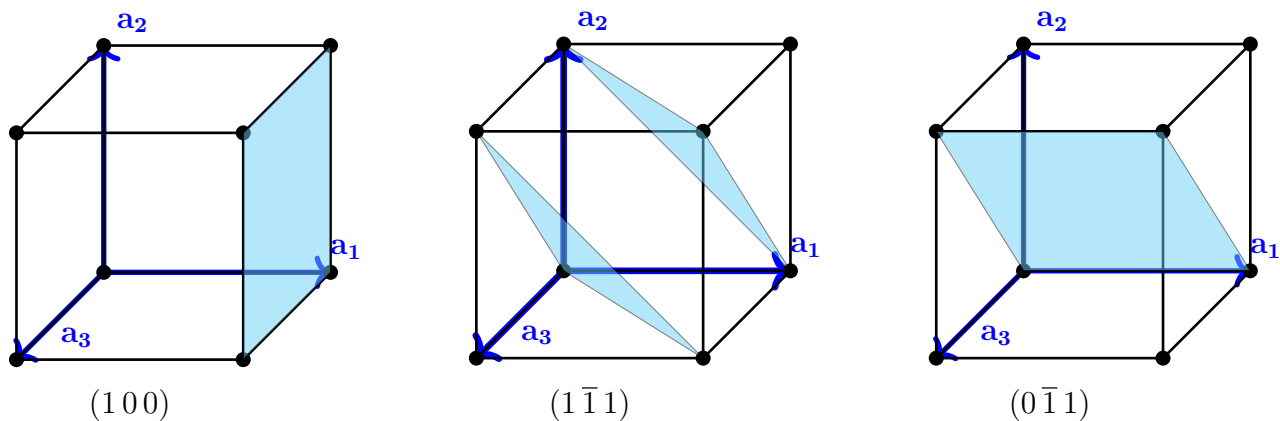


Figura 2.4: Ejemplos de índices de Miller

2.1.4. Red recíproca

Consideremos la siguiente señal (S) y su transformada de Fourier ($\mathcal{F}(S)$) (fig. 2.5)

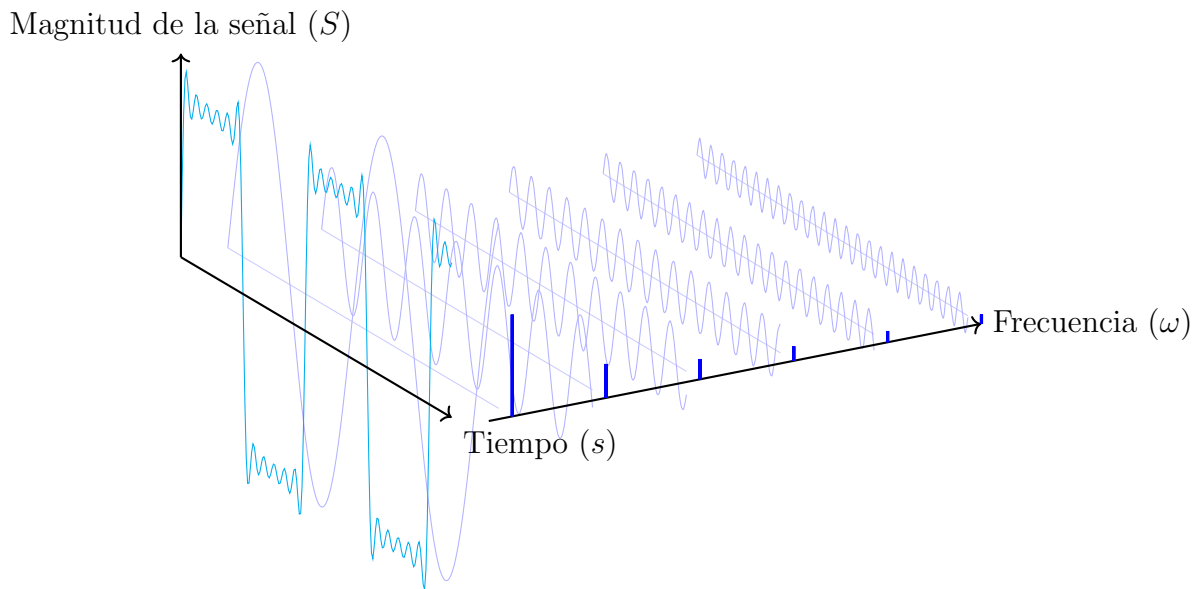


Figura 2.5: Una señal (S) y su transformada ($\mathcal{F}$) de Fourier

La transformada de Fourier en el tiempo nos dice cuánto de cada frecuencia (ω) está contenida en la señal. La frecuencia, tomada literalmente, nos indica qué tan frecuente ocurre un evento, en este caso un mínimo o máximo. Alternativamente, esta señal podría ser una señal en el espacio y la transformada de Fourier nos dice cuánto de cada frecuencia espacial (k) está contenida en nuestra señal, es decir, qué tan separados están los mínimos y máximos locales ($\lambda = \frac{2\pi}{k}$). Si estuviéramos en un máximo, podríamos esperar encontrar otro máximo si mirásemos una longitud de onda a la izquierda o derecha.

Esta forma de medir la periodicidad utilizando transformadas de Fourier es muy útil para entender las estructuras cristalinas. De la misma manera que pasamos de una distancia de repetición en una dimensión a tres vectores de red en tres dimensiones con la celda unitaria, pasamos de una línea de Fourier unidimensional a un espacio de Fourier tridimensional. En nuestro contexto, este espacio se llama espacio recíproco y contiene todas las posibles distancias y direcciones de repetición. De la misma manera que las señales periódicas producen transformadas de Fourier discretas en una dimensión, estructuras periódicas como los cristales producen puntos discretos en el espacio recíproco. Estos puntos forman una red llamada red recíproca y es similar a la red de Bravais en el espacio real en que hay tres vectores de red recíprocos, y los puntos en la red recíproca están ubicados en todas las posibles combinaciones lineales de los vectores de red recíprocos. Un aspecto elegante de los vectores de red recíprocos es que

pueden calcularse a partir de los vectores de red usando las ecuaciones:

$$\begin{aligned}\vec{a}^* &= \frac{2\pi}{V}(\vec{b} \times \vec{c}) \\ \vec{b}^* &= \frac{2\pi}{V}(\vec{c} \times \vec{a}) \\ \vec{c}^* &= \frac{2\pi}{V}(\vec{a} \times \vec{b})\end{aligned}\tag{2.1}$$

Ahora que tenemos una red recíproca podemos definir una celda unitaria recíproca. Hay varias formas de definir la celda unitaria recíproca, pero sólo abordaremos una.

Comenzando con un punto en la red recíproca, podemos trazar vectores a todos los demás puntos de la red recíproca. Luego colocamos planos normales a estos vectores, a medio camino entre los dos puntos. El poliedro sólido delimitado por estos planos es la celda unitaria recíproca. Esta celda unitaria es la primera zona de Brillouin o comúnmente abreviada como la zona de Brillouin. El método para construir la zona de Brillouin explicado anteriormente se ilustra en la Fig. 2.6 para el caso de dos dimensiones. La zona de Brillouin es la celda unitaria en el espacio de Fourier (recíproco).

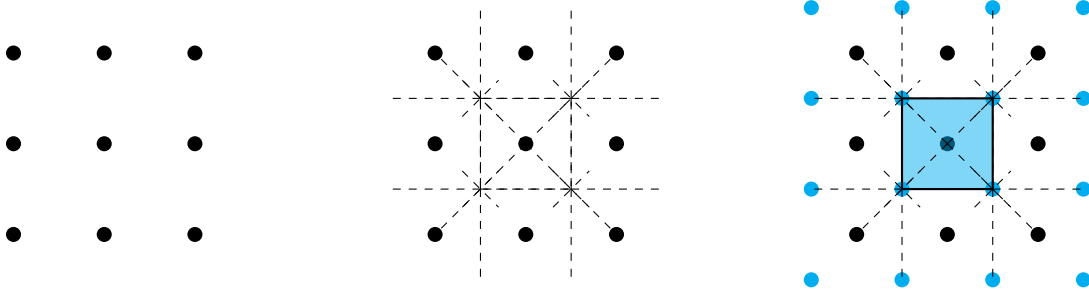


Figura 2.6: Construcción de la zona Brillouin en dos dimensiones utilizando líneas a puntos vecinos y planos normales para redes recíprocas cuadradas y hexagonales

2.1.5. Teorema de Bloch

Para describir la celda unidad de un sólido se necesita tener una función de onda que la describa la celda unidad y que cumpla con sus condiciones periódicas. Así, al trasladar un punto $\vec{r}$ aplicando el operador de traslación $\hat{T}$ a un punto $\vec{r} + \vec{R}$ de una celda replicada, tendremos la misma expresión de la función de onda salvo por un cambio de fase:

$$\hat{T}\Psi(\vec{r}) = \Psi(\vec{r} + \vec{R}) = e^{i\vec{k} \cdot \vec{R}}\Psi(\vec{r})\tag{2.2}$$

Así, el valor de la función de onda de un punto no se ve afectado al ser trasladada a su punto

equivalente en cada celda replicada. El responsable de que esto sea así es el potencial externo $V(\vec{r})$, que es el que presenta una periodicidad y determina la densidad electrónica dentro de la celda unidad y a la larga, toda la estructura electrónica.

El operador de traslación $\hat{T}$ conmuta con el Hamiltoniano (Hamiltoniano es invariante bajo traslaciones $[\hat{H}, \hat{\mathbf{T}}_{\mathbf{n}}] = 0$), por lo tanto las funciones de onda pueden ser funciones propias (eigenfunciones) de los dos operadores.

El teorema de Bloch postula que en un sistema periódico cada función de onda electrónica se puede expresar como un producto de dos funciones, una con la periodicidad intrínseca de la celda y otra que es una onda plana:

$$\Psi_j(\vec{r}) = e^{i\vec{k} \cdot \vec{r}} \zeta_j(\vec{r}) \quad (2.3)$$

La primera parte es una onda plana con vector de onda $\vec{k}$ y $\zeta(\vec{r})$ es la función periódica, que se puede expandir como una combinación lineal de ondas planas a su vez, pero cuyos vectores de onda $\vec{G}$ son recíprocos a los vectores de la red directa $\vec{a}$:

$$\zeta_j(\vec{r}) = \sum_{\vec{G}} c_{j,\vec{G}} e^{i\vec{G} \cdot \vec{r}}; \quad \vec{G} \cdot \vec{a}_i = 2\pi \quad \forall m \in \mathbb{Z}. \quad (2.4)$$

Considerando eq. (2.4) y el operador traslacional $\hat{T} = n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3$ obtenemos una expresión donde la función de onda se escribe como una suma de ondas planas:

$$\Psi_j(\vec{r}) = \sum_{\vec{G}} c_{j,\vec{k}+\vec{G}} e^{i(\vec{k}+\vec{G}) \cdot \vec{r}}. \quad (2.5)$$

De esta manera vemos la utilidad de representar las funciones de onda del sistema real como una suma de ondas planas del espacio recíproco. Cada onda plana viene caracterizada por una energía cinética concreta $\frac{\hbar^2}{2m} |\vec{k} + \vec{G}|^2$. De esta manera, se puede aumentar la base de ondas planas definiendo un límite para su energía cinética, y teniendo así una base finita de funciones base. Es usual aumentar la base hasta observar convergencia en un observable sensible, como es la energía.

2.2. Modelo teórico

Las primeras aplicaciones del formalismo de la mecánica cuántica, desarrollado durante los años 20's del siglo pasado y décadas antes del primer computador, fueron restringidas a sistemas atómicos, diatómicos o bien, altamente simétricos los cuales podían ser resueltos sin la necesidad de un computador, obteniendo en esos casos soluciones analíticas. Sin embargo, la mayoría de los sistemas de interés práctico son demasiado complejos para ser resueltos de manera analítica.

Hoy en día, el desarrollo de métodos numéricos aplicables a la resolución de las ecuaciones que surgen en el modelado molecular y que son aplicables a estos problemas mediante el uso de los computadores, han permitido que la mecánica cuántica pueda ser usada para efectuar cálculos sobre sistemas moleculares de interés real y práctico.

Con la mecánica cuántica es posible estudiar el comportamiento de los electrones en un sistema atómico o molecular y, por lo tanto, es posible obtener propiedades que dependen de la distribución electrónica, así como también investigar reacciones químicas en las cuales se pueden formar o romper enlaces. Hoy en día se han desarrollado varias aproximaciones a la teoría cuántica para tratar los sistemas moleculares con la finalidad de analizar las propiedades estructurales y electrónicas de nuestros sistemas. En este trabajo se utilizó la teoría del funcional de la densidad (DFT, por sus siglas en inglés), la cual es de crucial para describir la energía de los sistemas moleculares de interés en la ciencia de materiales.

2.2.1. Teoría del Funcional de Densidad

Uno de los enfoques más extendidos para investigar la estructura electrónica de los sistemas cuánticos de muchas partículas es la teoría del funcional de densidad (DFT), que es particularmente útil para metales y aleaciones, proporcionando un atractivo equilibrio entre la precisión de los cálculos y la velocidad computacional. Este método lo utilizamos en la investigación para calcular propiedades de aleaciones de Titanio y Tántalo mediante el uso del software *VASP* (presentado en [Subsec. 2.3.2](#)).

La Teoría del Funcional de la densidad se basa en los teoremas de Hohenberg-Kohn [8]:

Teorema 2.2.1 *El potencial externo (y por lo tanto la energía total) es una función única de la densidad de electrones.*

Teorema 2.2.2 *La función de energía que entrega la energía del estado fundamental del sistema, da la energía más baja si y solo si la densidad de entrada es la verdadera densidad del estado fundamental.*

Los teoremas de Hohenberg-Kohn ([Teorema 2.2.1](#) y [Teorema 2.2.2](#)) permiten reformular la ecuación de Schrödinger independiente del tiempo de un átomo para un sistema de N_a átomos (núcleos) con cargas Z_j y N_e electrones [eq. \(B.28\)](#) en términos de la función de densidad de electrones en lugar de la función de onda $\Psi(r_1, \dots, r_{N_e}, R_1, \dots, R_{N_a})$:

$$\begin{aligned}
 E[n(r)] = T[n(r)] &+ \sum_{j=1}^{N_{\text{átomos}}} \int \frac{e^2 Z_j n(r)}{|r - R_j|} d^3(r) \\
 &+ \int \int \frac{e^2 n(r) n(r')}{|r - r'|} d^3(r) d^3(r') \\
 &+ E_{XC}[n(r)] + \int V_{ext}(r) n(r) d^3(r)
 \end{aligned} \tag{2.6}$$

donde $T[n(r)]$ es el término de energía cinética, los integrales simples y dobles son la repulsión Coulombiana electrón-núcleo y electrón-electrón respectivamente, $V_{ext}(r)$ es el potencial externo, y el término $E_{XC}[n(r)]$ incluye otros factores no precisamente conocidos como la interacción de intercambio-correlación.

La densidad electrónica $n(r)$ se define como:

$$n(r) = \sum_{i=1}^{N_e} |\psi_i|^2.$$

Es importante tener en cuenta que si se consideran los grados de libertad de spin, se pueden introducir densidades de spin-up y spin-down, calculadas a partir de las correspondientes funciones de onda de spin-up y spin-down ψ_i .

Dado que la forma exacta del término $EXC[n(r)]$ es desconocida, podemos reescribir la ecuación [eq. \(2.6\)](#) que describe N_e partículas interactuantes en algún potencial externo $V_{ext}(r)$, como un sistema de ecuaciones de partículas no interactuantes ([Fig. 2.7](#)) en otro potencial, el potencial efectivo de Kohn-Sham

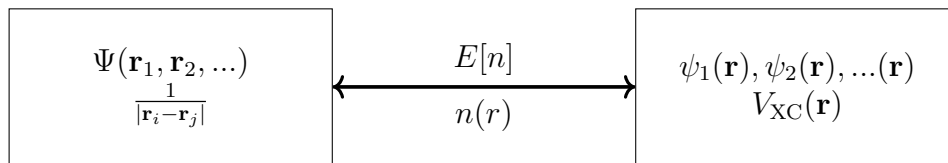


Figura 2.7: Sistema interactivo y no interactivo

donde el sistema de ecuaciones de partículas no interactuantes, que se conoce como ecuaciones de Kohn-Sham [26], se puede escribir como:

$$\left(-\frac{\hbar^2}{2m_e}\nabla_i^2 + V_{KS}(r)\right)\psi_i(r) = e_i\psi_i(r), i \leq Ne, \quad (2.7)$$

donde el potencial de Kohn-Sham incluye los funcionales correspondientes al potencial externo $V_{ext}(r)$, la interacción Coulombiana de electrones $\int \frac{e^2 n(r')}{|r-r'|} d^3 r'$ y la interacción de intercambio-correlación entre electrones $V_{XC}[n(r)]$.

$$V_{KS}(r) = V_{ext}(r) + \int \frac{e^2 n(r')}{|r-r'|} d^3 r' + V_{XC}[n(r)] \quad (2.8)$$

La forma de $V_{XC}[n(r)]$ aún se desconoce, y todos los efectos no incluidos explícitamente en eq. (2.8) se incluyen en $V_{XC}[n(r)]$.

De esta manera, el problema de muchas partículas eq. (B.28) se reemplaza por un conjunto de problemas de una sola partícula eq. (2.7) sin ninguna pérdida de generalidad. Para resolver eq. (2.7) necesitamos encontrar las funciones de onda ψ_i , asumiendo que conocemos $V_{KS}(r)$. El potencial $V_{KS}(r)$ depende de la distribución de densidad electrónica $n(r)$, que está definida únicamente por un conjunto de ψ_i . Por lo tanto, las ecuaciones eq. (2.7) deben resolverse de manera autoconsistente, es decir, iterativamente.

Usualmente se comienza con una suposición inicial para $n(r)$, luego se calcula $V_{KS}(r)$ y se resuelven las eq. (2.7) con respecto a ψ_i , lo que proporciona la nueva distribución de densidad $n(r)$. Estos pasos se repiten hasta alcanzar la convergencia. Esto se puede representar como un diagrama de flujo, como se muestra en la Fig. 2.8:

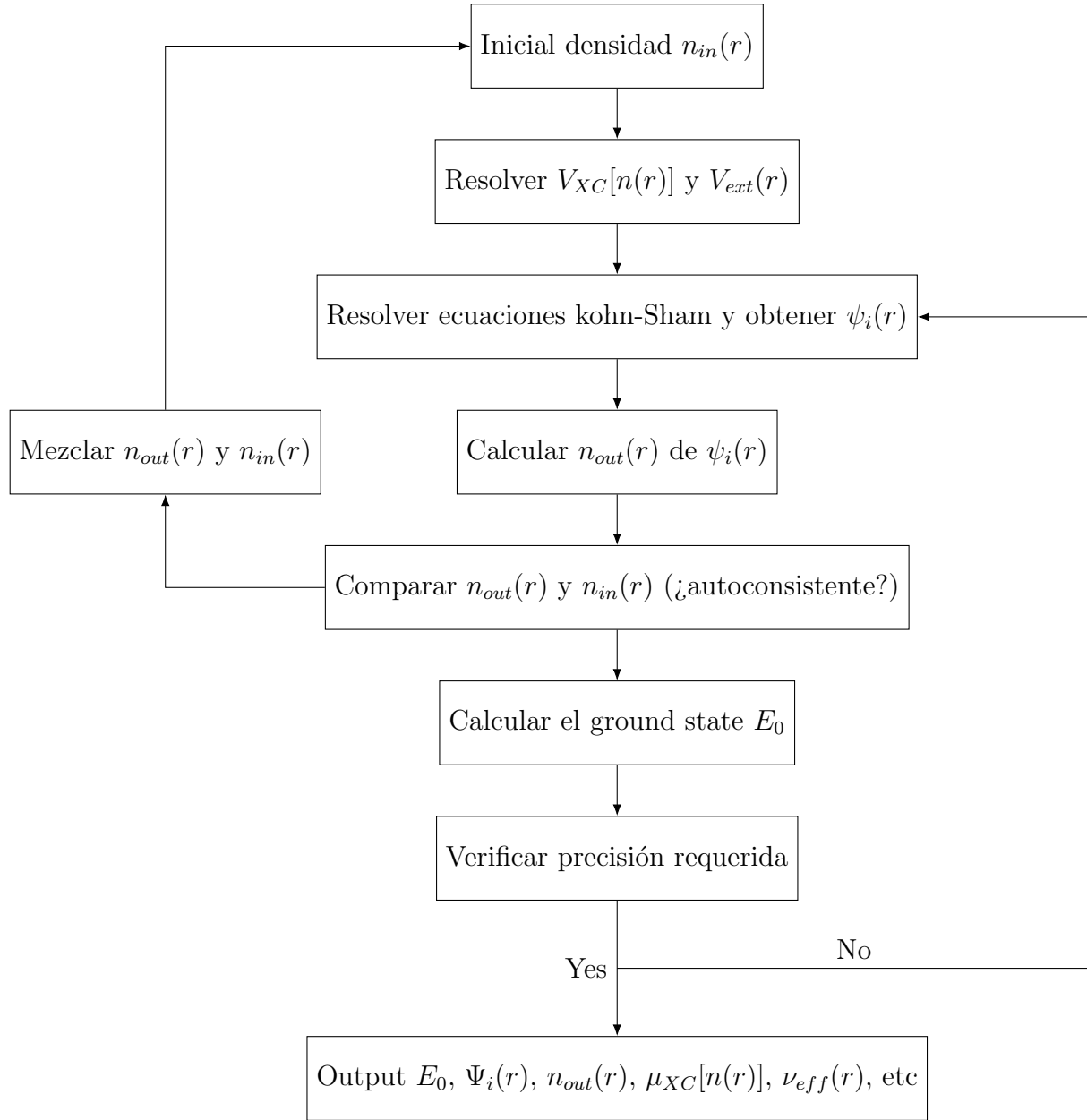


Figura 2.8: El diagrama de flujo para resolver de forma autoconsistente las ecuaciones de Kohn-Sham para obtener la energía del estado fundamental.

donde la autoconsistencia o *self-consistence* fue por primera vez introducida por Hartree en 1927 [7]:

“If the final field is the same as the initial field, the field will be called ‘self-consistent’, and the determination of selfconsistent fields for various atoms is the main object of this paper”.

Dada la forma desconocida de $V_{XC}[n(r)]$, se introducen diferentes aproximaciones. La “escalera de Jacob”, mostrada en la Fig. 2.9 es una forma de mostrar familias de aproximaciones funcionales de intercambio-correlación DFT [20].

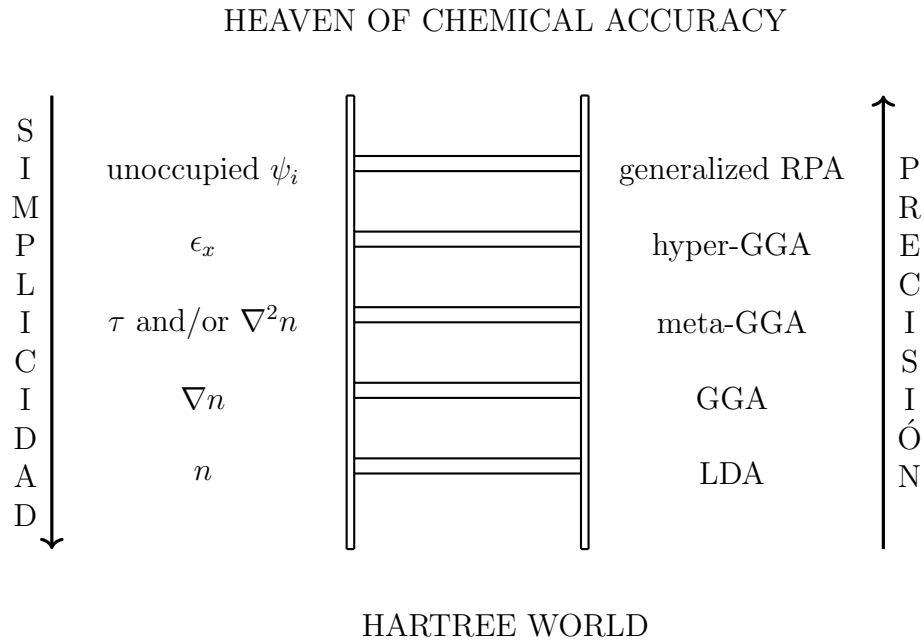


Figura 2.9: Escalera de Jacob

La más simple, y computacionalmente eficiente, es la aproximación de densidad local (LDA), que asume que $V_{XC}(r)$ depende solo de $n(r)$ en ese punto del espacio, es decir, es local: $V_{XC}(r) = V_{XC}(n(r))$. La aproximación de gradiente generalizado (GGA) es más exigente computacionalmente, pero en muchos casos mucho más precisa. Toma en cuenta la dependencia del derivado espacial de la densidad electrónica en un cierto punto: $V_{XC}(r) = V_{XC}(n(r), \nabla n(r))$.

Los cálculos *ab-initio* pueden ser extremadamente costosos en términos del requerimiento computacional necesarios. A pesar de eso, los avances en infraestructura de computación y la disponibilidad de softwares fáciles de manejar han ayudado a que los métodos *ab-initio* sean una herramienta computacional usada ampliamente.

2.3. Softwares

2.3.1. Generación de estructura

Hay muchos problemas en la ciencia de los materiales que involucran la solución de dos o más materiales cristalinos y la ocupación de los sitios de red. Muchas propiedades únicas de los materiales de

estado sólido aparecen solo en un estado desordenado y/o defectuoso. De los diversos tipos de desorden que existen a diferentes escalas de longitud en compuestos reales, las impurezas atómicas locales, las sustituciones y/o vacantes se encuentran entre las más importantes [19]. Estos defectos puntuales son en particular responsables de las propiedades únicas de una serie de semiconductores, superconductores de alta temperatura, aleaciones metálicas, cerámicas (incluidas las piezoeléctricas), catalizadores de zeolita y muchos otros tipos de materiales tecnológicamente importantes.

La búsqueda exhaustiva de posibles configuraciones atómicas en una *supercell* dada se basa en métodos de *combinatoria enumerativa*, donde el problema consiste en distribuir los átomos entre uno o varios conjuntos de sitios cristalográficos desordenados. Para Abordar este último problema existen distintos enfoques, los que explicaremos brevemente a continuación,

Grupos de simetría

Esta metodología se basa en la determinación y comparación del grupo espacial de las diferentes estructuras. Sin embargo, el hecho de que dos configuraciones tengan el mismo grupo espacial no implica que vayan a tener la misma energía [25] y por lo tanto, equivalentes. Tener el mismo grupo espacial es una condición necesaria pero no suficiente para que dos configuraciones sean simétricamente equivalentes.

Enfoque muestral

En un enfoque muestral el espacio configuracional se genera utilizando el método de Monte Carlo, donde las configuraciones se generan aleatoriamente para luego ser aceptadas o rechazadas en el conjunto de acuerdo con el algoritmo Metropolis. De esta manera, es posible obtener un conjunto representativo de configuraciones, que puede ser considerablemente más pequeño que el espacio configuracional completo.

Supercell

Muchas configuraciones combinatoriamente distintas son geoméricamente idénticas: están relacionadas por las operaciones de simetría de la red subyacente. La idea principal de la búsqueda de configuraciones irreducibles es generar una lista exhaustiva de configuraciones combinatoriamente distintas, eliminando luego los duplicados, es decir, las configuraciones simétricamente equivalentes.

El número de configuraciones posibles aumenta drásticamente con el tamaño de la supercell. Una limitación importante del enfoque de *supercell* es el costo computacional. Un enfoque alternativo para reducir el número de configuraciones es aprovechar la simetría del sistema: muchas configuraciones están relacionadas con la simetría y, por lo tanto, son idénticas, por lo que es posible limitar el espacio de configuración solo a las configuraciones simétricamente inequivalentes. El algoritmo implementado en el software SOD [23] es uno de los excelentes ejemplos [12].

SOD

Sod ocupa en el criterio de equivalencia entre dos configuraciones, que se basa en el concepto de transformaciones isométricas. Estas son operaciones geométricas (por ejemplo, traslaciones, rotaciones, reflexiones) que mantienen constantes todas las distancias y ángulos dentro del objeto transformado. Por lo tanto, dos configuraciones se consideran equivalentes si hay una transformación isométrica que convierte una en la otra. Cada configuración mantendrá los mismos operadores de simetría de la estructura principal, excepto por la ruptura de simetría causada por la propia sustitución.

2.3.2. Simulaciones de dinámica molecular *ab initio*

En los últimos años, los cálculos *ab initio* basados en Kohn-Sham (KS), la teoría del funcional de la densidad, han ganado un enorme interés no sólo entre los físicos del estado sólido, sino también entre los químicos. En parte, esto se debe al gran éxito de la aproximación de densidad local (LDA) [10]. Las simulaciones por computador basadas en una descripción de la mecánica cuántica de las interacciones entre los electrones y entre los electrones y los núcleos atómicos han desarrollado un impacto cada vez más importante en la química, en la física del estado sólido y en la ciencia de los materiales, promoviendo no solo una comprensión más profunda, sino también la posibilidad de contribuir significativamente al diseño de materiales para tecnologías futuras. Este desarrollo tiene dos pilares fundamentales:

- La descripción mejorada de los efectos electrónicos de muchos cuerpos dentro de la teoría del funcional de la densidad (DFT) y los métodos post DFT.
- La implementación de las nuevas funcionalidades y técnicas de muchos cuerpos dentro de códigos informáticos altamente eficientes, estables y versátiles, que permitan explotar el potencial de las arquitecturas informáticas modernas.

VASP

Vienna Ab initio Simulation Package (VASP) es un software computacional para realizar simulaciones de mecánica cuántica. VASP calcula una solución aproximada de la ecuación de Schrödinger de muchos cuerpos dentro del marco de la *teoría del funcional de la densidad* a través de las ecuaciones de Kohn-Sham o utilizando la aproximación de Hartree-Fock. Este paquete de software utiliza condiciones de contorno periódicas y un método pseudopotencial con una base de onda plana establecida.

VASP utiliza un conjunto básico de ondas planas en lugar de un conjunto básico localizado. El método PAW o *pseudopotenciales ultrablandos* se utiliza para describir la interacción entre el ion y el electrón de no valencia de modo que se pueda reducir el tamaño del conjunto de bases. Los ciclos de autoconsistencia y las opciones de métodos numéricos para calcular el estado fundamental electrónico de Kohn-Sham en VASP son bastante eficientes y robustos. Se utilizan condiciones de contorno periódicas para tratar un número infinito de átomos [21] [5]

Pseudopotenciales

Los pseudopotenciales aprovechan el hecho de que el enlace y la interacción entre átomos dependen principalmente de los electrones de valencia. Un pseudopotencial combina el potencial del núcleo con el potencial de los electrones en las capas llenas. Esto reduce el número de electrones que VASP considera, reduciendo la intensidad computacional. Hay varios métodos para generar los pseudopotenciales y para asegurar que los resultados sean consistentes y puedan relacionarse entre sí, el grupo VASP ha estandarizado los pseudopotenciales utilizados por VASP.

Base de Ondas Planas

Mediante el uso de transformadas de Fourier, la función de onda y la densidad de electrones pueden comprimirse expresándolas en términos de una base de ondas planas. Las ondas seno y coseno ($e^{i\omega t}$) en una dimensión se reemplazan por ondas planas ($e^{ik \cdot r}$) en tres dimensiones. VASP expresa tanto las funciones de onda como la densidad de carga en términos de una base de ondas planas, reduciendo enormemente la memoria requerida para ejecutar VASP. A diferencia de los pseudopotenciales, VASP requiere que el usuario especifique qué ondas planas quiere usar para su cálculo. Dado que las ondas planas se distinguen por su vector k , seleccionar ondas planas es idéntico a seleccionar puntos en la zona de Brillouin. Los puntos seleccionados se llaman puntos k , porque corresponden a un vector k particular. Para obtener una buena representación de toda la zona de Brillouin, los puntos k se eligen típicamente a intervalos regulares formando una matriz

tridimensional llamada la malla k . En VASP, estos puntos pueden ser especificados explícitamente o generados automáticamente. Al igual que con el sonido, si se usan más ondas planas, se obtiene un cálculo más preciso, pero que requiere más memoria y es más intensivo computacionalmente.

k-points

De lo discutido [Subsec. 2.1.5](#), cada vector recíproco de las zonas de Brillouin se puede expresar como el vector de la primera zona de Brillouin. Por lo tanto, sólo consideramos la primera zona de Brillouin para capturar todos los vectores. La primera zona de Brillouin puede ser reducida aún más mediante operaciones de simetría sin perder ninguna información.

Detalles sobre el manejo de archivos necesarios para ejecutar el software se encuentran en [Apéndice C](#).

2.4. Aprendizaje Automático

2.4.1. Flujo de trabajo de Aprendizaje Automático en el diseño de aleaciones

La búsqueda de métodos efectivos para diseñar aleaciones de alto rendimiento es altamente deseada. Se requiere mucho tiempo y esfuerzo para calcular las propiedades de las aleaciones con principios de primeros principios. El aprendizaje automático (AA) [Apéndice A.1](#) se puede utilizar para construir un modelo a partir de los datos proporcionados. El propósito principal del modelo de AA es predecir las propiedades objetivo de las aleaciones candidatas que pueden ser sintetizadas. El flujo de trabajo de AA en el diseño de aleaciones incluye identificación de objetivos, preparación de datos, ingeniería de características, selección de modelos y aplicación de modelos. El flujo de trabajo se puede ver en la [Fig. 2.10](#).

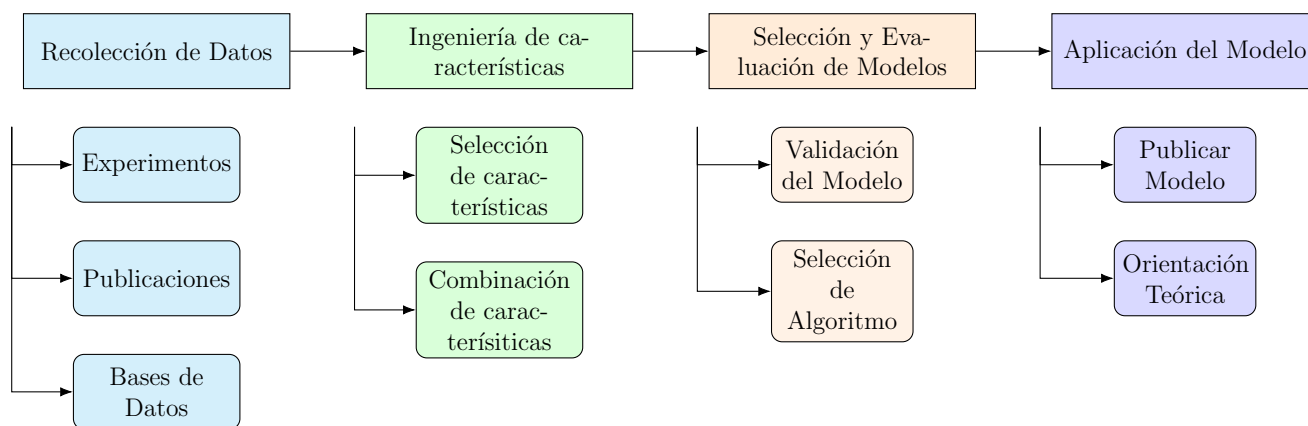


Figura 2.10: Workflow de aprendizaje automático en el diseño de aleaciones

Identificación de objetivos

El primer paso de cualquier tecnología de AA es determinar los objetivos de investigación en cuanto a las propiedades del material, generalmente dependiendo de la experiencia del experto y del conocimiento del dominio, así como de los requisitos de desarrollo del material. Este paso es posiblemente el más importante, ya que las propiedades objetivo deben ser potencialmente aprendibles a partir de la información disponible. Además, las propiedades objetivo deben estar definidas de manera unívoca según los experimentos y/o simulaciones. Por lo tanto, es crucial identificar las propiedades objetivo de las aleaciones antes de la selección de las herramientas de AA.

Preparación de datos

Los datos se pueden obtener de bases de datos públicas, publicaciones y autogeneración. Específicamente, los datos de las aleaciones incluyen entradas (características) y salidas (propiedades objetivo). Las entradas consisten en composiciones elementales, parámetros atómicos, parámetros estructurales y condiciones de procesamiento, etc. Las propiedades objetivo de las aleaciones típicamente contienen dureza, resistencia a la corrosión, conductividad eléctrica, resistencia a la tracción y capacidad de formación de fase, etc.

El rendimiento de los modelos de AA está altamente correlacionado con la calidad y cantidad de conjuntos de datos. Los requisitos de cantidad de datos generalmente dependen de la selección de algoritmos de AA. El algoritmo de aprendizaje profundo y las redes neuronales requieren un tamaño de conjunto de datos mucho mayor, mientras que las máquinas de soporte vectorial (SVM) tienden a ser la selección óptima de AA para un tamaño pequeño de conjunto de datos. Además, la calidad del conjunto

de datos está asociada con la incertidumbre relacionada con los datos. Generalmente, los errores de datos podrían ser causados por cálculos o experimentos.

Dado que los datos originales a menudo se muestran con incompletitud y ruido, el preprocesamiento de datos para adquirir modelos de AA confiables debe realizarse antes de la implementación del modelo. Generalmente, el preprocesamiento de datos incluye la eliminación de datos erróneos o incompatibles, la normalización de datos originales, etc. Además, la fuente de datos de publicaciones o bases de datos debe ser cuidadosamente verificada para garantizar la precisión y autenticidad.

Ingeniería de características

Las excelentes características proporcionan una base firme para los modelos robustos de AA. Las características razonables deben tener los siguientes tres estándares:

- Únicas para las características de la aleación
- Sensibles a los valores del objetivo
- Fácilmente accesibles.

La selección de características tiene como objetivo encontrar un subconjunto efectivo para entrenar el modelo, conservando características útiles y eliminando características redundantes o irrelevantes. El subconjunto efectivo posee las ventajas de acortar el tiempo de entrenamiento, evitar la maldición de la dimensionalidad, reducir el sobreajuste y hacer que el modelo sea simple, interpretable y universal. Los métodos de selección de características ampliamente utilizados son la selección de características supervisada y no supervisada.

Selección de modelos

Es fundamental construir modelos óptimos que representen los datos existentes sin incurrir en sobreajuste, que ocurre cuando un modelo de AA se ajusta demasiado bien a los datos de entrenamiento y pierde capacidad para generalizar a nuevos datos no vistos anteriormente, o subajuste, que significa que el modelo es demasiado simple para capturar la información general del conjunto de entrenamiento, lo que resulta en una gran desviación. Es importante destacar **que las Máquinas de Soporte Vectorial (SVM) son un algoritmo especialmente apropiado para conjuntos de datos de pequeño**

tamaño, que se pueden usar para evitar el sobreajuste en dichos conjuntos de datos. Los algoritmos de AA más utilizados están representados en la [fig. 2.11](#).

En general, la selección de un algoritmo adecuado es clave para construir modelos de *Aprendizaje Automático* (AA). Según los estilos de aprendizaje, los algoritmos de AA podrían clasificarse en tres tipos:

- Aprendizaje Supervisado
- Aprendizaje No supervisado
- Aprendizaje Semi-supervisado
- Aprendizaje por Refuerzo

Cuando se realizan procesos de AA, los datos se pueden dividir en conjunto de entrenamiento y conjunto de prueba basados en la proporción. El conjunto de entrenamiento se utiliza generalmente para aprender patrones ocultos, mientras que el conjunto de prueba se aplica para evaluar la capacidad de generalización del modelo. La validación cruzada y el muestreo con reemplazo se aplican para evaluar la confiabilidad de los modelos de AA. Para los modelos de regresión, el coeficiente de determinación (R^2), el coeficiente de correlación de Pearson (r) y el error cuadrático medio (ECM) se ven como métricas de evaluación. Generalmente, un ECM más bajo sugiere un mayor rendimiento del modelo, mientras que un R^2 y r más altos representan un mejor ajuste del modelo. Para los modelos de clasificación, la matriz de confusión, la precisión y la curva característica operativa del receptor se utilizan como índices de evaluación. El modelo óptimo se obtiene de la validación cruzada, y el conjunto de prueba se explota aún más para evaluar la generalizabilidad del modelo al predecir las propiedades de nuevos materiales.

Aplicación del modelo

La aplicación del modelo significa cómo aplicar completamente el modelo para descubrir aleaciones candidatas, proporcionando orientación teórica para los experimentadores. Una forma de aplicar modelos de AA es se es diseñar las posibles muestras virtuales en el espacio de muestra original. Luego, los modelos de AA se utilizan para predecir las propiedades de las muestras virtuales. Finalmente, las muestras seleccionadas se proporcionan para síntesis experimental y caracterización para descubrir nuevas aleaciones con propiedades extraordinarias.

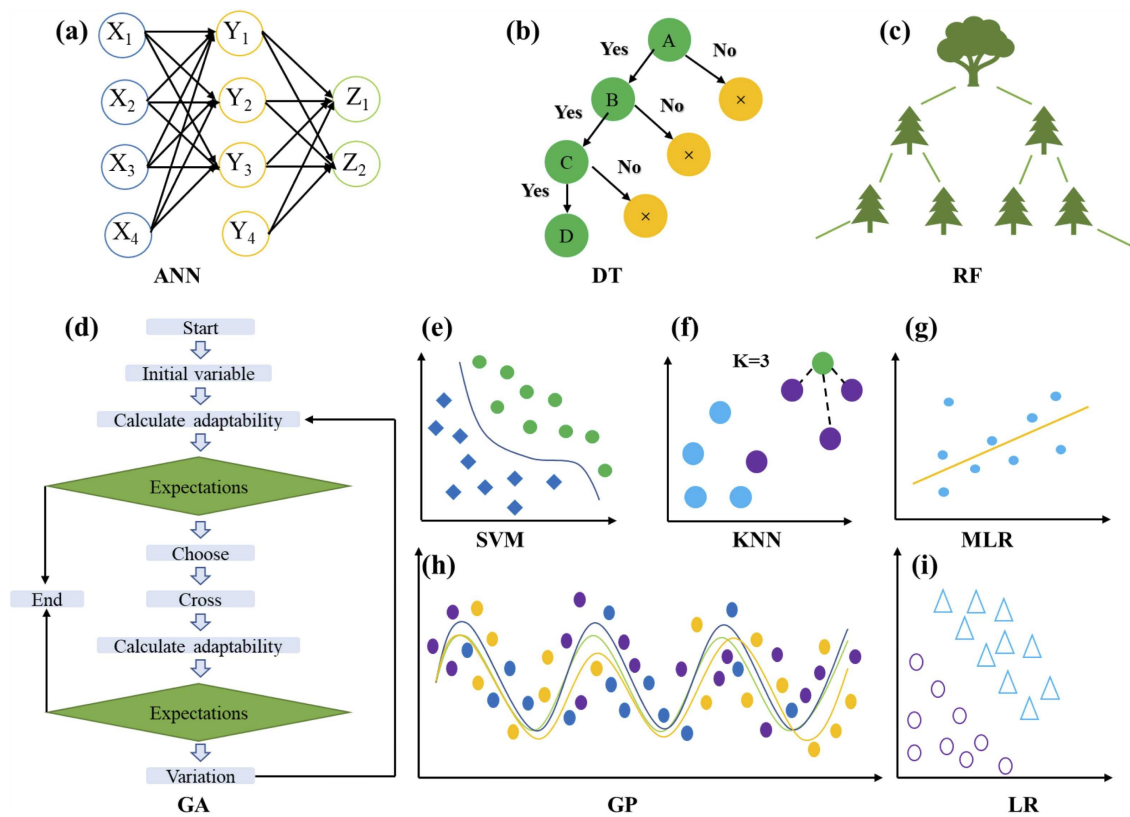


Figura 2.11: [14] Diagramas esquemáticos de los algoritmos de aprendizaje automático comunes en el diseño de aleaciones. ANN: Redes neuronales artificiales. DT: Árbol de decisión. RF: Bosques aleatorios. GA: Algoritmo genético. SVM: máquina de vectores de soporte. KNN: K-regresión del vecino más cercano. MLR: Regresión logística múltiple. GP: regresión del proceso gaussiano. LR: Regresión lineal

2.4.2. Tipos

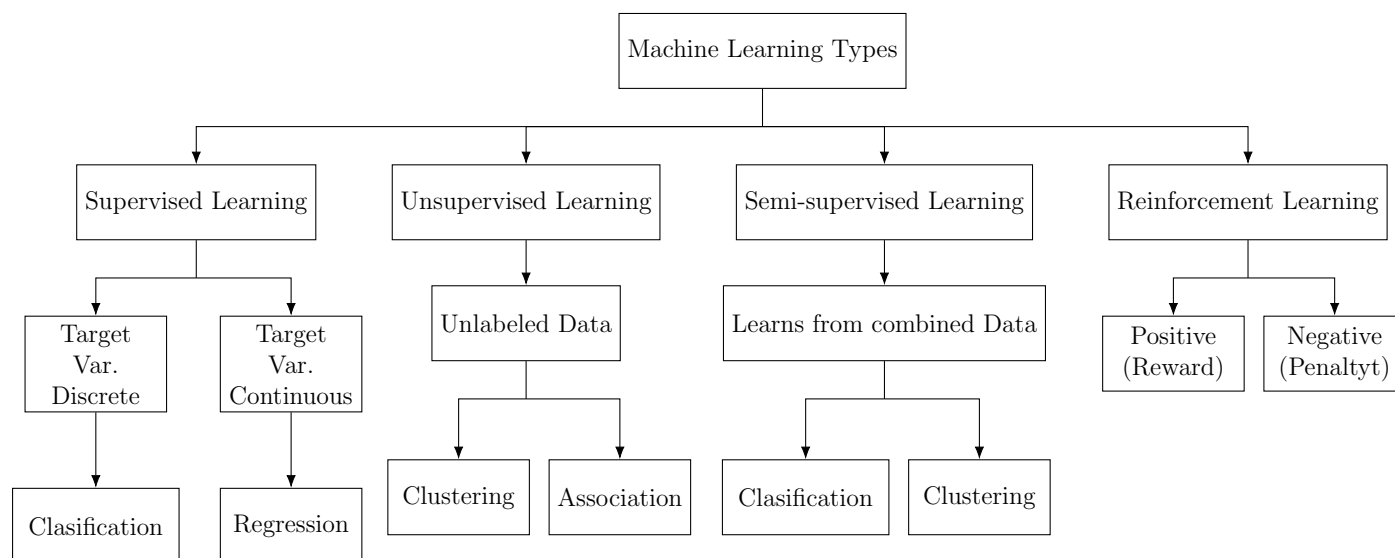


Figura 2.12: Tipos de aprendizaje automático

Supervisado

El *aprendizaje supervisado* es típicamente la tarea de aprender una función que mapea una entrada a una salida basada en pares de ejemplos de entrada-salida. Utiliza datos de entrenamiento etiquetados y una colección de ejemplos de entrenamiento para inferir una función. El aprendizaje supervisado se lleva a cabo cuando se identifican ciertos objetivos a lograr a partir de un conjunto determinado de entradas, es decir, un enfoque orientado a tareas. Las tareas supervisadas más comunes son la “clasificación”, que separa los datos, y la “regresión”, que ajusta los datos. Por ejemplo, predecir la etiqueta de clase o el sentimiento de un texto, como un tweet o una reseña de producto, es decir, la clasificación de texto, es un ejemplo de aprendizaje supervisado.

No Supervisado

El *aprendizaje no supervisado* analiza conjuntos de datos no etiquetados sin la necesidad de intervención humana, es decir, un proceso impulsado por datos. Se utiliza ampliamente para extraer características generativas, identificar tendencias y estructuras significativas, agrupaciones en resultados y fines exploratorios. Las tareas de aprendizaje no supervisadas más comunes son el agrupamiento, la estimación de densidad, el aprendizaje de características, la reducción de dimensionalidad, la búsqueda de reglas de asociación, la detección de anomalías, etc.

Semi-supervisado

El *aprendizaje semi-supervisado* puede definirse como una hibridación de los métodos supervisados y no supervisados mencionados anteriormente, ya que opera tanto en datos etiquetados como no etiquetados. Por lo tanto, se sitúa entre el aprendizaje “sin supervisión” y el aprendizaje “con supervisión”. En el mundo real, los datos etiquetados pueden ser raros en varios contextos y los datos no etiquetados son numerosos, donde el aprendizaje semi-supervisado es útil. El objetivo final de un modelo de aprendizaje semi-supervisado es proporcionar un mejor resultado para la predicción que el producido utilizando solo los datos etiquetados del modelo. Algunas áreas de aplicación donde se utiliza el aprendizaje semi-supervisado incluyen la traducción automática, la detección de fraude, el etiquetado de datos y la clasificación de texto.

Reforzamiento

El *aprendizaje por refuerzo* es un tipo de algoritmo de aprendizaje automático que permite a agentes de software y máquinas evaluar automáticamente el comportamiento óptimo en un contexto o entorno particular para mejorar su eficiencia, es decir, un enfoque impulsado por el entorno. Este tipo de aprendizaje se basa en la recompensa o penalización, y su objetivo final es utilizar las percepciones obtenidas de las actividades ambientales para tomar medidas que aumenten la recompensa o minimicen el riesgo. Es una herramienta poderosa para entrenar modelos de IA que pueden ayudar a aumentar la automatización u optimizar la eficiencia operativa de sistemas sofisticados como la robótica, tareas de conducción autónoma, manufactura y logística de la cadena de suministro, sin embargo, no se prefiere su uso para resolver problemas básicos o sencillos.

Capítulo 3

Metodología

Nuestra metodología de trabajo se puede resumir en la [Fig. 3.1](#), donde se muestra cada una de las etapas de esta investigación y que se discutirán en las siguientes secciones.

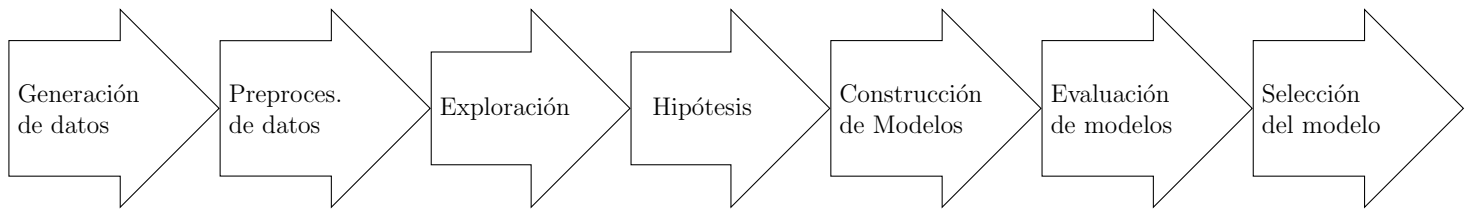


Figura 3.1: Metodología

3.1. Generación de datos

Luego de definir el problema, nuestro trabajo fue generar los archivos de las configuraciones posibles dado un número de impurezas (átomos de Ti). Inicialmente generamos estos archivos para 1, 2, 3 y 4 impurezas por medio del uso del software SOD [\[23\]](#). Este software, como fue discutido en [Subsec. 2.3.1](#) reduce considerablemente la cantidad de configuraciones resultantes (ver [Fig. 3.2](#)).

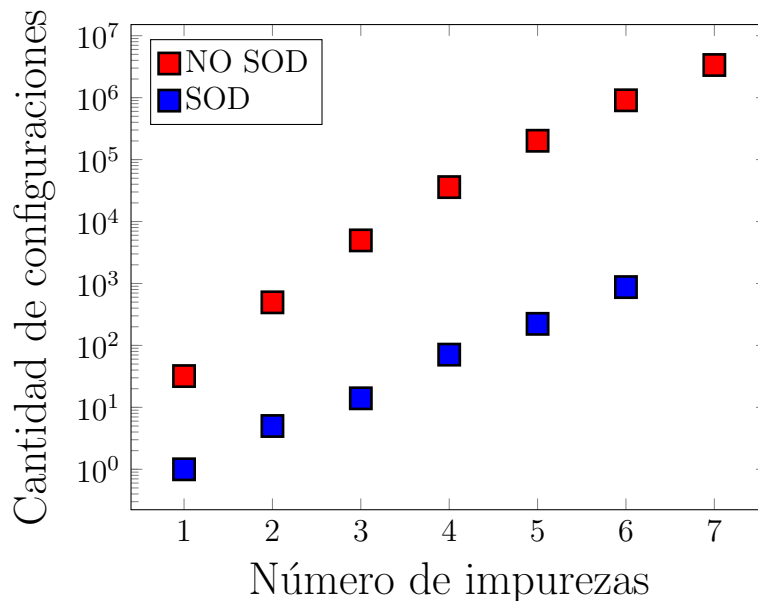


Figura 3.2: Cantidad de configuraciones posibles para 1, 2, 3 y 4 impurezas.

Estos archivos, generados por SOD, fueron usados para calcular las respectivas energías de cada configuración por medio de simulaciones de dinámica molecular hechas por el software VASP.

3.2. Preprocesamiento de datos

Con nuestros datos generados (archivos para las configuraciones y archivos resultantes de las simulaciones), lo siguiente es preprocesar nuestros datos iniciales con el uso de herramientas computacionales. En particular se usó el language de programación [Python](#) y sus librerías [Pandas](#) [24], [ASE](#) [11] y [NumPy](#) [6], pudiendo crear las funciones necesarias para preprocesar los archivos de configuraciones y los archivos de energías.

Como resultado del preprocesamiento obtenemos un archivo `.csv` con las configuraciones y sus respectivas características, donde inicialmente trabajamos con mas de 20 características, pero las principales son:

- `n_impurezas`: Número de impurezas
- `Efree`: Energía de la configuración entregada por VASP

- **vectors**: Vectores posición de los átomos
- **distances_i**: Distancias entre las impurezas
- **angles_i**¹: Ángulos entre los vectores que unen las impurezas
- **meandistance**: Distancia promedio entre las impurezas

De las características más importantes tenemos a **meandistance**, que es la distancia promedio entre las impurezas según nuestro criterio, el cual es discutido a continuación ([Subsec. 3.2.1](#)).

3.2.1. El descriptor **meandistance**

El descriptor **meandistances** se calcula como la media aritmética de las distancias entre las impurezas, por lo tanto cada configuración tiene asociado un valor de **meandistance**. Para explicar esto de mejor manera, debemos observar las dos subfiguras superiores en la [Fig. 3.3](#), donde en la izquierda superior está la representación de una configuración en particular de $\text{Ti}_{29}\text{Ta}_3$ en una celda descrita por vectores ortogonales y de igual magnitud, donde las esferas azul representan las impurezas (átomos de Ta) y las grises las de Ti. Para poder mostrar de manera más clara la construcción de este descriptor, en la esquina superior derecha se encuentra la misma configuración anterior, pero sólo con las impurezas.

¹Este descriptor sólo es válido para un número de impurezas mayor a 2, ya que para un número de impurezas igual a 2 o 1 no existe tal ángulo.

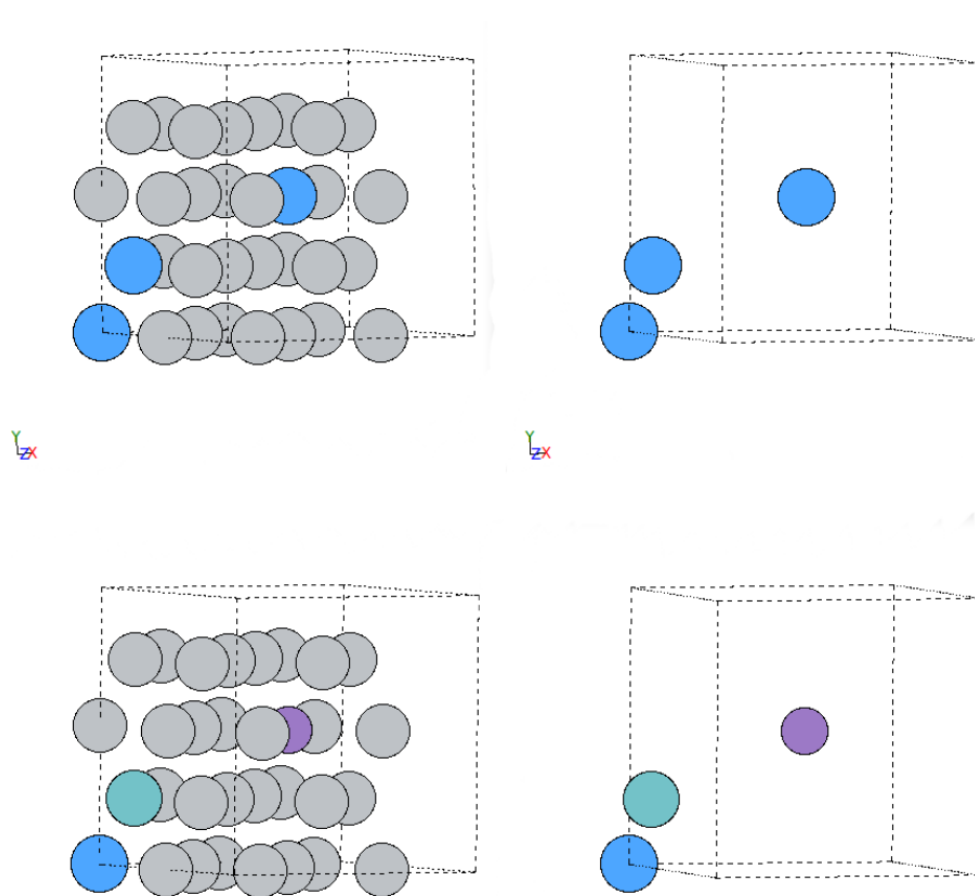


Figura 3.3: La figura corresponde a la configuración c00010 para 3 impurezas, donde en la columna izquierda se observan los 32 átomos (esferas grises corresponden a los átomos de T_i y las azules a los átomos de T_a) y en la columna derecha sólo las impurezas.

Para entender cómo se calcula **meandistance**, primero etiquetamos las impurezas con “colores”, lo cual se puede observar en las dos subfiguras inferiores de Fig. 3.3, donde cada una de ellas es la subfigura superior, pero con las impurezas etiquetadas por “color”. Luego, con el fin de poder explicar el cálculo del descriptor en cuestión, las etiquetamos con un número y así obtenemos la configuración del lado izquierdo en la Fig. 3.4 donde además mostramos las distancias mínimas entre cada par de impurezas de distinto “color” considerando condiciones de borde periódicas, lo cual significa que debemos tomar en cuenta la menor distancia entre esferas de esos colores que pueden haber en las celdas vecinas, es decir, la distancia mínima se puede formar con una posición equivalente en una celda vecina.

La búsqueda de la distancia mínima (**meandistance**) es equivalente construir una superestructura de $2 \times 2 \times 2$ y en esa superestructura se calculan todas las distancias entre las impurezas de distinto “color” y se eligen las menores de ellas.

Es importante mencionar que el criterio es distinto a si sólo buscamos las `n_impurezas` menores distancias dentro de todas las distancias posibles entre las impurezas (lo cuál sería equivalente a no considerar el “color” asignado), ya que como vemos en la parte derecha de la Fig. 3.4, las distancias (magnitud de las líneas discontinuas en la figura) va a ser distintas que las distancias que explicamos anteriormente.

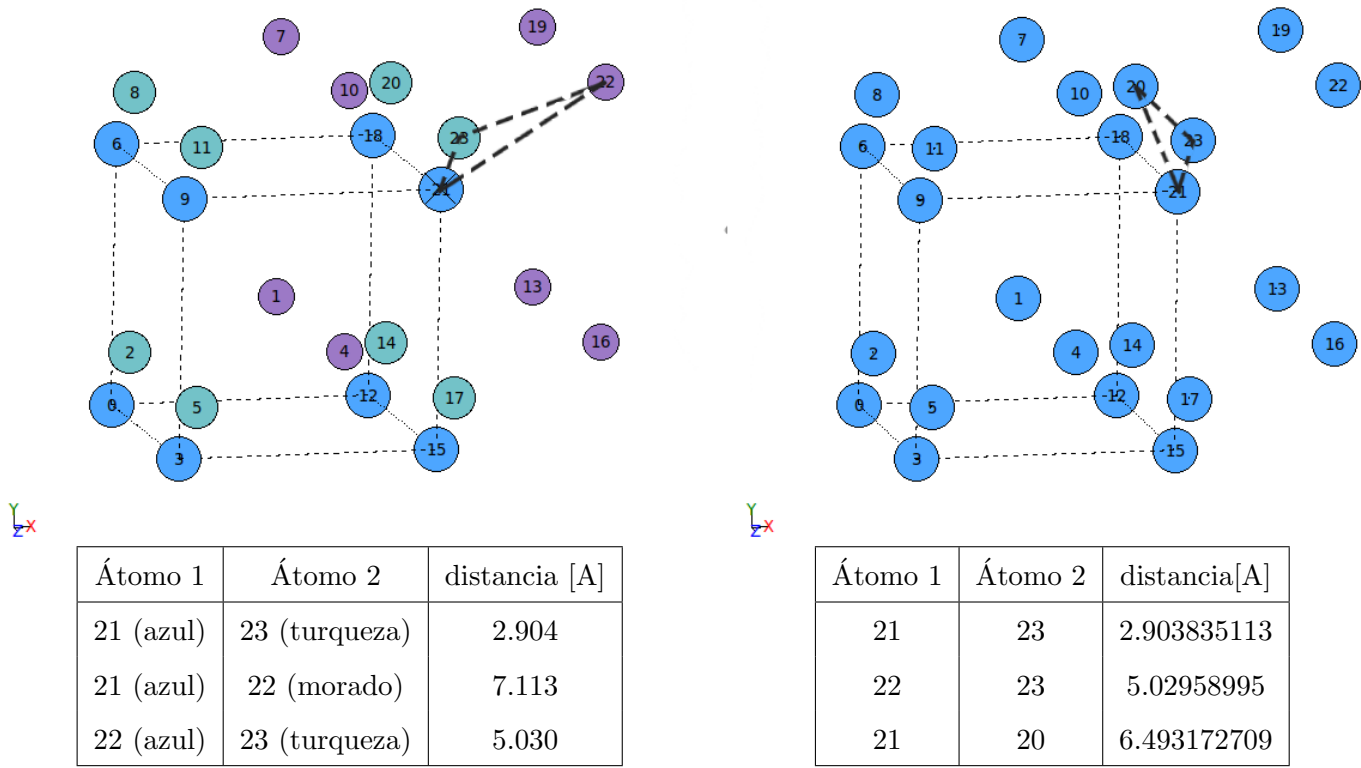


Figura 3.4: La imagen corresponde a una superestructura de tamaño $2 \times 2 \times 2$ de una configuración en particular con tres impurezas (c00010) y que tiene `cantmax=1`. El Átomo 21 es el que se muestra en cada figura azul en la parte superior a la derecha.

Volviendo a la Fig. 3.5, donde en las líneas discontinuas representan los vectores entre las impurezas, cuya magnitud es la distancia medida que se muestra en el Tabla 3.1 (cuando se escoge el átomo 21), se observa un caso en el que las condiciones de borde periódicas afectan a nuestro descriptos `meandistance`, ya que si sólo consideraríamos una celda las distancias deberían ser:

$$\left\| \vec{21} - \vec{22} \right\|, \quad \left\| \vec{21} - \vec{23} \right\| \quad \text{y} \quad \left\| \vec{22} - \vec{23} \right\|$$

pero en este caso se consideran las distancias

$$\|\vec{21} - \vec{22}\|, \quad \|\vec{21} - \vec{20}\| \quad \text{y} \quad \|\vec{22} - \vec{20}\|$$

donde la impureza 20 corresponde a una celda vecina.

Notar que al considerar la superestructura de tamaño $2 \times 2 \times 2$ las etiquetas son $\{0, 1, \dots, 255\}$ debido a que son $32 \cdot 2^3$ átomos. Al mostrar sólo las impurezas, ocupamos las etiquetas $\text{label} \in \{0, 1, \dots, 23\}$. Para volver a la etiqueta original debemos sumarle $32 \cdot \lfloor \frac{\text{label}}{3} \rfloor$

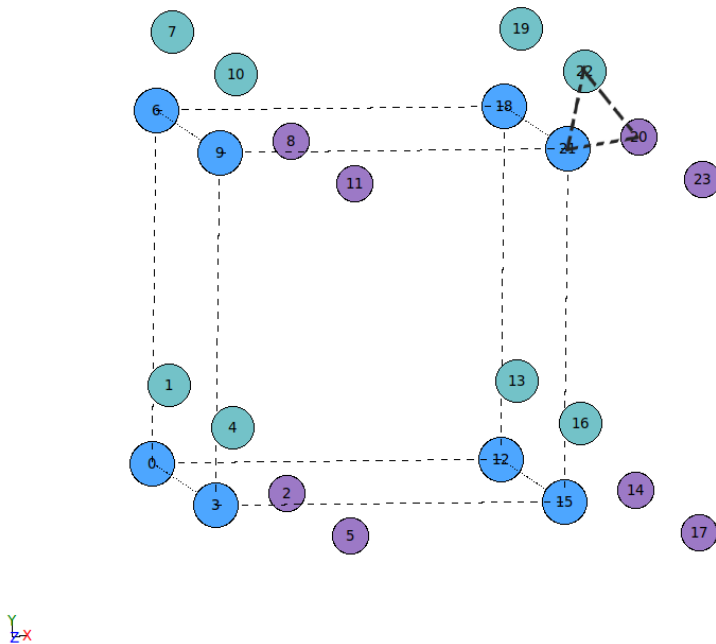


Figura 3.5: Configuración c00012 (tiene `cantmax=0`) para 3 impurezas y `repeat = (2,2,2)`.

Átomo 1	Átomo 2	distancia [Å]
21	22	2.904
21	20	2.904
22	20	5.030

Cuadro 3.1: Distancias medidas con `distances_matrix` (criterio de los colores)

3.3. Exploración

Al graficar `meandistance` vs `Efree` (ver Fig. 3.6) pudimos observar una correlación lineal para un determinado número de impurezas. Al hacer una regresión lineal para cada número de impurezas obtenemos el Tabla 3.2.

Efree-Efree_min vs meandistance para 4 impureza:

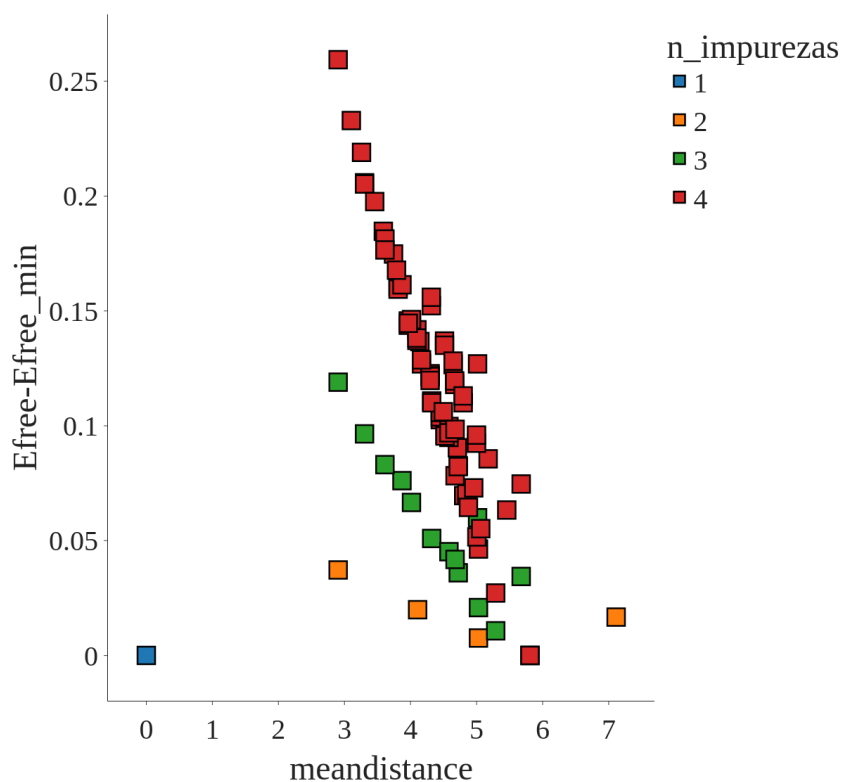


Figura 3.6: Efree vs meandistanceMicTrue para 1, 2, 3 y 4 impurezas

n.i	Intercepto	Pendiente	R2
2	0.045179	-0.005786	0.436
3	0.213231	-0.035721	0.869
4	0.451454	-0.075146	0.868

Cuadro 3.2: Tabla de interceptos, pendientes y R2 para 1, 2, 3 y 4 impurezas NO considerando `cantmax`
Fig. 3.6

Notamos que para cada número de impurezas se podían agrupar los puntos que tenían una correlación lineal mayor, lo cual se puede ver en al Fig. 3.7 donde incluimos nuevo descriptor `cantmax`, el cual es discutido en Subsec. 3.3.1.

Efree-Efree_min vs meandistance para 4 impurezas

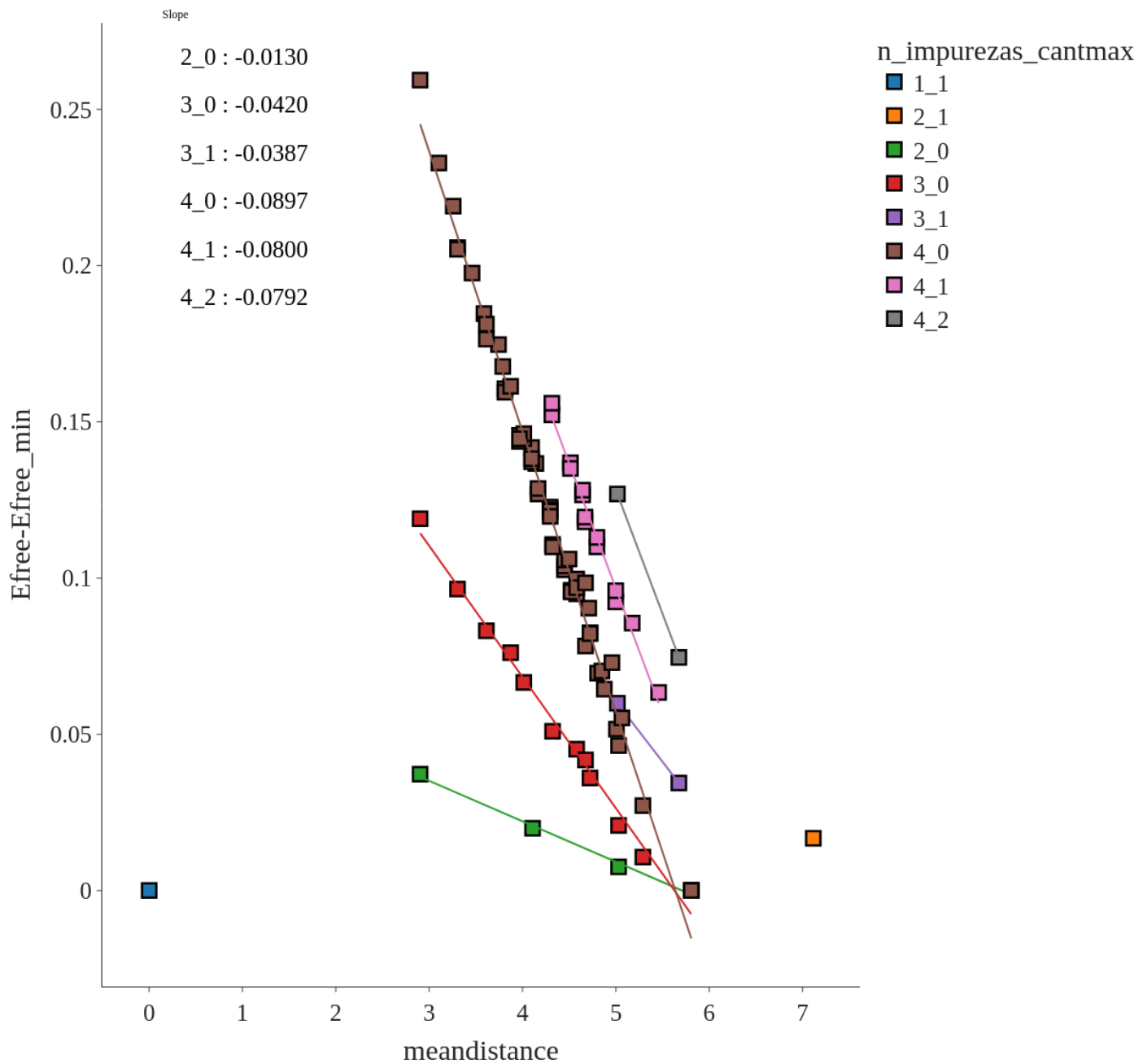


Figura 3.7: Efree-Emin vs meandistanceMicTrue para 1, 2, 3 y 4 impurezas considerando cantmax

Al considerar este descriptor, notamos que obtenemos un mejor ajuste en nuestras regresiones (ver [Tabla 3.3](#)), donde por ejemplo para 2 impurezas, el R2 aumenta de 0.436 a 0.994.

n_imp_cantmax	cantmax	interceptos	Pendiente	R2
2_0	0	0.074031	-0.012964	0.994281
2_1	1	-	-	-
3_0	0	0.236383	-0.042007	0.989730
3_1	1	0.254255	-0.038735	1.000000
4_0	0	0.505722	-0.089710	0.989273
4_1	1	0.496745	-0.080015	0.985975
4_2	2	0.524301	-0.079227	1.000000

Cuadro 3.3: Tabla de coeficientes, pendientes y R2 para 1, 2, 3 y 4 impurezas considerando `cantmax` y con OLS [Fig. 3.7](#)

La división de conjunto de puntos agrupados por formar una recta cuando sólo consideramos `meandistace` y `Efree`, se logra ocupando máquinas de soporte vectorial (SVM), donde separamos los puntos en 2 grupos al encontrar el hiperplano que maximiza la distancia entre los puntos de cada grupo, logrando así clasificarlos en 2 o más grupos dependiendo de la cantidad de hiperplanos que encontremos. El procedimiento de esta división, y por lo tanto construcción del descriptor `cantmax`, se explica en la [Subsec. 3.3.1](#).

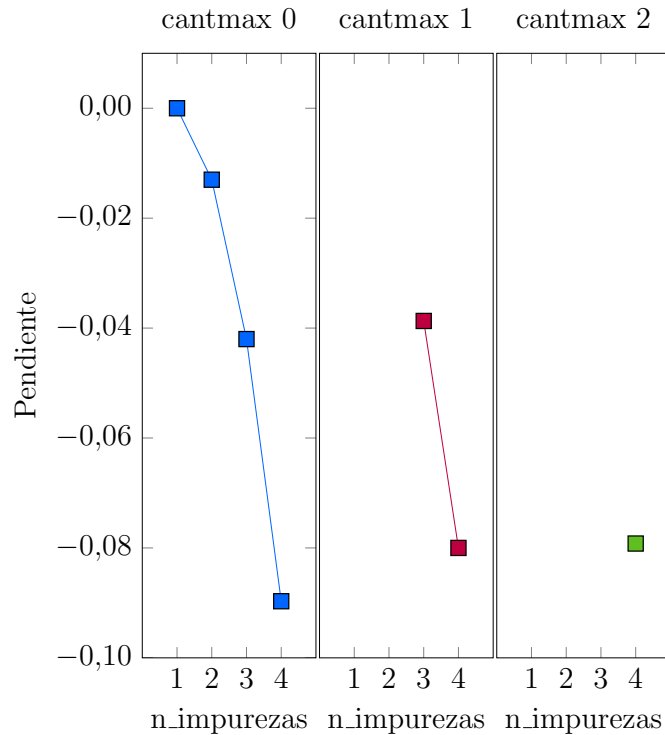


Figura 3.8: Pendiente vs `n_impurezas` para 1, 2, 3 y 4 impurezas

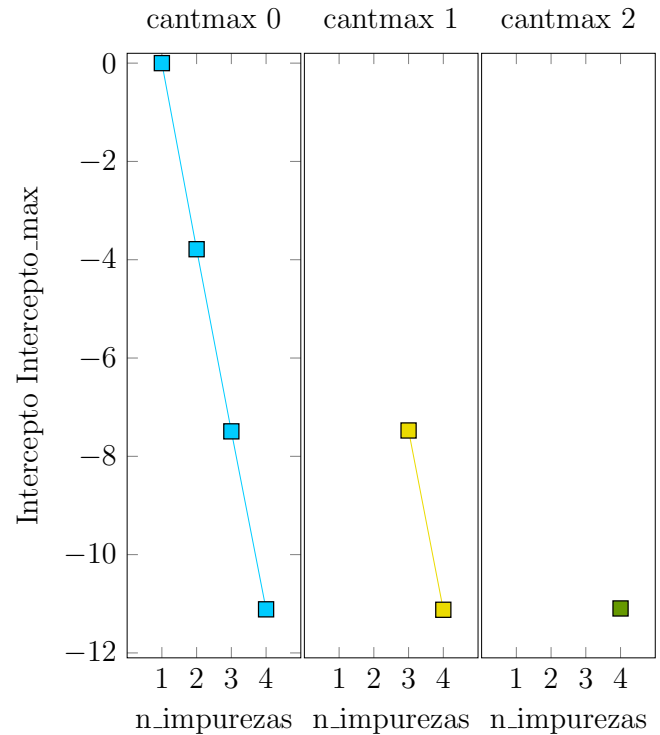


Figura 3.9: Intercepto ajustado a la mayor energía para 1, 2, 3 y 4 impurezas

3.3.1. El descriptor `cantmax` y SVM

El descriptor `cantmax` es igual a la frecuencia absoluta de la distancia máxima que puede haber entre dos impurezas en el conjunto de distancias posibles entre impurezas. La magnitud de la distancia máxima en este problema es igual a la mitad de la diagonal del cubo formado por los vectores de la celda y, en nuestro conjunto de distancias entre impurezas, no puede existir una distancia mayor a esta, debido a que consideramos condiciones de borde periódicas. A continuación, explicaremos el proceso de construcción de este descriptor.

Luego del análisis exploratorio, observamos en la [Fig. 3.10](#) que hay agrupaciones que siguen un comportamiento lineal en `meandistance`. Al intentar clasificarlas buscando alguna característica (o un conjunto de ellas) que pudiera clasificar el grupo de comportamiento lineal exitosamente para una configuración, sólo encontramos que había una correlación entre medidas de dispersión en el conjunto de distancias entre impurezas y el grupo de configuraciones de comportamiento lineal. En lo siguiente, explicaremos el proceso que seguimos, donde escogimos el caso de 4 impurezas.

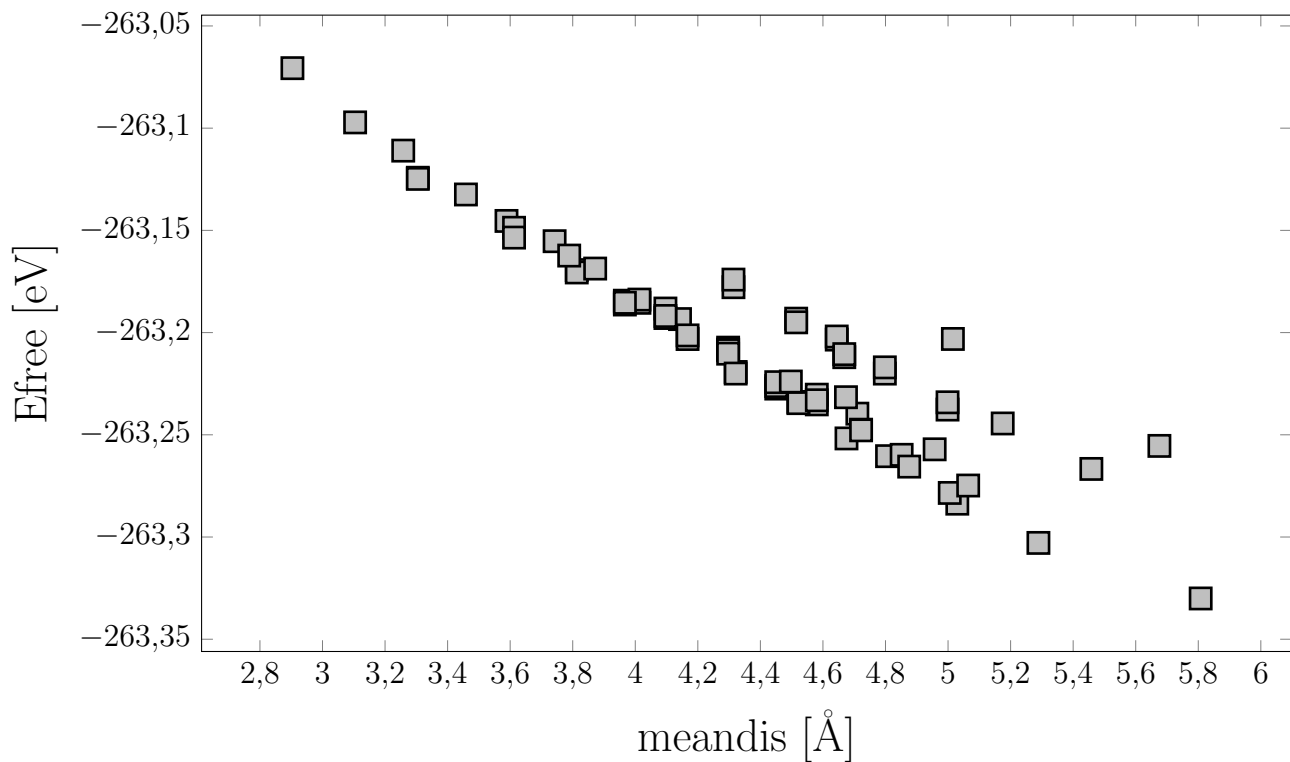


Figura 3.10: Energía vs meandis para 4 impurezas

De nuestros conocimientos de ML (AA), el primero enfoque fue ocupar técnicas de aprendizaje no supervisado que pudieran agrupar los datos, pero no tuvimos éxito debido a las características (en nuestro caso descriptores) que teníamos.

La siguiente estrategia fue ocupar un enfoque semi-supervisado, donde construimos un clasificador usando SVM con kernel lineal, donde básicamente buscamos un hiperplano que separe los puntos (los clasifique). Es importante mencionar que hablamos de un hiperplano y no de sólo una recta, ya que en el proceso de entrenar las máquinas de soporte vectorial (SVM) también consideramos otras características (descriptores) además de `meandistance`.

Lo que hicimos fue escoger un pequeño conjunto de datos y etiquetarlos (“pintarlos”) [Fig. 3.11](#) y dentro de ese conjunto aplicar las técnicas de ML para clasificar con éxito los puntos. Esto lo llevamos a cabo mediante el uso de SVM ([Apéndice A.3](#)) con autoetiquetado, lo cual consiste en ir construyendo el hiperplano clasificador de manera iterativa: en la primera iteración se ocupan los datos etiquetados manualmente obteniendo la recta [Fig. 3.12](#),

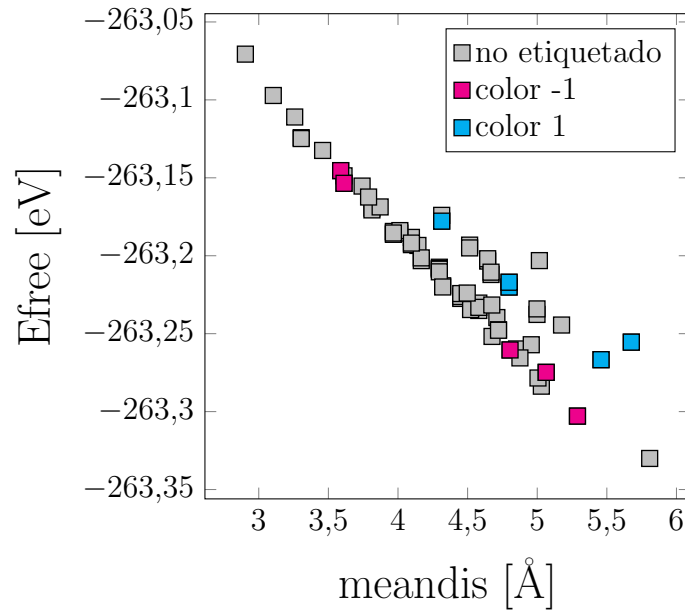


Figura 3.11: Energía vs meandis para 4 impurezas, data etiquetada con frontera de decisión

y luego de la construcción de este primer clasificador, el algoritmo ocupa este mismo clasificador para etiquetar más puntos que hace a través la frontera de decisión. Un ejemplo de este primer clasificador lo podemos ver en representado como una recta en la [Fig. 3.12](#). En la siguiente iteración construirá un nuevo clasificador (en nuestro ejemplo 'actualizará' la recta de frontera) de acuerdo al nuevo conjunto y así sucesivamente hasta que se clasifiquen todos los puntos y sea mínima la métrica de error considerada. En la [Fig. 3.13](#) podemos ver el resultado de la clasificación.

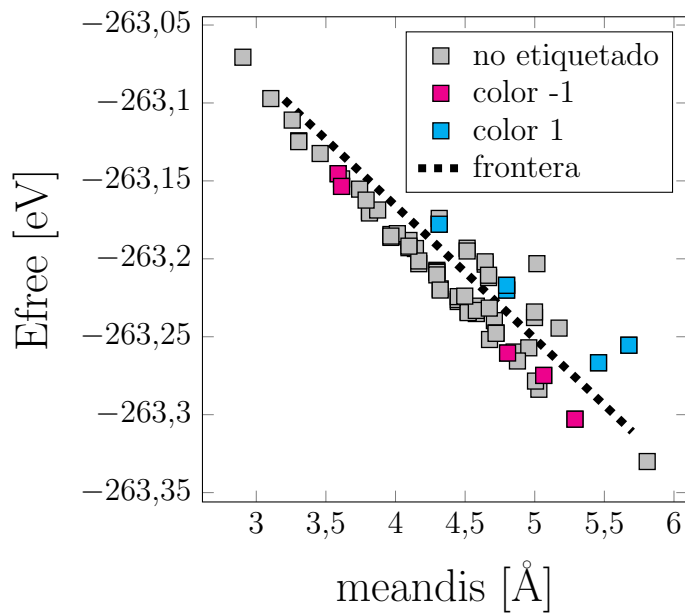


Figura 3.12: Energía vs meandis para 4 impurezas, data etiquetada con frontera de decisión

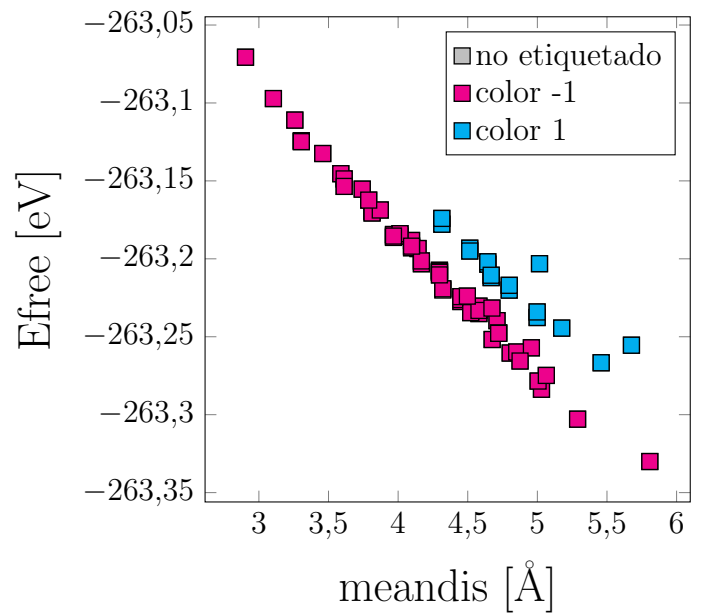


Figura 3.13: Energía vs meandis para 4 impurezas, data clasificada

Al ya poder agregar una etiqueta que “pinte” todos los datos del conjunto, empezamos a analizar manualmente algunas configuraciones mediante el uso de ASE y VESTA. Como anteriormente mencionamos, al tener información sobre la influencia de la dispersión en el conjunto de distancias entre impurezas y el ‘color’, agregamos más características a nuestras configuraciones, siendo una de ellas la frecuencia absoluta de la distancia máxima que pueden tener 2 impurezas (cantmax) de distinto ‘color’ en el cristal cuando consideramos condiciones de borde periódicas.

Incluyendo el nuevo descriptor, notamos que cantmax clasifica con un 100 % de éxito (de acuerdo a los “grupos” que nosotros observamos). Si ahora graficamos (Fig. 3.14)

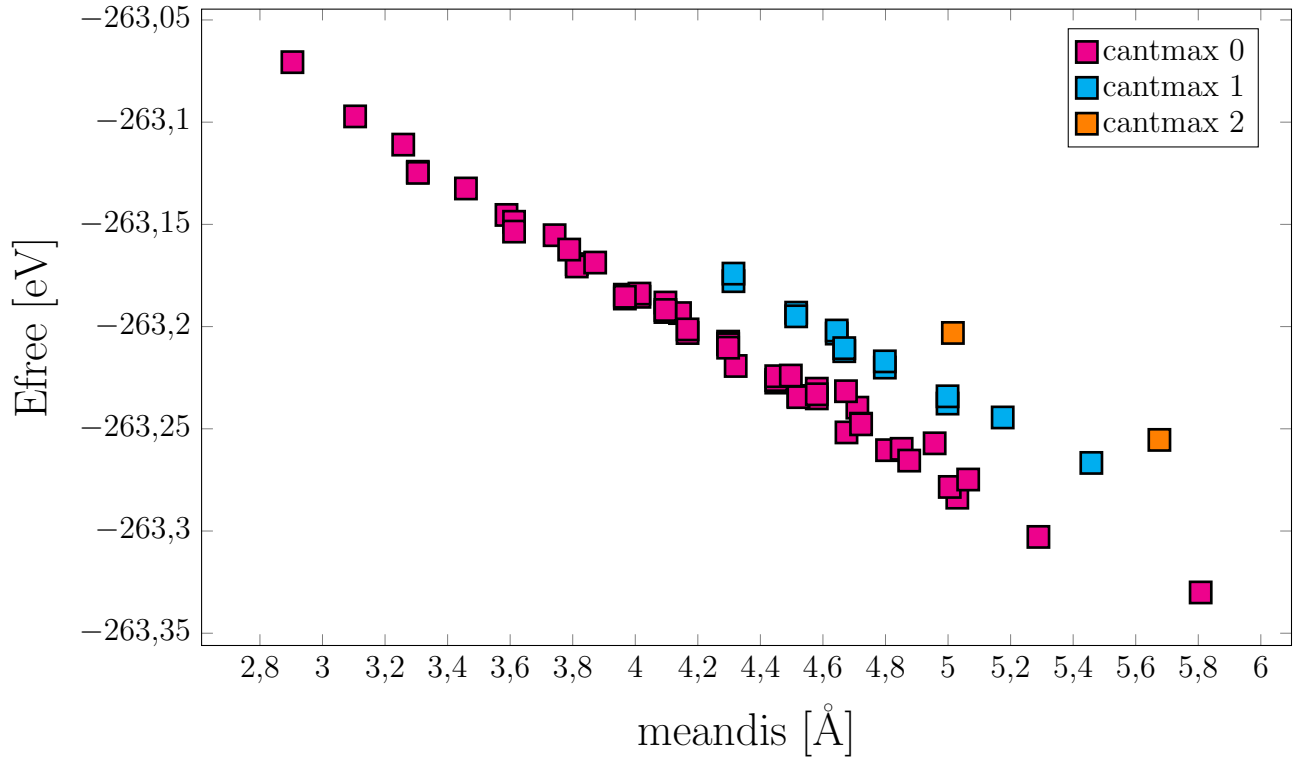


Figura 3.14: Energía vs meandis para 4 impurezas, color por `cantmax`

En Fig. 3.15 podemos ver con más detalle un ejemplo del agrupamiento hablado en la Sec. 3.3.

La configuración mostrada en la Fig. 3.4 es una configuración con valor 1 para el descriptor `cantmax`. Esto se debe a que 1 de las 3 distancias con las que se calcula `meandistance` es igual a la distancia máxima que puede haber entre dos impurezas de acuerdo a nuestro criterio. En este caso donde los vectores de la celda son

$$\begin{aligned}\vec{a} &= \begin{bmatrix} 8,213286 & 0,000000 & 0,000000 \end{bmatrix} [\text{\AA}] \\ \vec{b} &= \begin{bmatrix} 0,000000 & 8,213286 & 0,000000 \end{bmatrix} [\text{\AA}] \\ \vec{c} &= \begin{bmatrix} 0,000000 & 0,000000 & 8,213286 \end{bmatrix} [\text{\AA}]\end{aligned}$$

la distancia es máxima es:

$$\begin{aligned}
 d_{max} &= \frac{\|\vec{a}\|\sqrt{3}}{2} \\
 &= \frac{8,213286[\text{\AA}]\sqrt{3}}{2} \\
 &= 7,113[\text{\AA}]
 \end{aligned} \tag{3.1}$$

Entonces, la distancia entre las impurezas azul y morado en Fig. 3.15, donde mostramos sólo las impurezas (Átomos de Ta coloreados de acuerdo a su equivalencia en la estructura de $1x1x1$) de una superestructura de $2 \times 2 \times 2$ para una configuración en particular, es la máxima posible.

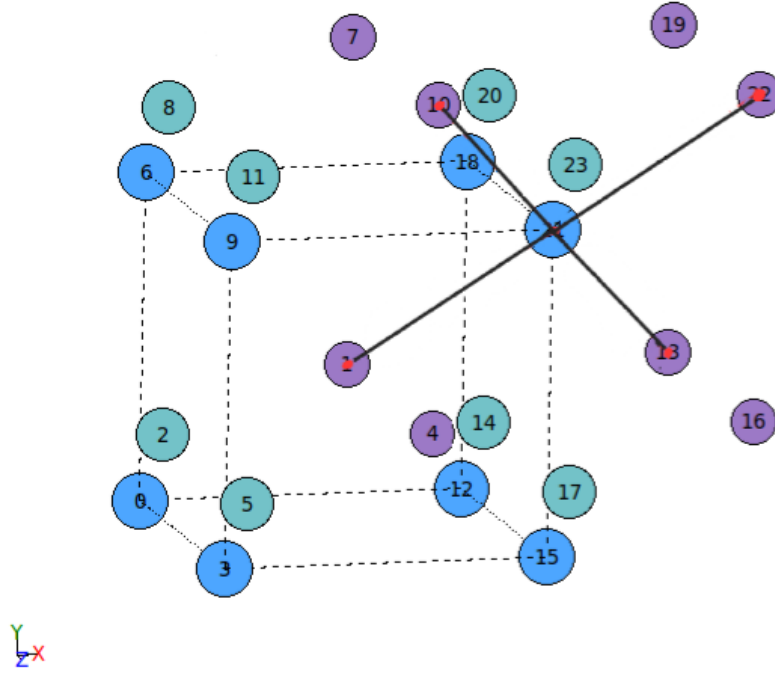


Figura 3.15: Superestructura de $2 \times 2 \times 2$ de la configuración c00010 (la misma de la Fig. 3.4). Podemos ver que las 4 distancias (azul-morado) miden $7.113\text{\AA}$.

3.4. El Modelo lineal: $E_{0_{n_{\text{impurezas}}}} \propto \text{meandistance}$. Hipótesis y evaluación.

Luego de el análisis exploratorio y las observaciones comentadas sobre la correlación entre el descriptor `meandistance` y la energía para cada par `n_impurezas` - `cantmax`, nuestra hipótesis fue de que

la energía es lineal en `meandistance` y la pendiente e intercepto que definen ese modelo lineal dependen del número de impurezas de manera cuadrática.

Luego de plantear nuestra hipótesis, lo siguiente fue construir un modelo lineal para cada par `n_impurezas` - `cantmax` siguiendo los principios del Machine Learning (ML). De igual manera, se construirán otros modelos para comparar los resultados obtenidos.

A continuación, sólo detallaremos el procedimiento para el caso de 4 impurezas, pero se realizó de igual manera para 1, 2 y 3 impurezas.

3.4.1. División de entrenamiento y de prueba

Necesitamos dividir nuestros datos en entrenamiento y prueba para poder evaluar el desempeño de nuestro modelo. Estratificamos por el descriptor `cantmax` para evitar que en el conjunto de entrenamiento y/o prueba se concentren configuraciones con un valor de `cantmax` en particular, ya que como se ve en [Fig. 3.16](#) hay una gran diferencia entre la cantidad de configuraciones con `cantmax` = 0 y `cantmax` = 1. Debido a la poca cantidad de configuraciones con `cantmax` = 2 para este número de impurezas, no las consideramos en nuestras pruebas.

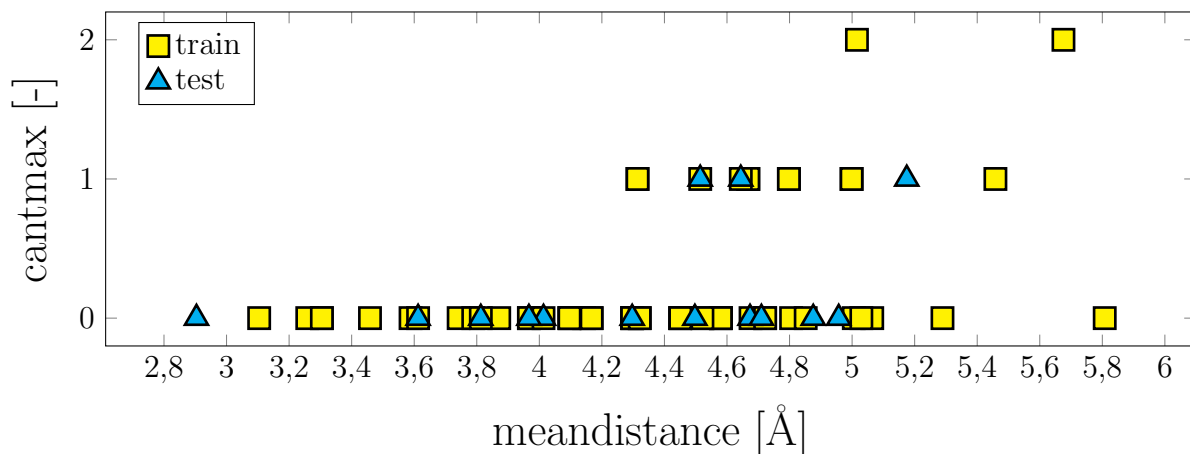


Figura 3.16: `cantmax` vs `meandistance` para 4 impurezas color para train y test considerando el 80 % de los datos para entrenamiento.

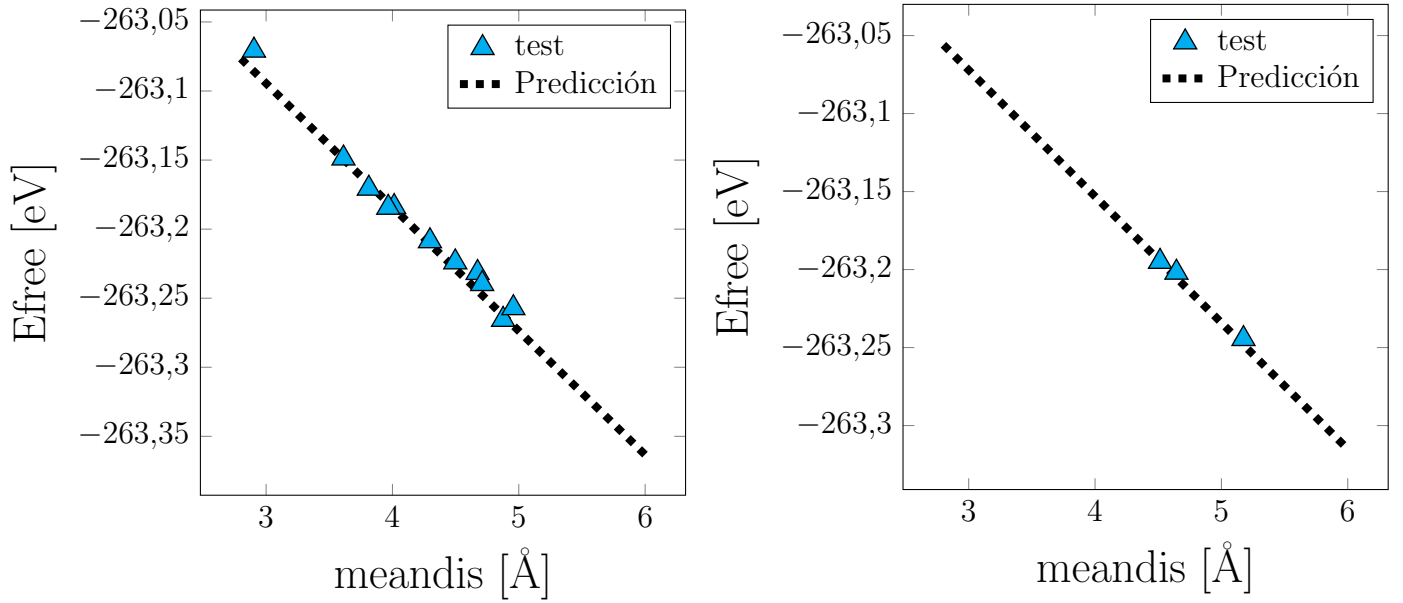


Figura 3.17: resultados de la predicción (y_{pred}) versus los target de la prueba (y_{test}) donde ponemos notar que el modelo predice con un no mayor a 10^{-2} [eV] los valores de energía para configuraciones con `cantmax` = 0 (figura en el lado izquierdo) y `cantmax` = 1 (figura en el lado derecho).

En la Fig. 3.16 se observa que el modelo predice con un error menor a 10^{-2} [eV] los valores de energía para configuraciones con `cantmax` = 0 y 1.

El test se realizó de igual manera para 1, 2, 3 y 4 impurezas, con distintos porcentajes de training set y test set, confirmando además que el tamaño de entrenamiento (dentro de los límites que permite la estratificación) no afecta el desempeño de nuestro modelo y se confirma la hipótesis de linealidad en `meandistance`.

3.4.2. Construcción de modelos para números de impurezas desconocidos

Luego de haber confirmado la linealidad en `meandistance` de la energía cuya precisión se vió aumentada considerablemente al encontrar, con nuestros análisis de ML, el nuevo descriptor `cantmax`, buscamos un modelo que pueda predecir la energía que tendrá una configuración para números de impurezas para los cuáles no tenemos mediciones. Para eso, buscamos un modelo que pueda predecir cuál es la pendiente e intercepto que tendrá la función lineal en `meandistance` que predice la energía (`Efree`).

3.4.3. Modelo para la pendiente y el intercepto

Una observación importante hecha en [Sec. 3.3](#) es la relación que hay entre el número de impurezas y las pendientes e interceptos de las regresiones. Al observar [Fig. 3.8](#) y [Fig. 3.9](#) propusimos modelos que se ajusten a lo observado:

Regresión polinomial de grado 1

$$\beta_0 + \beta_1 \cdot \text{n_impurezas} \quad (3.2)$$

Regresión polinomial de grado 2

$$\beta_0 + \beta_1 \cdot \text{n_impurezas} + \beta_2 \cdot \text{n_impurezas}^2 \quad (3.3)$$

Polinomio interpolador de *Lagrange*

$$P(x) = \sum_{i=0}^n y_i L_i(x) \quad (3.4)$$

donde en la [eq. \(3.4\)](#) $L_i(x)$ es el i -ésimo polinomio base de *Lagrange*, definido como $L_i(x) = \prod_{j=0, j \neq i}^n \frac{x-x_j}{x_i-x_j}$

Regresión ley de potencias

$$\beta_0 + \beta_1 \cdot \text{n_impurezas}^{\beta_2} \quad (3.5)$$

construimos modelos que se ajusten a las observaciones en [Subsec. 3.4.2](#). Es importante notar que en el caso del *Polinomio Interpolador de Lagrange* estamos contruyendo un polinomio que pasa por todos los puntos, por lo que no es un modelo que se pueda generalizar a otros datos. Sin embargo, lo incluimos para poder comparar los resultados con los otros modelos.

3.4.4. Evaluación de modelos

De los modelos contruidos con los datos de 1, 2, 3 y 4 impurezas, los 3 de mejor desempeño fueron:

Model 1

$$\beta_0 + \beta_1 \cdot \text{n_impurezas}^{\beta_2}$$

Model 2

$$P_3(\text{n_impurezas}) = \sum_{i=1}^4 y_i L_i(\text{n_impurezas})$$

Model 3

$$\beta_0 + \beta_1 \cdot \text{n_impurezas} + \beta_2 \cdot \text{n_impurezas}^2$$

donde el desempeño es el ajuste de sus predicciones con los datos de 5, 6 y 7 impurezas (donde para 5 y 6 impurezas fueron enviadas a calcular la energía con nuestras predicciones y para 7 impurezas se enviaron a calcular configuraciones aleatorias). Estos resultados de muestran en [Tabla 3.4](#).

Efree	cantmax	Efree-Efree_model_1	Efree-Efree_model_2	Efree-Efree_model_3	n_impureza
-267.150822	1	0.081252	0.066841	0.042329	5
-267.145714	0	0.017992	0.004162	-0.018420	5
-267.121023	0	0.017782	0.004163	-0.017716	5
-267.115114	0	0.021374	0.007774	-0.014039	5
-267.112360	0	0.012853	-0.000651	-0.022146	5
Efree	cantmax	Efree-Efree_model_1	Efree-Efree_model_2	Efree-Efree_model_3	n_impureza
-271.002543	0	0.079083	0.035136	-0.040698	6
-270.991335	0	0.077040	0.033303	-0.041734	6
-270.987331	0	0.070258	0.026692	-0.047697	6
-270.986114	2	0.188047	0.142629	0.061235	6
-270.965643	3	0.230748	0.184978	0.102247	6
Efree	cantmax	Efree-Efree_model_1	Efree-Efree_model_2	Efree-Efree_model_3	n_impureza
-274.832839	0	0.354387	0.256816	0.071361	7
-274.778394	0	0.285788	0.190854	0.016843	7
-274.708790	0	0.278724	0.185433	0.018553	7

Cuadro 3.4: Diferencia de energía calculada y predicción (en las columnas la diferencia para cada modelo) para 5, 6 y 7 impurezas (orden descendente). Notamos que la diferencia de energía es menor para la configuración con cantmax = 1 y nuestros modelos están ajustados para cantmax = 0. Debido a lo anterior es que explicamos cuál será nuestro criterio de selección dado un conjunto de configuraciones sin su energía calculada (sólo con predicción)

Capítulo 4

Resultados

En este capítulo se presentarán los principales resultados obtenidos en este trabajo y se discutirán sus consecuencias.

4.1. Confirmación del modelo lineal en `meandistance` para la Energía (E_0) dado un número de impurezas

Con los test realizados primero se confirmó el comportamiento lineal en los conjuntos donde conocemos la energía para todas las configuraciones (1, 2, 3 y 4 impurezas) y luego se demostró que con una cantidad pequeña (2 - 3) de pares `config-Efree` podemos predecir la energía de configuraciones con 5 y 6 impurezas. Esto lo logramos ajustando el modelo lineal predicho y demostrando que se cumple para 2, 3 y 4 impurezas.

4.2. Un modelo para la Energía (E_0) construido con los datos reales de 1, 2, 3 y 4 impurezas

De los distintos modelos construidos en [Subsec. 3.4.4](#) con los datos reales de 1, 2, 3 y 4 impurezas, elegimos el que tiene menor error en la predicción de la energía para 5, 6 y 7 impurezas, cuyo error se muestra en [Tabla 3.4](#). Este modelo es el lineal en `meandistance` cuya pendiente e interceptos son funciones cuadráticas en el número de impurezas, donde la función fue obtenida mediante regresión polinomial de grado 2.

4.2.1. Modelo para la Energía (E_0) en términos de `meandistance` y `n_impurezas`

La función construída para predecir la energía E_0 en términos de `meandistance` y `n_impurezas` es

$$E_0 = s \cdot \text{meandistance} + i$$

donde

$$s = \underbrace{\beta_0}_{-0,00505} + \underbrace{\beta_1}_{0,0136} \cdot \text{n_impurezas} + \underbrace{\beta_2}_{-0,0087} \cdot \text{n_impurezas}^2 \quad \left[\frac{eV}{\text{\AA}} \right]$$

$$i = \underbrace{\beta_0}_{-247,841} + \underbrace{\beta_1}_{-3,9114} \cdot \text{n_impurezas} + \underbrace{\beta_2}_{0,0414} \cdot \text{n_impurezas}^2 \quad [eV]$$

que en una sólo ecuación es

$$E_0(\text{meandis}, \text{n_imp}) = (-0,00505 + 0,0136 \cdot \text{n_imp} - 0,0087 \cdot \text{n_imp}^2) \cdot \text{meandis} \quad (4.1)$$

$$+ (-247,841 - 3,9114 \cdot \text{n_imp} + 0,0414 \cdot \text{n_imp}^2) \quad (4.2)$$

la cual nos permite hacer una predicción de la energía cuando no tenemos dato alguno de la energía de configuraciones de mayor número de impurezas. Notamos que, considerando las dificultades discutidas en la introducción sobre hacer cálculos para configuraciones a medida que aumentamos el número de impurezas, este modelo nos permite hacer una predicción de la energía de configuraciones de mayor número de impurezas sin necesidad de hacer cálculos de energía para configuraciones de mayor número de impurezas.

4.3. Descriptor `cantmax` y su importancia en la selección de configuraciones candidatas

Con ayuda de las técnicas de *Machine learning*, encontramos un descriptor (`cantmax`) que explica la “degeneración” observada en el comportamiento de la energía vs `meandistance` para un número de

impurezas dado, permitiéndonos hacer una precisa predicción de las configuraciones candidatas a ser la de menor energía.

4.3.1. Criterio de selección de la configuración de mínima energía

Al confirmar que la energía es lineal en `meandistance` podríamos pensar que sólo necesitamos calcular `meandistance` de las configuraciones y simplemente escoger la de mayor `meandistance` como la configuración con menor energía, pero cómo mostramos en Fig. 4.1, donde la configuración de menor energía no es la de mayor `meandistance` por lo tanto ese criterio por sí sólo no es suficiente.

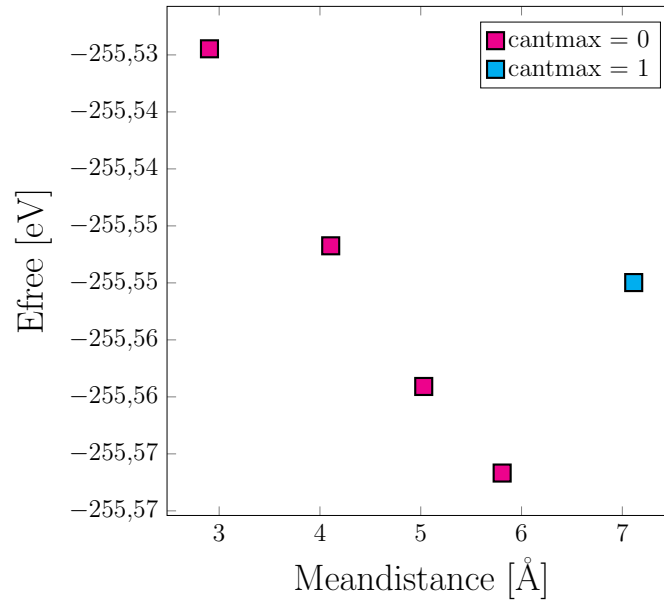


Figura 4.1: Energía en función de `meandistance` para 2 impurezas.

Al identificar esto, notamos que nuestro criterio para escoger el conjunto de configuraciones candidatas debe considerar el descriptor `cantmax`, ya que para un número de impurezas para el cuál no tenemos información sobre la energía el criterio de escoger la configuración de mayor `meandistances` o las x configuraciones de mayor `meandistance` (donde x se escoge bajo algún criterio, por ejemplo el 5 % de las configuraciones de mayor `meandistance`), podría fallar debido al comportamiento observado.

4.3.2. Sin muestras para un número de impurezas dado

Si no tenemos información experimental (obtenida por simulación) sobre la energía para un número dado de impurezas, con nuestros resultados igualmente podemos abordar el problema.

Al tener en cuenta **cantmax** y la pendiente predicha podemos hacer una elección más precisa sobre la configuración de menor energía. El criterio para la selección de la configuración de mínima energía está formado por los siguiente pasos:

Paso 1

Calcular **meandistance** ($\overline{\mathcal{D}_i}$), **cantmax** (f_i) para todas las configuraciones $c_i : c_i \in \mathcal{C}$ y almacenarlas en un conjunto $\mathcal{A}$

Paso 2

Ordenar el conjunto $\mathcal{A}$ (resultado del paso 1) de mayor **meandistance** a menor **meandistance**

Paso 3

Recorrer el $\mathcal{A}$ y agregar a_i $a_i \in \mathcal{A}$ a un nuevo conjunto $\mathcal{M}$ si y sólo si no se ha agregado una con su mismo **cantmax**

Paso 4

Si el conjunto $\mathcal{M}$ tiene un sólo elemento, entonces esa es la configuración de menor energía. Si tiene más de un elemento, entonces calculamos la energía de todo el conjunto $\mathcal{M}$, el cual será de tamaño a lo más igual que los valores posibles de **cantmax**

Con este criterio se disminuye notablemente la cantidad de configuraciones a las que se les debe calcular la energía por medio de *simuladores de dinámica molecular*, ya que pasamos de calcular la energía de todas las configuraciones a calcular a lo más la energía de $|\mathcal{V}|$ configuraciones, donde $\mathcal{V}$ es el conjunto de valores que puede tomar **cantmax**.

Algoritmos para la selección de configuración estable

En las siguientes figuras ([fig. 4.2](#), [fig. 4.3](#) y [fig. 4.5](#)) se muestran los diagramas de flujo de los algoritmos necesarios para implementar nuestro criterio, donde los bloques trapezoidales son los datos de entrada, los bloques rectangulares son procesos y los bloques de rombo son decisiones.

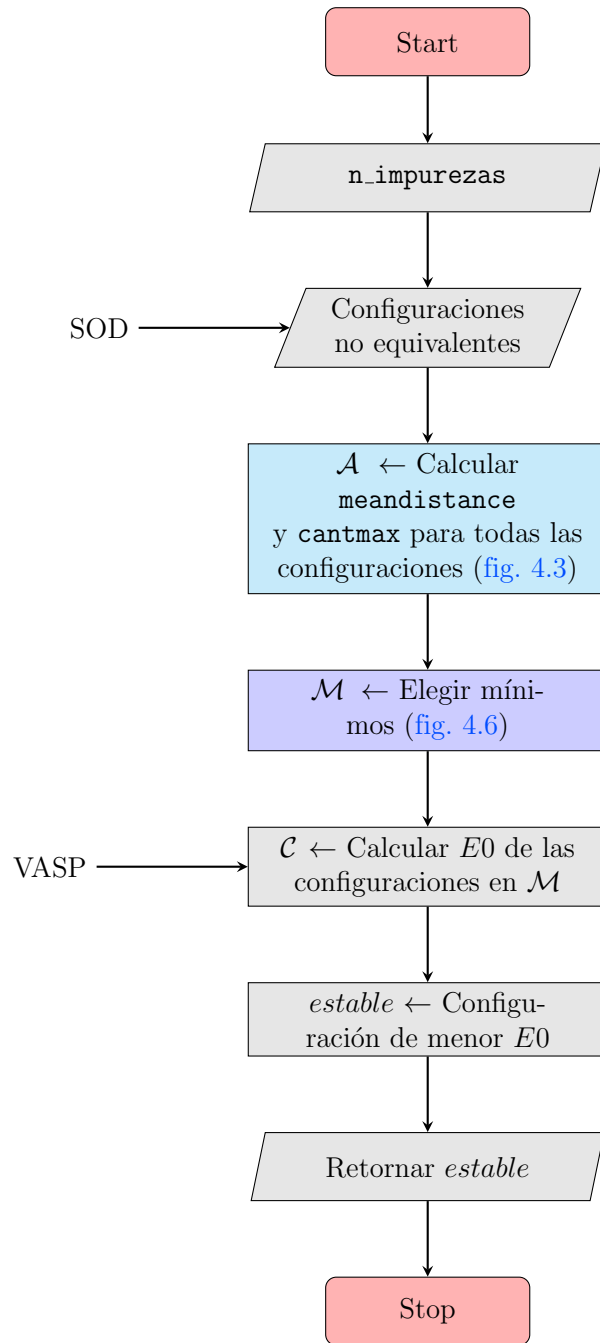


Figura 4.2: Diagrama de flujo del criterio encontrado. C es el conjunto cuya cardinalidad es igual a la cantidad de valores que puede tomar cantmax para un número de impurezas dado.

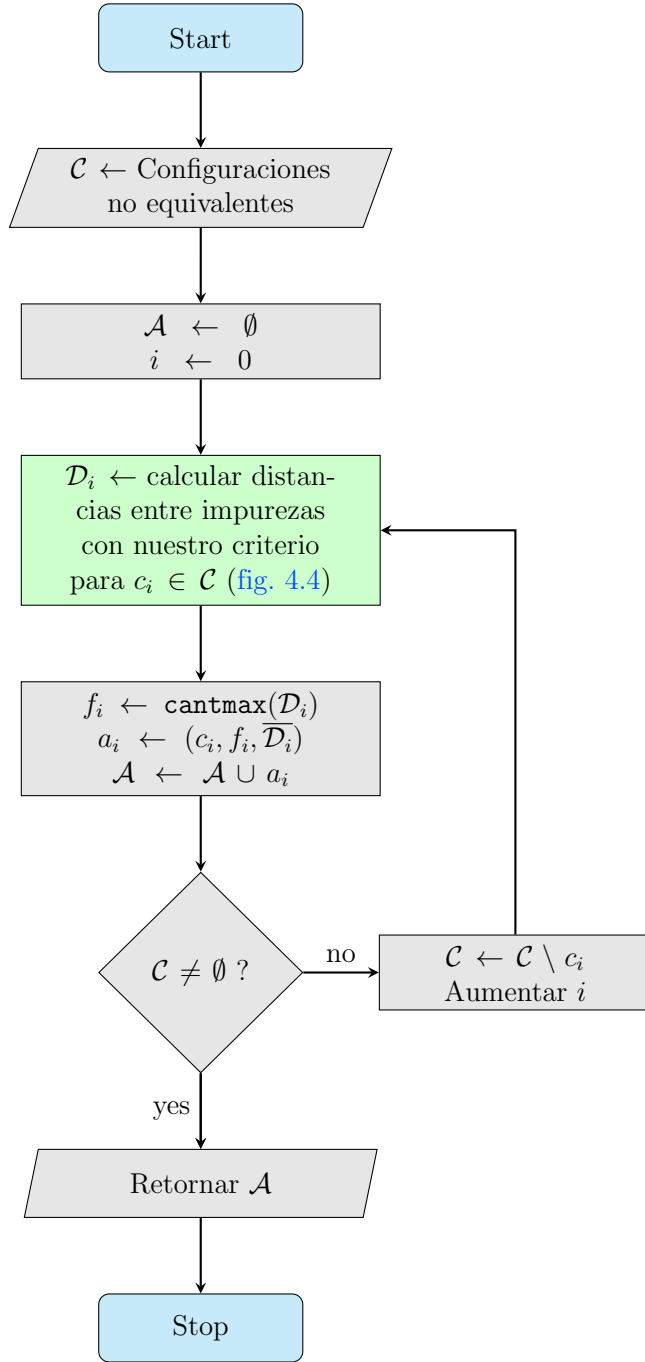


Figura 4.3: Diagrama de flujo cálculo descriptores.

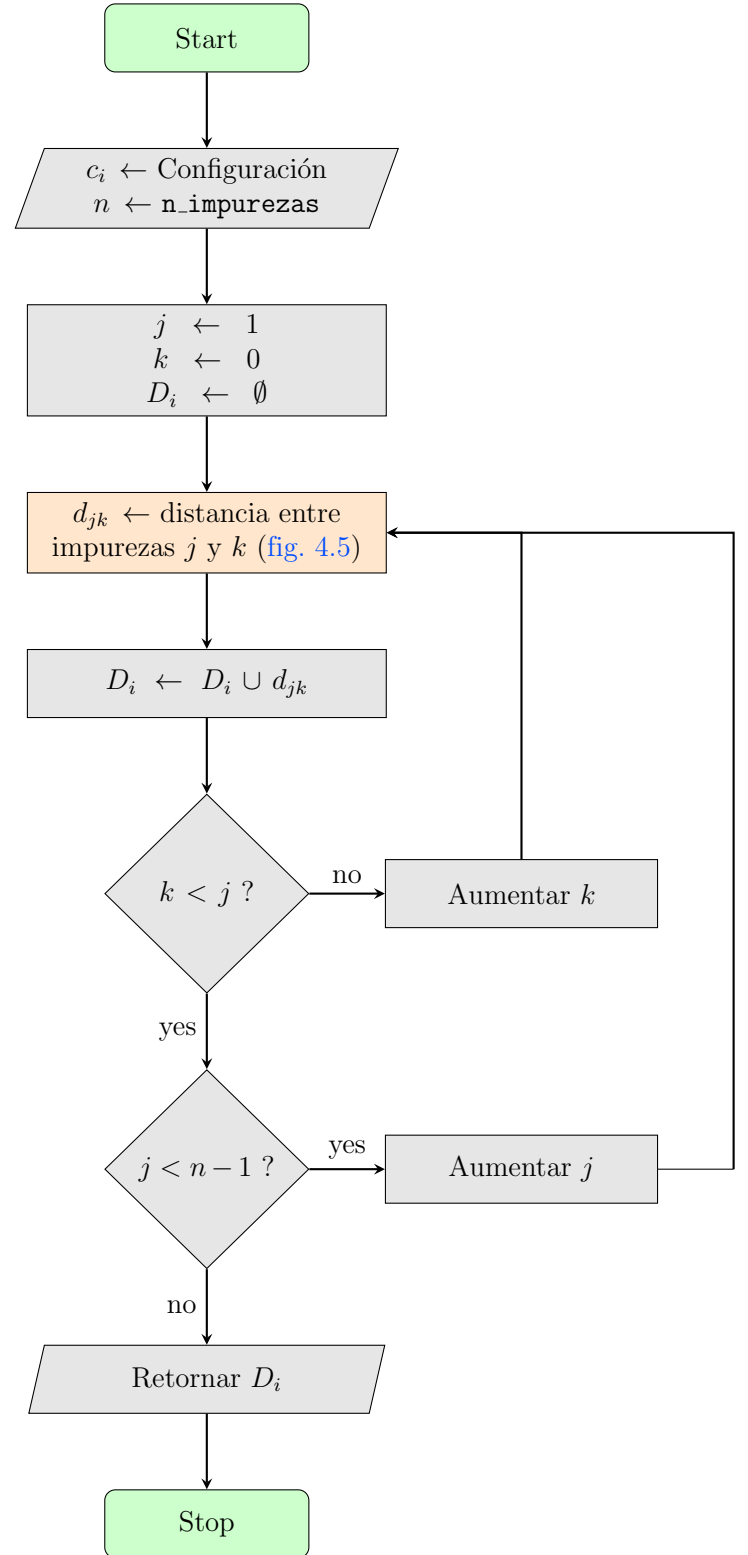


Figura 4.4: Diagrama de flujo cálculo descriptores. Se ha omitido el caso $n_impurezas_1$, que es trivial.

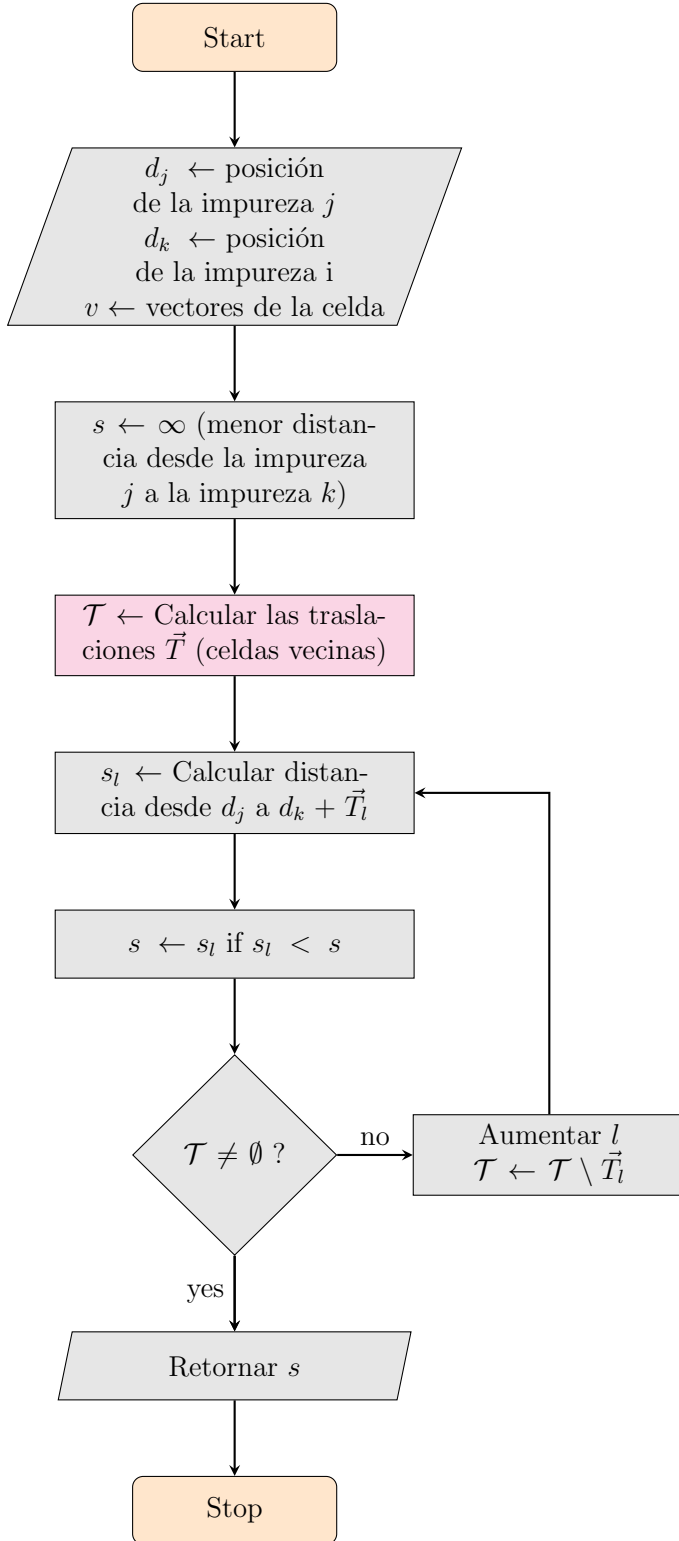


Figura 4.5: Diagrama de flujo cálculo de distancia entre impurezas: nuestro criterio

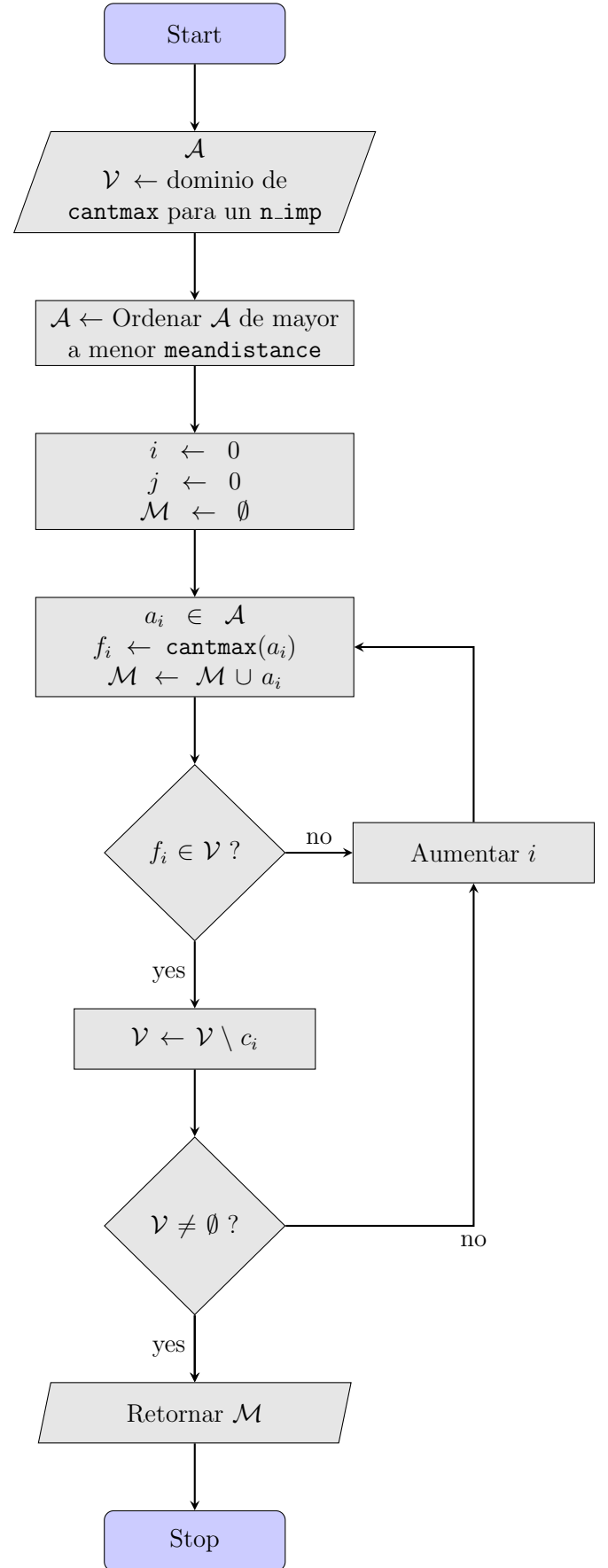


Figura 4.6: Diagrama de flujo para la selección de estructuras estables.

Es importante mencionar que en el subproceso **calcular traslaciones** (fig. 4.5), es el proceso donde se calculan todas las traslaciones posibles y se ha omitido su diagrama por motivo de presentar el algoritmo en un diagrama sin mayor complejidad.

4.3.3. Con muestras para un número de impurezas dado

Si tenemos información sobre la energía para un número de impurezas, con nuestros resultados podemos encontrar la configuración de menor energía fácilmente.

Mostramos que con sólo 2-3 pares **config-Efree** podemos obtener el modelo lineal para la energía en **meandistance** para ese número de impurezas, de lo cual obtenemos:

- Si esa cantidad, reducida de pares en comparación con la cantidad total de pares **config-Efree**, no los escogemos nosotros, los podemos usar para ajustar nuestro modelo descrito en el criterio anterior.
- Si esos pocos pares **config-Efree** los escogemos nosotros, entonces podemos elegirlos de configuraciones del conjunto $\mathcal{M}$ obtenido en el criterio anterior.

Una pequeña muestra es suficiente para ajustar el modelo. Esto se menciona, ya que podría ser un enfoque para tratar con otras impurezas que sepamos que se comportan de manera similar a las impurezas de Tántalo, y así ver si nuestro modelo se puede ajustar a ellas.

4.4. Simetría de las configuraciones y Energía

Observamos que si para un determinado número de impurezas hay dos configuraciones con el mismo **meandistance**, entonces la de menor **cantmax** (y la más simétrica entre las dos) es la de menor energía.

Idealmente lo que se busca es una configuración en la que la distancia entre pares de impurezas sea igual (tengamos una desviación estándar igual a 0) y la media sea lo más grande posible. Si para un número de impurezas mayor a 2 una configuración tiene $\text{cantmax} \geq 0$, entonces las distancias entre pares de impurezas ya no podrán ser iguales debido a que las posiciones en las que pueden estar las impurezas no lo permiten.

4.4.1. Extremos de Energía

En el [fig. 4.7](#), $\Delta E_x(21)$ es la energía diferencia de energía entre 2 y 1 impureza de los valores extremos x (máximo o mínimo).

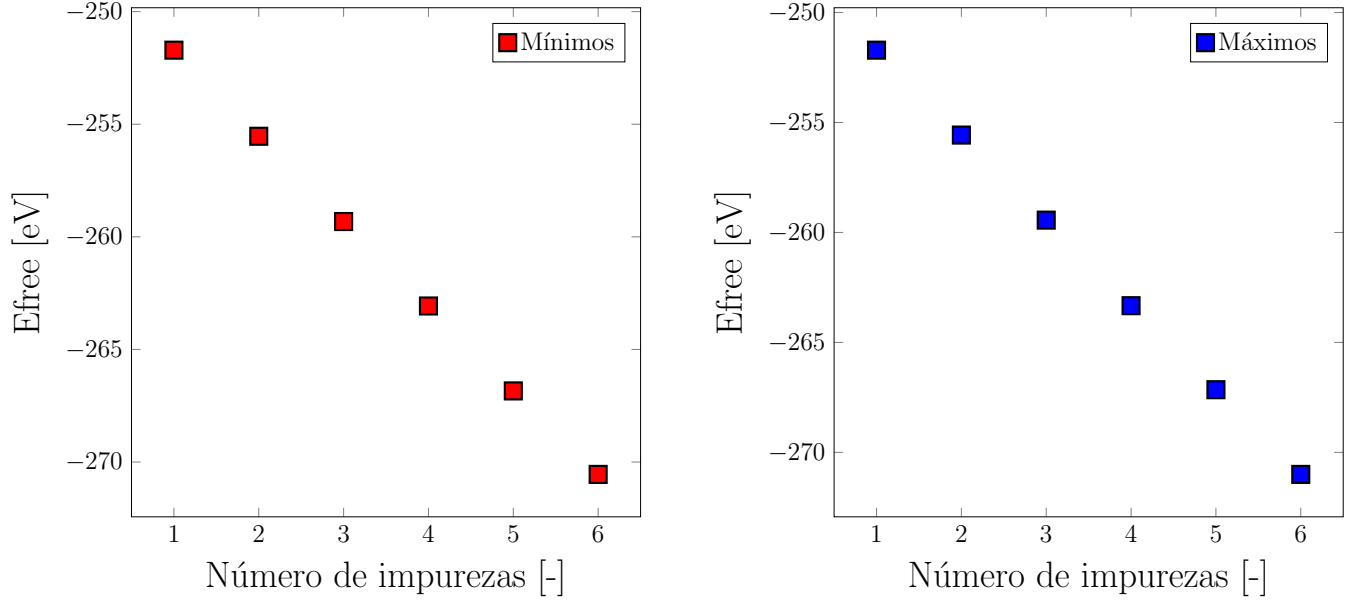


Figura 4.7: Valores externos. Detalles en [tabla 4.1](#)

	$\Delta E_x(2, 1)$	$\Delta E_x(3, 2)$	$\Delta E_x(4, 3)$	$\Delta E_x(5, 4)$	$\Delta E_x(6, 5)$
$x = \min$	-3,86049115	-3,86807182	-3,89030108	-3,8207749	-3,85172054
$x = \max$	-3,82327194	-3,78629938	-3,74988729	-3,77115805	-3,70754377

Cuadro 4.1: diferencias de valores extremos de energía

Podemos notar que la diferencia de energía mínima a medida que incluimos una impureza más (aumentamos la concentración de impurezas) es aproximadamente constante, observándose lo mismo para las energías máximas.

La energía del sistema escala con el número de átomos, por lo que si reemplazamos un átomo de Titanio por uno de Tántalo la diferencia observada en la energía es aproximadamente constante. Debemos notar que bajo la suposición de que los átomos no interactúan entre sí, la diferencia de energía al hacer el reemplazo debe ser constante.

Además, como discutimos en el capítulo anterior, la pendiente de la energía en `meandistance` disminuye (pero aumenta en magnitud) a medida que aumentamos el número de impurezas, por lo tanto

la diferencia de `meandistance` entre configuraciones de mayor número de impurezas es menor.

Capítulo 5

Conclusiones

En este trabajo logramos presentar un exitoso ejemplo del uso de técnicas de aprendizaje automático para resolver problemas de la ciencia de materiales. En particular se mostró que el uso de estas técnicas nos permitió la construcción de un criterio (o algoritmo) que tiene como consecuencia una notable disminución del uso de recursos computacionales para la determinación de las configuraciones más estables luego de una sustitución en una red cristalina.

Nuestro criterio (o algoritmo) para la determinación de las configuraciones más estables, al considerar sólo el cálculo de los descriptores `meandistance` y `cantmax` de cada configuración (los cuales dependen únicamente de la geometría), tiene como consecuencia calcular las energías mediante simulación de sólo unas pocas configuraciones y no la todas las configuraciones no equivalentes.

El criterio encontrado permite reducir la cantidad de simulaciones de dinámica molecular abruptamente, ya que el costo en tiempo real del calculo de la energía de una sólo configuración por primeros principios es varios órdenes de magnitud mayor al tiempo real requerido en generar esa configuración [Tabla 5.1](#) (considerando el mismo hardware). Así, nuestro enfoque se presenta como **alternativa viable para el estudio de las configuraciones estables** de esta aleación.

Estequiometrías	Conf. totales	Conf. irreducibles	<i>disorder</i>	<i>supercell</i>	<i>enumlib</i>	<i>SOD</i>
1:31	32	1	0.0218	13.354	0.0745	0.0681
2:30	496	5	0.0207	13.542	0.1517	0.0699
3:29	4960	14	0.0226	13.746	0.3102	0.0879
4:28	35960	71	0.0385	13.565	1.0856	21.173
5:27	201376	223	0.0812	13.721	3.2949	451.29
6:26	906192	874	0.2459	14.304	12.447	10248
7:25	3365856	2706	0.7622	16.553	39.866	≥ 2 días
8:24	10518300	8043	2.2275	23.266	119.01	/
9:23	28048800	20123	6.9478	36.764	300.72	/
10:22	64512240	45497	18.398	65.990	701.43	/
11:21	129024480	88716	41.692	116.29	1377.3	/
12:20	225792840	154379	81.531	191.36	2333.6	/
13:19	34737600	234803	137.11	286.38	3559.0	Crashed
14:18	471435600	318348	209.04	379.59	4857.5	Crashed
15:17	565722720	379926	283.44	454.09	6836.6	Crashed
16:16	601080390	404582	306.24	4743.10	6123.4	Crashed
Acumulativo	2240823842	1658311	1087.7	2125.7	25264	/

Cuadro 5.1: Los resultados del cálculo de las configuraciones por medio de los softwares *disorder*, *supercell*, *enumlib* y *SOD* para todas las estequiometrías distintas combinatoriamente en una red fcc de $2 \times 2 \times 2$. El tiempo de ejecución reportado de los cuatro códigos es en tiempo real en un procesador Intel[®] Xeon[®] E5-2678, y solo se utiliza un hilo durante la prueba. [13]

5.1. Trabajos futuros

Con el criterio encontrado y con la identificación de la importancia de poder encontrar la configuración de menor energía para un número de impurezas, para el cual no tenemos datos de energía, se abren opciones a alternativas al software SOD (u a otros que generen las estructuras no equivalentes como los de la tabla [Tabla 5.1](#)) que puedan hacer el cálculo de los dos descriptores mencionados anteriormente y retornar sólo las configuraciones de menor energía para un número de impurezas dado.

Independiente de las optimizaciones que pueda tener SOD o algún software con el que generemos todas las configuraciones no equivalentes para un número de impurezas dado, justamente estamos “armando” todas las configuraciones.

Con el criterio encontrado en este trabajo, se pueden proponer los siguientes enfoques:

Algoritmos de búsqueda con heurísticas

Un problema de búsqueda del óptimo global, que en nuestro caso es la búsqueda de la configuración que minimiza la energía. Al tener una función que predice la energía podemos usar algoritmos de búsqueda con heurísticas y ocupar como función objetivo nuestro modelo (eq. (4.1)). En este caso es muy útil el predecir el valor aproximado que tendrán las de máxima energía, ya que así podemos acotar nuestra búsqueda.

Incluir el criterio en la generación de estructura

Podría incluirse el criterio antes mencionado en algún software que ya exista y que sea de código abierto, como por ejemplo el software que se usó en este trabajo para la generación de configuraciones (SOD). De esta forma, al momento de generar las configuraciones, se podría ir calculando `meandistance` y `cantmax` y sólo retornar las configuraciones de menor energía para un número de impurezas dado.

Acotar el espacio de búsqueda

Este enfoque implica identificar cuáles son los movimientos, que en el caso de SOD son las operaciones sobre una estructura, que disminuyen `meandistance` y evitarlos.

5.2. Limitaciones de los resultados

Nuestro modelo sólo ha sido validado para las configuraciones de impurezas de Tántalo y para el rango de impurezas que hemos considerado. No podemos asegurar que nuestro modelo sea válido para configuraciones de impurezas de otros materiales, ni para porcentajes de impurezas considerablemente más altos donde podrían observarse un comportamiento distinto, por lo tanto se hace necesario probar con otras sustituciones y obtener una dependencia general.

Anexos

Apéndice A

Aprendizaje Automático

A.1. Definiciones

En el aprendizaje automático, el proceso de desarrollar y evaluar modelos requiere dividir los datos disponibles en tres conjuntos principales: conjunto de entrenamiento (training set), conjunto de validación (validation set) y conjunto de pruebas (test set). Cada uno de estos conjuntos tiene un rol específico en el proceso:

Conjunto de entrenamiento (training set)

Es el conjunto de datos utilizado para entrenar el modelo de aprendizaje automático. Durante el entrenamiento, el modelo ajusta sus parámetros internos para aprender patrones y relaciones en los datos de entrada y sus correspondientes etiquetas o valores objetivo. El objetivo es que el modelo aprenda a hacer predicciones precisas a partir de estos datos. El conjunto de entrenamiento suele ser la mayor parte de los datos disponibles, generalmente alrededor del 60-80 % del total.

Conjunto de validación (validation set)

Este conjunto se utiliza para evaluar el rendimiento del modelo durante el proceso de entrenamiento y para ajustar los hiperparámetros del modelo. Los hiperparámetros son parámetros externos al modelo que no se aprenden durante el entrenamiento, pero que influyen en su rendimiento, como la tasa de aprendizaje, la profundidad máxima de un árbol de decisión o el número de capas en una red neuronal. El conjunto de validación permite realizar una evaluación objetiva del modelo sin utilizar el conjunto de pruebas, y ayuda a evitar el sobreajuste al conjunto de entrenamiento. Por lo general, representa alrededor del 10-20 % del total de datos disponibles.

Conjunto de pruebas (test set)

El conjunto de pruebas se utiliza para evaluar el rendimiento final del modelo en datos no vistos previamente, después de que se haya completado el entrenamiento y la selección de hiperparámetros. Este conjunto proporciona una estimación imparcial de cómo el modelo se desempeñará en el “mundo real” al enfrentarse a nuevos datos. Al igual que el conjunto de validación, el conjunto de pruebas suele representar alrededor del 10-20 % del total de datos disponibles.

La división de datos en estos tres conjuntos es crucial para garantizar que el modelo de aprendizaje automático tenga un buen rendimiento en datos no vistos, y permite evaluar su capacidad de generalización de manera objetiva. Además, ayuda a evitar problemas comunes como el sobreajuste y permite ajustar los hiperparámetros de manera efectiva.

A.2. Aprendizaje supervisado

Sea $\mathbf{x}_m \in \mathbb{R}^n$ el vector de características del m -ésimo dato de entrenamiento, también llamado input. y_m es la etiqueta del m -ésimo dato de entrenamiento (también llamado target). Un par $(\mathbf{x}_m, y_m)$ es un dato de entrenamiento (train example) y un grupo de M datos de entrenamientos

$$S_M = \{(\mathbf{x}_m, y_m)\}_{m=1}^M$$

es un training set. $\mathcal{X}$ es el espacio de características y $\mathcal{Y}$ es el espacio de etiquetas (output space). Nuestro objetivo es encontrar una función (Comunmente llamada *hypothesis* o *learner*)

$$f : \mathcal{X} \rightarrow \mathcal{Y},$$

$$\mathcal{X} \subset \mathbb{R}^n,$$

$$y \in \mathcal{Y} \subset \mathbb{R}$$

que aproxime la relación entre los datos de entrenamiento. Un *learning algorithm* es una función del espacio de datos de entrenamiento S_M al espacio de hipótesis $\mathcal{H}$ de posibles funciones f :

$$\mathcal{A} : S \rightarrow \mathcal{H}.$$

$$S_M \rightarrow \mathcal{A}(S_M) = f.$$

A.2.1. Regresión polinomial

El problema de regresión polinomial en *Aprendizaje automático* consiste en encontrar una función polinomial p que aproxime una relación entre una variable independiente x y una variable dependiente y de la forma:

$$y = p(x) + \epsilon,$$

donde ϵ es un error aleatorio. En el caso de que p sea un polinomio de grado n , el problema se reduce a encontrar los coeficientes β_i de la forma:

$$p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n.$$

tal que el error cuadrático medio (MSE) sea mínimo. El error cuadrático medio se define como:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - p(x_i))^2,$$

en python podemos definir una función que calcule el MSE de la siguiente manera:

```
1 def MSE(y, y_pred):
2     return np.mean((y - y_pred)**2)
```

donde y es un vector con los valores reales y y_pred es un vector con los valores predichos por el modelo.

Notamos que este problema es distinto al de interpolación, ya que en este caso no se busca que la función p pase por todos los puntos, sino que se busca una función que aproxime la relación entre x e y .

Para buscar el mínimo de la función MSE usamos algún método de optimización.

A.3. SVM

SVM es un algoritmo de aprendizaje supervisado que se puede usar tanto para clasificación como para regresión. El objetivo de SVM es encontrar el hiperplano óptimo en un espacio N-dimensional que

clasifique claramente los puntos de datos. Buscamos un hiperplano que separe los datos, un problema de clasificación binaria, por lo que buscamos un hiperplano que separe los datos en dos clases (y_m , que puede ser +1 o -1), de la mejor manera posible.

Clasificación binaria

El problema de clasificación binaria es un problema de clasificación en el que los datos solo pueden clasificarse en dos clases. Por ejemplo, si queremos clasificar los correos electrónicos como spam o no spam, es un problema de clasificación binaria donde los correos electrónicos solo pueden clasificarse en dos clases, spam y no spam.

Hiperplano

Un hiperplano es un subespacio de dimensión $n - 1$ de un espacio vectorial de dimensión n . Se puede definir mediante una ecuación lineal de la forma:

$$\beta \mathbf{v}^T + \beta_0 = 0,$$

donde $\beta \in \mathbb{R}^n$ y $\beta_0 \in \mathbb{R}$. Es de dimensión $n - 1$ porque hay una sola ecuación y n variables (las componentes de $\mathbf{v}$).

Datos linealmente separables

Los datos son linealmente separables si existe un hiperplano que los separe.

Margen

El margen se calcula como la distancia perpendicular desde el hiperplano a los puntos más cercanos de cada clase entre las dos clases en el conjunto de datos de entrenamiento. En el caso de los datos linealmente separables queremos maximizar la distancia mínima de cualquier punto al hiperplano, que es el “margen”.

Distancia de un punto al hiperplano

La distancia desde un punto $(\mathbf{x}_m, y_m)$ del conjunto de entrenamiento al hiperplano definido por los parámetros β y β_0 se puede calcular utilizando la fórmula de distancia entre un punto y un hiperplano en un espacio euclidiano. La fórmula de distancia entre un punto $\mathbf{x}$ y un hiperplano definido por el vector de pesos β y el sesgo β_0 es:

$$\text{dist}(\mathbf{x}, \beta, \beta_0) = \frac{|\beta^T \mathbf{x} + \beta_0|}{\|\beta\|}$$

El problema de encontrar el hiperplano separador óptimo es equivalente a encontrar el hiperplano que maximiza el margen entre las dos clases en el conjunto de datos de entrenamiento. En el caso de los datos linealmente separables, queremos maximizar la distancia mínima de cualquier punto al hiperplano, que es el “margen”.

El criterio $D^*(\beta, \beta_0) = \sum_{m=1}^M y_m(\beta \mathbf{x}_m^T + \beta_0)$ representa la suma de los márgenes funcionales sobre todo el conjunto de entrenamiento. Lo que buscamos es maximizar la suma de los márgenes funcionales.

Apéndice B

Teoría del funcional de densidad

B.1. DFT

Se propone,

$$F[\rho] = \int f(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \quad (\text{B.1})$$

La función $f(\mathbf{r})$ usualmente depende de otras funciones bien definidas. En el caso de la DFT, la función depende de la densidad electrónica, lo cual haría Q una funcional de $\rho(\mathbf{r})$; en el caso más simple, $f(\mathbf{r})$ sería equivalente a la densidad [$f(\mathbf{r}) = \rho(\mathbf{r})$]. Si la función $f(\mathbf{r})$ dependiera de alguna manera de los gradientes (o de sus derivadas de orden superior) de $\rho(\mathbf{r})$, entonces al funcional se le conoce como “no local” o con corrección de gradiente. En contraste, una funcional “local” solamente tiene una dependencia simple sobre $\rho(\mathbf{r})$. En la DFT, la funcional de la energía se escribe como una suma de dos términos:

$$E[\rho] = \int F[\rho(\mathbf{r})]d\mathbf{r} = \int (E_{\text{KE}}[\rho(\mathbf{r})] + E_H[\rho(\mathbf{r})] + E_{\text{XC}}[\rho(\mathbf{r})]) d\mathbf{r} \quad (\text{B.2})$$

- $E[\rho]$: Esta es la energía total del sistema electrónico, que es el objetivo principal de la teoría del funcional de la densidad. Es la cantidad que se desea calcular y representa la energía total de todos los electrones en el sistema.
- $E_{\text{KE}}[\rho(\mathbf{r})]$: Este término representa la energía cinética electrónica. Es la energía asociada al movimiento de los electrones y se calcula en función de la densidad electrónica $\rho(\mathbf{r})$ y las funciones de

onda electrónicas.

- $E_H[\rho(\mathbf{r})]$: Este término representa la energía electrostática de Hartree. Representa la energía de repulsión electrostática entre los electrones. En la aproximación de Hartree, se asume que los electrones no están correlacionados y se trata cada electrón como si estuviera interactuando con la carga promedio de los otros electrones.
- $E_{XC}[\rho(\mathbf{r})]$: Este término representa la energía de intercambio y correlación. Es una corrección que tiene en cuenta las interacciones cuánticas entre los electrones, incluyendo los efectos de intercambio y correlación. La forma específica de esta corrección depende de la elección de la aproximación funcional en la DFT y es una de las partes más importantes de la teoría del funcional de la densidad.

Entonces tenemos que primer término de la ecuación anterior surge de la interacción de los electrones con un potencial externo $V_{\text{ext}}(\mathbf{r})$, el cuál típicamente es debido a la interacción de Coulomb con el núcleo. $F[\rho(\mathbf{r})]$ es la suma de la energía cinética de los electrones y la contribución de las interacciones electrónicas. El valor mínimo en la energía corresponde a la densidad electrónica del estado base exacto, entonces se puede hacer uso de una aproximación variacional; la mejor solución que corresponde al mínimo de la energía. Si se tiene una densidad incorrecta, esta dará una energía superior a la energía verdadera. Además, tenemos que existe una restricción sobre la densidad electrónica, esta tiene que ver con el número fijo de electrones,

$$N = \int \rho(\mathbf{r}) d\mathbf{r} \quad (\text{B.3})$$

Con el fin de minimizar la energía, introducimos esta restricción como un multiplicador de Lagrange $(-\mu)$ obteniendo,

$$\delta \left(F[\rho] - \mu \int \rho(\mathbf{r}) d\mathbf{r} \right) = 0 \quad (\text{B.4})$$

donde δ es la variación funcional, la cual es una operación matemática que se utiliza para encontrar cómo un funcional cambia cuando se varía la función en la que depende. En este caso, el funcional es $F[\rho]$, que es la energía en función de la densidad electrónica $\rho(r)$. La variación funcional δ se aplica tanto a $F[\rho]$ como a la restricción integrada. De lo cual resulta entonces la expresión,

$$\frac{\delta F[\rho]}{\delta \rho(\mathbf{r})} = \mu \quad (\text{B.5})$$

La ecuación anterior es equivalente en la DFT a la ecuación de Schrödinger, el subíndice V_{ext} indica que esto es ante las condiciones de potencial externo constante (i.e., posiciones nucleares fijas). El multiplicador de Lagrange μ se puede identificar con el potencial químico de una nube electrónica en su núcleo, el cual también está relacionado con la electronegatividad,

$$\chi^- = -\mu = -\frac{\delta E[\rho]}{\delta N} \quad (\text{B.6})$$

El segundo artículo de suma importancia para el desarrollo de la teoría del funcional de la densidad fue escrito por Khon y Sham quienes sugirieron una manera práctica de resolver el teorema de Hohenberg-Kohn para un conjunto de electrones interactuando. La dificultad con la ecuación (B.2) es que no conocemos la función $F[\rho(\mathbf{r})]$. Khon y Sham sugirieron que $F[\rho(\mathbf{r})]$ debería ser aproximada como la suma de tres términos:

$$F[\rho] = E_{\text{KE}}[\rho] + E_H[\rho] + E_{\text{XC}}[\rho] \quad (\text{B.7})$$

en donde E_{KE} es la energía cinética, E_H es la energía Coulombiana de la interacción electrón-electrón y E_{XC} contiene las contribuciones del intercambio y correlación. El primer término de la ecuación está definido como la energía cinética de un sistema de electrones no interactuantes con la misma densidad del sistema real,

$$F_{\text{KE}}[\rho] = \sum_i \int \frac{\hbar^2}{2m} |\nabla \phi_i(\mathbf{r})|^2 d\mathbf{r} - \sum_i \int \frac{e^2 |\phi_i(\mathbf{r})|^2}{|\mathbf{r} - \mathbf{R}_i|} d\mathbf{r} \quad (\text{B.8})$$

El segundo término $E_H[\rho(\mathbf{r})]$ es también conocido como la energía electrostática de Hartree. La aproximación de Hartree para resolver la ecuación de Schrödinger falla ya que no reconoce que los movimientos electrónicos están correlacionados. En la aproximación de Hartree esta energía electrostática surge de la interacción clásica entre dos densidades de carga, cuando se suma sobre todas las posibles interacciones a pares, se obtiene

$$E_H[\rho] = \frac{1}{2} \iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \quad (\text{B.9})$$

Combinando estas dos ecuaciones y añadiendo la interacción electrón-núcleo lleva a la expansión completa para la energía de un sistema de N-electrones dentro del esquema de Khon-Sham,

$$E[\rho] = \sum_i \int \frac{\hbar^2}{2m} |\nabla \phi_i(\mathbf{r})|^2 d\mathbf{r} - \sum_i \int \frac{e^2 |\phi_i(\mathbf{r})|^2}{|\mathbf{r} - \mathbf{R}_i|} d\mathbf{r} + \frac{1}{2} \iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + E_{XC}[\rho] - \sum_I \frac{Z_I e^2}{|\mathbf{R}_I|} \quad (\text{B.10})$$

Esta ecuación define la funcional de la energía de intercambio y correlación $E_{XC}[\rho(\mathbf{r})]$, la cual contiene no solamente contribuciones debidas al intercambio y la correlación sino también una contribución debida a la diferencia entre la energía cinética real del sistema y $E_{KE}[\rho(\mathbf{r})]$.

Khon y Sham escribieron la densidad $\rho(\mathbf{r})$ de un sistema como la suma del módulo cuadrado de un conjunto de orbitales ortonormales de un electrón,

$$\rho(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2 \quad (\text{B.11})$$

Introduciendo esta expresión para la densidad electrónica y aplicando la condición variacional se obtienen las siguientes ecuaciones de Khon-Sham de un electrón,

$$-\frac{\hbar^2}{2m} \nabla^2 \phi_i(\mathbf{r}) + V_{\text{eff}}(\mathbf{r}) \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}) \quad (\text{B.12})$$

En la ecuación (B.12) hemos escrito el potencial externo en la forma apropiada para la interacción con M núcleos. ϵ_i son las energías orbitales y V_{XC} es conocido como el funcional de intercambio-correlación, relacionado a la energía de intercambio-correlación por

$$E_{XC}[\rho] = \int \rho(\mathbf{r}) V_{XC}(\mathbf{r}) d\mathbf{r} \quad (\text{B.13})$$

La energía total es entonces calculada de la ecuación (B.10).

Antes de discutir el esquema de cálculo definimos un potencial efectivo V_{EF} de la siguiente manera:

$$V_{\text{EF}} = - \sum_I \frac{Z_I e^2}{|\mathbf{R}_I|} + \int V_{XC}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \quad (\text{B.14})$$

Y reescribimos la ecuación (B.12) como

$$-\nabla^2 \phi(r) + V_{\text{eff}}(r) = \epsilon \phi(r) \quad (\text{B.15})$$

La ecuación de Schrödinger (B.12) y la forma para la densidad electrónica (B.11) que se conocen como las ecuaciones de Kohn-Sham deben ser resueltas de manera autoconsistente. En la práctica se lleva a cabo el procedimiento de la siguiente manera:

1. Se propone una densidad electrónica inicial $\rho(\mathbf{r})$.
2. Se construye el potencial efectivo V_{eff} de la expresión (B.14).
3. A partir de las ecuaciones (B.15) y (B.11), se obtiene una nueva densidad $\rho(\mathbf{r})$.
4. Utilizando la ecuación (B.10), se calcula la energía total.

Si la densidad de carga final es igual a la inicial, la auto-consistencia ha sido llevada a cabo.

B.1.1. Aproximación de Hartree-Fock.

Otra aproximación que se realiza se basa en que la función de onda molecular total para un sistema de n electrones es muy difícil de obtener, sin embargo, una forma aproximada se puede construir a partir de un producto de *funciones anti-simétricas*¹ normalizadas arregladas en forma de un *determinante de Slater*[22] como sigue:

$$\Psi = \frac{1}{\sqrt{n!}} \begin{bmatrix} \phi_1(\mathbf{r}_1)\alpha(1) & \phi_1(\mathbf{r}_1)\beta(1) & \cdots & \phi_{n/2}(\mathbf{r}_1)\alpha(1) \\ \phi_1(\mathbf{r}_2)\alpha(2) & \phi_1(\mathbf{r}_2)\beta(2) & \cdots & \phi_{n/2}(\mathbf{r}_2)\alpha(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{r}_n)\alpha(n) & \phi_1(\mathbf{r}_n)\beta(n) & \cdots & \phi_{n/2}(\mathbf{r}_n)\alpha(n) \end{bmatrix}, \quad (\text{B.16})$$

en donde las funciones de onda ϕ son orbitales moleculares de un solo electrón mientras que las funciones α y β representan los dos estados de espín accesibles del electrón. A esta simplificación, a través de la cual se cambia el problema de n electrones interactuantes por el de n ecuaciones de un

¹Una función $f : X^n \rightarrow \mathbb{F}$ se llama *antisimétrica* si cambia su signo al intercambiar cualesquiera dos de sus argumentos, esto es, si para cualesquiera i, j en $\{1, \dots, n\}$ tales que $i \neq j$ y para cualesquiera $x_1, \dots, x_n$ en X se tiene la igualdad:

$$f(\dots, x_i, \dots, x_j, \dots) = -f(\dots, x_j, \dots, x_i, \dots).$$

Aquí x_i es el i -ésimo argumento y x_j es el j -ésimo argumento.

solo electrón se le denomina aproximación de Hartree-Fock [7] y modifica la forma para el Hamiltoniano electrónico en la Con esto, el sistema es descrito por n ecuaciones tipo Schrödinger de la forma,

$$f_i(1)\phi_i(1) = \epsilon_i\phi_i(1), \quad (\text{B.17})$$

las cuales son llamadas ecuaciones de Hartree-Fock. El operador f es el operador Fockiano del sistema (y es similar al operador Helec), el cual es definido como:

$$f_i(1) = -\frac{1}{2}\nabla_1^2 - \sum_{A=1}^N \frac{Z_A}{r_{1A}} + \sum_{j=1}^n \left[\int \phi_j^*(2) \frac{1}{r_{12}} \phi_j(2) d\tau_2 \right], \quad (\text{B.18})$$

en donde el término de interacción electrón-electrón de la ecuación (2.10) se ha extendido a dos términos los cuales están contenidos dentro de los paréntesis cuadrados.

B.1.2. Segundo Teorema de Hohenberg-Kohn

Del [teorema B.1.2](#) tenemos que

$$\frac{\delta}{\delta n} E[n(\mathbf{r})] \big|_{n=n_0} = 0 \quad (\text{B.19})$$

y

$$E_0 = E[n_0(\mathbf{r})]. \quad (\text{B.20})$$

Además la [eq. \(B.32\)](#):

$$\begin{aligned} E_v[\tilde{\rho}(\mathbf{r})] &= \int \tilde{\rho}(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} + F_{HK}[\tilde{\rho}(\mathbf{r})] \\ &= \langle \tilde{\psi} | H | \tilde{\psi} \rangle \end{aligned}$$

pero por el principio variacional:

$$\begin{aligned}
\langle \tilde{\psi} | H | \tilde{\psi} \rangle &\geq \langle \psi | H | \psi \rangle \\
&= E_v[\rho(\mathbf{r})] \\
&= E_0
\end{aligned}$$

B.1.3. Hamiltoniano molecular

Una molécula es un arreglo estable de un grupo de núcleos y electrones interactuantes. Cuando una molécula se forma, los electrones de las capas internas de cada átomo permanecen fuertemente ligados al núcleo original y casi no son perturbados. Sin embargo, los electrones de las capas exteriores tienen una menor interacción con el núcleo del átomo, por lo que son fuertemente perturbados e influenciados por todas las partículas del sistema. Al acercarse los átomos entre sí, sus funciones de onda se modifican significativamente hasta conducir al enlace, donde se obtienen configuraciones de mínima energía muy precisas y así determinando la mayoría de las propiedades observadas a escala macroscópica. La utilidad de las funciones de onda es de tal importancia que se justifica el trabajo requerido para obtenerlas.

El punto de partida para definir y resolver cualquier problema mecanicocuántico es plantear la ecuación de Schrödinger dependiente del tiempo. La forma completa de esta ecuación es:

$$\left[-\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V \right] \Psi_{rt}(r, t) = i\hbar \frac{\partial \Psi_{rt}(r, t)}{\partial t}. \quad (\text{B.21})$$

La [eq. \(B.21\)](#) describe una partícula de masa m que está influenciada de un campo externo V . Cuando se tiene el caso particular en que el potencial externo no depende del tiempo, la función de onda de la partícula puede ser escrita como el producto de una función espacial y una parte temporal

$$\Psi_{rt}(r, t) = \underbrace{\psi_r(r)}_{\text{espacial}} \underbrace{T_t(t)}_{\text{temporal}} \quad (\text{B.22})$$

Por lo tanto, la ecuación [eq. \(B.21\)](#) es separable en una parte espacial y una temporal y podemos entonces obtener una ecuación de Schrödinger independiente del tiempo definida como,

$$\left[-\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V \right] \psi(r) = E\psi(r) \quad (\text{B.23})$$

donde E es la energía de la partícula.

Podemos ver que al lado izquierdo de la eq. (B.23) se encuentra un operador actuando sobre la función de onda $\psi(r)$, el cual es el Hamiltoniano del sistema y se define como,

$$H = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V \quad (\text{B.24})$$

quedado finalmente la eq. (B.23) expresada en forma compacta como:

$$H\psi(r) = E\psi(r) \quad (\text{B.25})$$

La eq. (B.25) es claramente una ecuación de valores propios y sólo se puede resolver de una manera exacta para casos muy particulares del potencial V como por ejemplo: *Una partícula en una caja, una partícula sujeta al potencial de oscilador armónico o una partícula en un anillo, la partícula sobre una esfera*. Estas condiciones son ideales, sin embargo, en condiciones prácticas, la solución de dicha ecuación se obtiene de manera aproximada.

El caso más sencillo es el átomo de hidrógeno, debido a que es el átomo más simple que existe y el único que admite una solución analítica exacta. Además del átomo de hidrógeno, la ecuación también puede ser resuelta para átomos con núcleos de la forma Ze que sólo tienen un electrón, en los cuales la energía potencial dependerá únicamente de la distancia existente entre el electrón y su núcleo por medio de la ecuación de Coulomb. Con ello el Hamiltoniano del sistema definido en la eq. (B.24) toma la forma,

$$H = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) - \frac{Ze^2}{4\pi\epsilon_0 r} \quad (\text{B.26})$$

el cual se puede escribir en unidades atómicas como,

$$H = -\frac{1}{2}\nabla^2 - \frac{Z}{r}. \quad (\text{B.27})$$

La complejidad aumenta cuando se intenta describir a un sistema que contiene dos o más electrones. En estos casos, la ecuación de Schrödinger independiente del tiempo [eq. \(B.23\)](#) para un sistema de n electrones y N núcleos adopta la forma general,

$$\left(\overbrace{-\frac{1}{2} \sum_{i=1}^N \nabla_i^2 - \frac{1}{2} \sum_{i=1}^n \nabla_i^2}^{\text{cinética de los núcleos y electrones}} \underbrace{- \sum_{i=1}^n \sum_{A=1}^N \frac{Z_A}{R_{Ai}}}_{\text{atracción electrón-núcleo}} \underbrace{+ \sum_{i<j}^n \frac{1}{r_{ij}}}_{\text{repulsión electrones}} + \underbrace{\sum_{A<B}^N \frac{Z_A Z_B}{R_{AB}}}_{\text{repulsión núcleos}} \right) \psi(\mathbf{R}, \mathbf{r}) = E \psi(\mathbf{R}, \mathbf{r}). \quad (\text{B.28})$$

En la [eq. \(B.28\)](#), el Hamiltoniano molecular H contiene los términos de energía cinética de los núcleos (primer término) y los electrones (segundo término), la energía de atracción electrón-núcleo (tercer término), la interacción repulsiva electrón-electrón (cuarto término), y la interacción de repulsión entre los diferentes núcleos (último término); $\psi(\mathbf{R}, \mathbf{r})$ es la función de onda molecular total, $\mathbf{R}$ es el vector es la función de onda molecular total, E es la energía molecular total, A y B designan a un par de núcleos cualesquiera, i y j a un par de electrones arbitrarios, Z_A y Z_B son los números atómicos de los núcleos.

Para resolver [eq. \(B.23\)](#) y [eq. \(B.28\)](#) cuando nos enfrentamos a sistemas moleculares como los considerados en este trabajo, se deben introducir algunas aproximaciones para simplificar los cálculos. Una de estas aproximaciones es la de *Born-Oppenheimer* la cual se describe a continuación ([apéndice B.1.4](#)).

B.1.4. Aproximación de Born-Oppenheimer

En esta aproximación se considera que los electrones se mueven mucho más rápido que los núcleos, por lo que se asume que los electrones se mueven a través de un potencial generado por núcleos estacionarios. Matemáticamente, esto tiene como consecuencia que la solución de la ecuación es de variables separables en r y R , por lo tanto la solución $\psi(R, r)$ se puede escribir como un producto de la forma,

$$\psi(R, r) = \underbrace{\psi(R)}_{\text{nuclear}} \underbrace{\phi(r)}_{\text{electrónica}}, \quad (\text{B.29})$$

en donde $\psi(R)$ es la función de onda nuclear y $\phi(r)$ es la función de onda electrónica.

Esta aproximación permite que el Hamiltoniano electrónico sea considerado por un conjunto de coordenadas nucleares fijas y así poder plantear un Hamiltoniano solo para los electrones H_{elec} ,

$$H_{elec} = -\frac{1}{2} \sum_{i=1}^n \nabla_i^2 - \sum_{i=1}^n \sum_{A=1}^N \frac{Z_A}{R_{Ai}} + \sum_{i < j}^n \frac{1}{r_{ij}} \quad (\text{B.30})$$

en el cual los núcleos no contribuyen a la energía cinética del sistema y la interacción núcleo-núcleo se convierte en una constante del sistema para una posición fija de los iones.

B.1.5. Teoremas Hohenberg-Kohn

El aspecto clave de la teoría funcional de densidad radica en reemplazar la función de onda multidimensional con una densidad de electrones 3D a partir de la cual se pueden derivar todas las propiedades del estado fundamental, según lo propuesto por Hohenberg y Kohn en [9]. Los teoremas son:

Teorema B.1.1 (Hohenberg-Kohn) *El potencial externo $V_{ex}(r)$ es un único funcional de la densidad electrónica $\rho(r)$ y a su vez $V_{ex}(r)$ fija el hamiltoniano $\hat{H}$, concluyendose que el completo estado fundamental es un funcional único de la densidad $\rho(r)$.*

Teorema B.1.2 (Hohenberg-Kohn 2) *El funcional que entrega la energía del estado fundamental del sistema proporciona la energía más baja si y sólo si la densidad de entrada es la verdadera densidad del estado fundamental.*

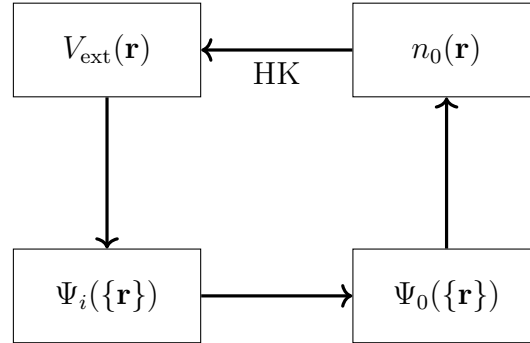


Figura B.1: [15] Representación del Teorema de Hohenberg-Kohn.

El [teorema B.1.2](#) es equivalente a decir que para una densidad cualquiera $\tilde{n}(\mathbf{r})$, que satisface las condiciones $\tilde{n}(\mathbf{r}) \geq 0$ y $\int \tilde{n}(\mathbf{r}) d\mathbf{r} = N$, se cumple la desigualdad (demostrada en [apéndice B.1.2](#)):

$$E_0 \leq E_v[\tilde{n}(\mathbf{r})] \quad (\text{B.31})$$

donde $E_v[\tilde{n}(\mathbf{r})]$ es el funcional de la energía calculada como:

$$E_v[\tilde{n}(\mathbf{r})] = \int \tilde{n}(\mathbf{r})v(\mathbf{r})d\mathbf{r} + F_{HK}[\tilde{n}(\mathbf{r})] \quad (\text{B.32})$$

entonces, del [teorema B.1.1](#) y [teorema B.1.2](#):

- Cada observable para el estado fundamental de un sistema electrónico en cada estado estacionario incluyendo la energía, puede ser calculado en principio de forma exacta únicamente a partir de la densidad, es decir, que cada observable puede ser expresado como un funcional de la densidad.
- La densidad del estado fundamental puede ser calculada, en principio de forma exacta, usando argumentos de tipo variacional.
- La energía del estado fundamental de un sistema de n electrones es un funcional único de la densidad electrónica $n(\mathbf{r})$: $E_0 \sim E[n(\mathbf{r})]$.
- La energía del estado fundamental se puede obtener en la densidad de carga de equilibrio $n_0(\mathbf{r})$ minimizando la energía funcional con respecto a la densidad electrónica, que es

$$E_0 = \min\{E[n(\mathbf{r})]\} = E[n_0(\mathbf{r})].$$

donde la energía total funcional del estado fundamental de un sistema de n electrones se puede escribir como

$$E_0[n_0] = T[n_0] + V_{N-e}[n_0] + V_{e-e}[n_0], \quad (\text{B.33})$$

donde T , V_{N-e} y V_{e-e} representan la energía cinética electrónica, las interacciones electrón-núcleo y las interacciones electrón-electrón, respectivamente. El efecto de los núcleos sobre los electrones, $V_{N-e}[n_0]$, puede considerarse como un campo externo que también está determinado por la densidad electrónica,

$$V_{N-e}[n_0] = \int n_0(\mathbf{r})v_{\text{ext}}(\mathbf{r})d\mathbf{r}.$$

la ecuación [eq. \(B.33\)](#) se puede reescribir como

$$E_0[n_0] = \int n_0(\mathbf{r})v_{\text{ext}}(\mathbf{r})d\mathbf{r} + F[n_0],$$

donde el funcional HK

$$F[n_0] = T[n_0] + V_{e-e}[n_0] \quad (\text{B.34})$$

que es independiente del potencial externo.

Los teoremas de Hohenberg-Kohn dan a la teoría del funcional de la densidad una poderosa base física pero, desafortunadamente, se desconoce la forma exacta del funcional de Hohenberg-Kohn.

B.1.6. Método Kohn-Sham

El método de Kohn-Sham proporciona una manera práctica de llevar a cabo la búsqueda. La estrategia dentro del método de Kohn-Sham se puede resumir como:

- Simplificar un problema de un sistema de n -cuerpos a n problemas de sistemas de un cuerpo.
- Separar un sistema interactivo en un sistema no interactivo efectivo y términos de corrección adicionales.

Primero se introduce un sistema de referencia con n electrones no interactuantes. Se asume que la densidad electrónica $n(\mathbf{r})$ del estado fundamental de un sistema auxiliar es la misma que la densidad electrónica del estado fundamental del sistema interactivo original. El Hamiltoniano del sistema auxiliar se escribe entonces como,

$$\hat{H}_r = \sum_{i=1}^n \hat{h}_{r_i} = \sum_{i=1}^n \left[-\frac{\hbar^2}{2m_i} \nabla_i^2 + v_{\text{eff}}^r(\mathbf{r}_i) \right], \quad (\text{B.35})$$

donde $\hat{h}_r$ se refiere al Hamiltoniano de Kohn-Sham de un electrón. La ecuación de Schrödinger para cada electrón en el sistema auxiliar es,

$$\hat{h}_{r_i} \psi_{r_i} = \epsilon_{r_i} \psi_{r_i}. \quad (\text{B.36})$$

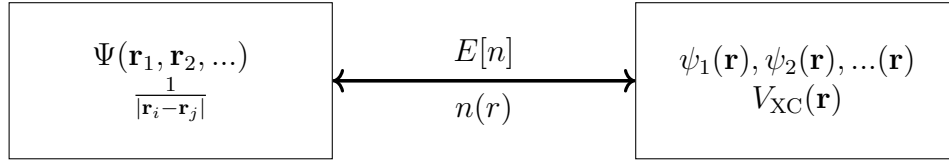


Figura B.2: Sistema interactivo y no interactivo

Basado en el sistema auxiliar, el funcional $F[n_0]$ (eq. (B.34)) del sistema interactivo puede reescribirse como

$$F[n_0] = T[n_0] + V_{e-e}[n_0] \quad (\text{B.37})$$

$$= (T_r[n_0] + T'[n_0]) + \left(\frac{e^2}{2} \iint \frac{n_0(\mathbf{r})n_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + V'_{e-e}[n_0] \right) \quad (\text{B.38})$$

$$= (T_r[n_0] + \frac{e^2}{2} \iint \frac{n_0(\mathbf{r})n_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}') + E_{xc}[n_0]. \quad (\text{B.39})$$

Donde la energía cinética $T[n_0]$ y la interacción electrón-electrón $V_{e-e}[n_0]$ pueden ser explícitamente divididas en un término de referencia y un término extra, respectivamente. Los dos términos extra, $T'[n_0]$ y $V'_{e-e}[n_0]$, constituyen la energía funcional de intercambio-correlación,

$$E_{xc}[n_0] = T'[n_0] + V'_{e-e}[n_0]. \quad (\text{B.40})$$

Luego, la energía total del estado fundamental funcional en la eq. (B.33) con la suposición de que $n_r(\mathbf{r}) = n_0(\mathbf{r})$ será,

$$E_0[n_r] = T_r[n_r] + \int n_r(\mathbf{r})v_{\text{ext}}(\mathbf{r})d\mathbf{r} + \underbrace{\frac{e^2}{2} \iint \frac{n_r(\mathbf{r})n_r(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}'}_{E_H[n_r(\mathbf{r})]^2} + E_{xc}[n_r]. \quad (\text{B.41})$$

En el sistema de referencia (auxiliar) es fácil calcular la densidad electrónica a partir de las funciones de onda de un electrón. Luego nos queda el problema de encontrar las funciones de onda de un electrón apropiadas ψ_{r_i} que minimizan la energía total funcional del sistema en el que estamos interesados. Según el algoritmo de Kohn-Sham, la energía del estado fundamental E_0 se puede obtener de la eq. (B.36) resolviendo n ecuaciones de Schrödinger de una partícula (ecuaciones de Kohn-Sham):

²Energía de Hartree, interacción de Coloumb electrón-electrón.

$$\left\{ -\frac{\hbar^2}{2m_i} \nabla_i^2 + v_{\text{eff}}^r(\mathbf{r}) \right\} \psi_{r_i}(\mathbf{r}) = \epsilon_{r_i} \psi_{r_i}(\mathbf{r}),$$

estableciendo que

$$n_r(\mathbf{r}) = \sum_{i=1}^N n_{r_i}(\mathbf{r}) = \sum_{i=1}^N |\psi_{r_i}(\mathbf{r})|^2,$$

donde

$$v_{\text{eff}}^r(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + e^2 \int \frac{n_r(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + \mu_{\text{xc}}[n_r(\mathbf{r})],$$

y

$$\mu_{\text{xc}}[n_r(\mathbf{r})] = \frac{\delta E_{\text{xc}}[n_r(\mathbf{r})]}{\delta n_r(\mathbf{r})},$$

es llamado el potencial de intercambio-correlación.

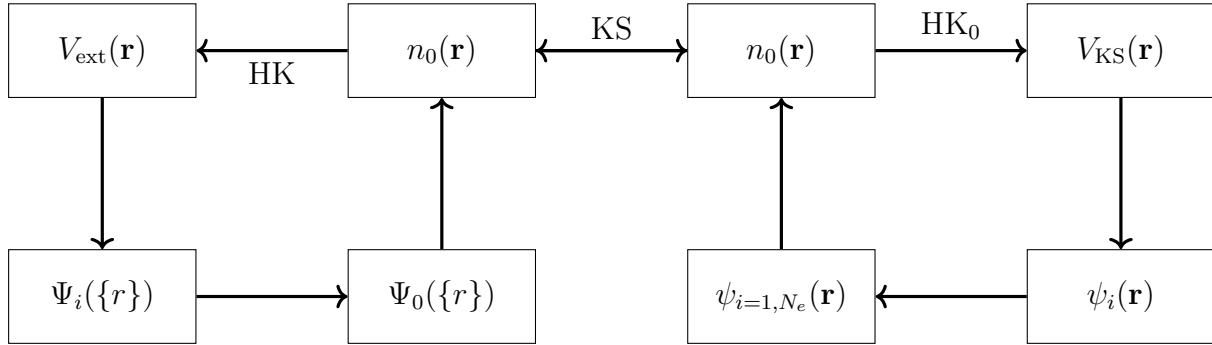


Figura B.3: Representación de Kohn-Sham.

La parte izquierda de la [fig. B.3](#) representa el sistema de muchos cuerpos y la derecha representa el sistema de partículas independientes. Entre estos dos sistemas hay conexión en ambas direcciones, lo que proporciona Kohn-Sham. De la solución del sistema de partículas independientes podemos recibir todas las propiedades del sistema de muchos cuerpos.

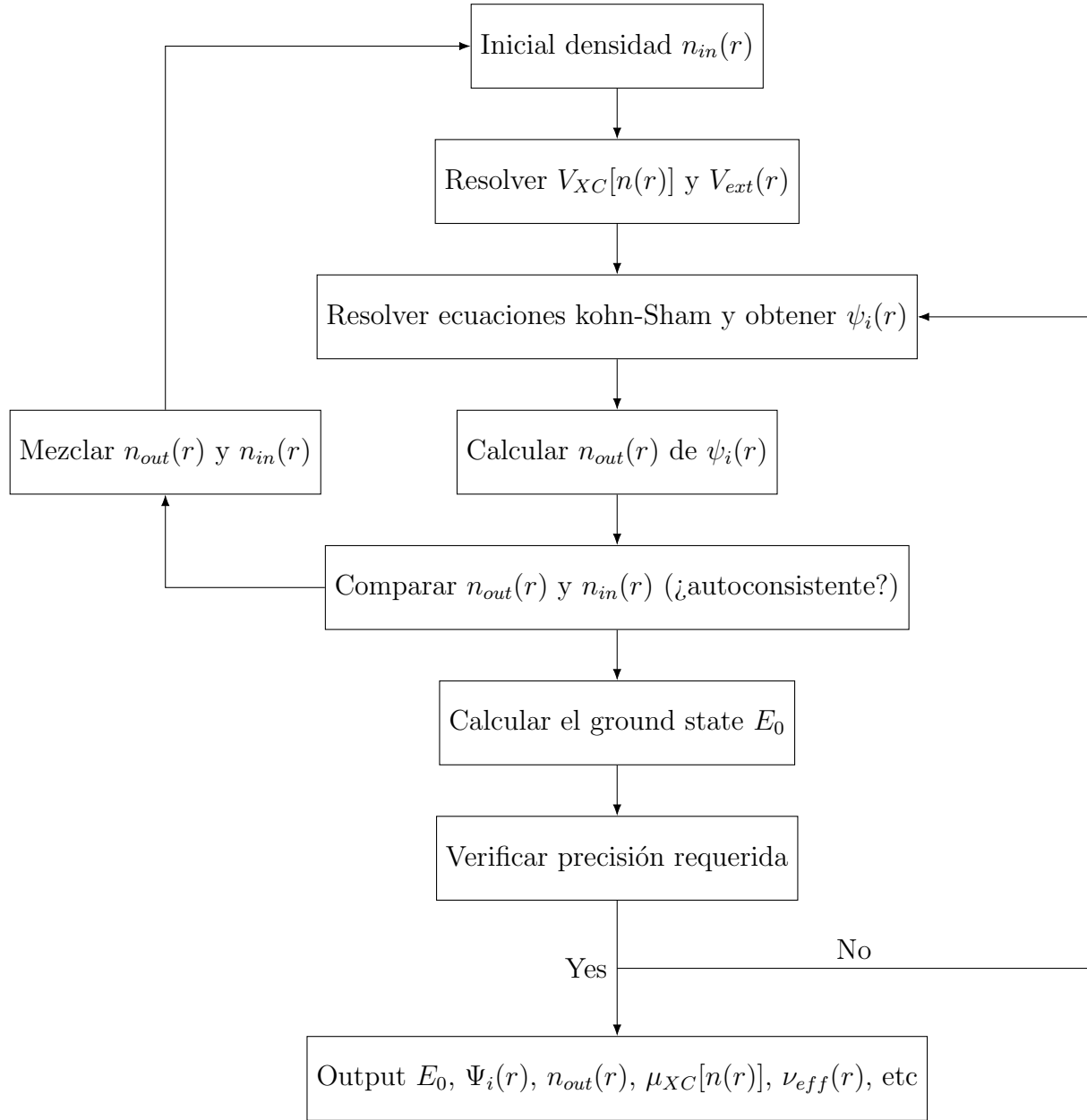


Figura B.4: El diagrama de flujo para resolver de forma autoconsistente las ecuaciones de Kohn-Sham para obtener la energía del estado fundamental.

B.1.7. Funcionales de Correlación-Intercambio

Para el enfoque de Kohn-Sham necesitamos conocer E_{XC} y, debido a ello, necesitamos aproximar el funcional de intercambio-correlación E_{XC} . Para lograr esto existen varias aproximaciones, ya que las expresiones analíticas exactas para los funcionales de la energía de intercambio y correlación sólo se conocen para el caso especial de un gas de electrones libres.

La “escalera de Jacob”, mostrada en la [fig. B.5](#) es una forma de mostrar familias de aproximaciones

funcionales de intercambio-correlación DFT [20].

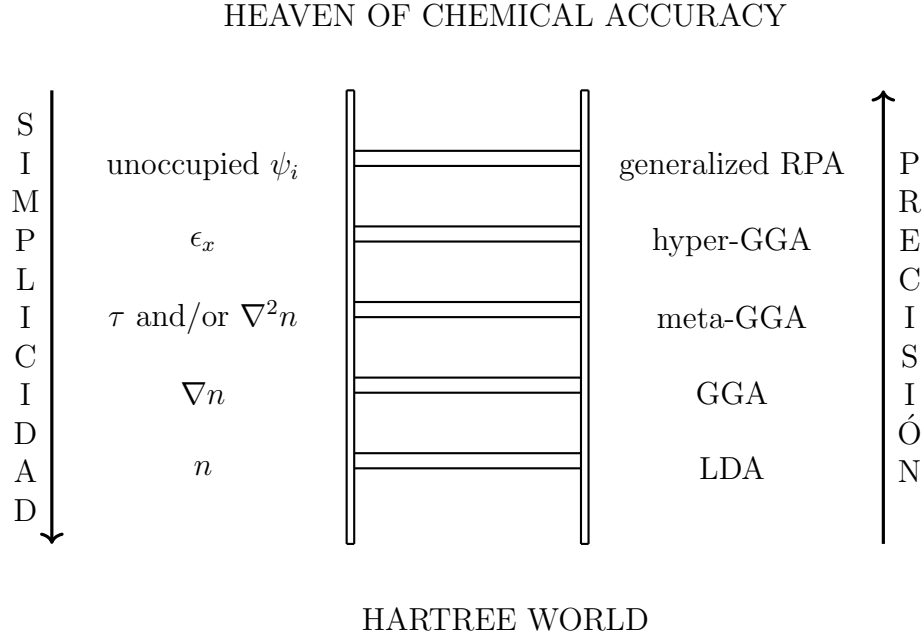


Figura B.5: Escalera de Jacob

Las más simples son la aproximación *homogeneous electron gas* (HEG) y la *local density approximation* (LDA), donde esta última es la utilizada por el programa *ab-initio* utilizado en este trabajo, *VASP*.

La aproximación gas homogéneo de electrones (HEG) asume que se conoce el potencial de correlación intercambio en cada posición [8],

$$V_{XC}^{\text{HEG}}(r) = V_{XC}^{\text{HEG}}[n(r)], \quad (\text{B.42})$$

La idea de la **aproximación de densidad local** (LDA) es construir E_{xc} a partir del conocimiento de la energía de intercambio-correlación por partícula, ϵ_{xc} , del gas de electrones homogéneo [9]. Se asume que la energía de intercambio-correlación depende sólo de la densidad electrónica local, la cual es uniforme al rededor de un punto en el espacio.

$$E_{XC}^{\text{LDA}}[n(\mathbf{r})] = \int n(\mathbf{r}) \underbrace{\epsilon_{XC}^{\text{HEG}}[n(\mathbf{r})]}_{\epsilon_{XC}^{\text{LDA}}} d\mathbf{r}, \quad (\text{B.43})$$

donde $\epsilon_{XC}^{\text{LDA}}[n(\mathbf{r})]$ es la energía de intercambio-correlación por electrón de un gas electrónico homogéneo con la densidad local $n(\mathbf{r})$ constante en todos los puntos del espacio. Es exacto solo para sistemas cuya variación de densidad electrónica es suficientemente lenta.

$$\epsilon_{XC}^{\text{LDA}}[n(\mathbf{r})] = \epsilon_X^{\text{HEG}}[n(\mathbf{r})] + \epsilon_C^{\text{HEG}}[n(\mathbf{r})] \quad (\text{B.44})$$

donde $\epsilon_X^{\text{HEG}}[n(\mathbf{r})]$ es la energía de intercambio de un sistema homogéneo de electrones, obtenido por las aproximaciones de *Hartree-Fock* y $\epsilon_C^{\text{HEG}}[n(\mathbf{r})]$ es la energía de correlación de el sistema que puede ser calculada con el método *Quantum Monte Carlo*[15].

La variación de la densidad electrónica es suficientemente lenta en algunos sistemas. El rendimiento real muestra que la LDA puede proporcionar resultados satisfactorios para sistemas con enlaces covalentes, iónicos y metálicos, especialmente para aquellos con densidad electrónica homogénea. Sin embargo, una deficiencia bien conocida de la LDA es la sobreestimación de la fuerza del enlace. Las constantes de red y las energías del estado fundamental atómico obtenidas con LDA, por lo tanto, generalmente se subestiman mientras que las energías de cohesión se sobreestiman. Las brechas de banda en semiconductores y aislantes se subestiman seriamente.

Los cálculos *ab-initio* pueden ser extremadamente costosos en términos del requerimiento computacional necesarios. A pesar de eso, los avances en infraestructura de computación y la disponibilidad de softwares fáciles de manejar han ayudado a que los métodos *ab-initio* sean una herramienta computacional usada ampliamente.

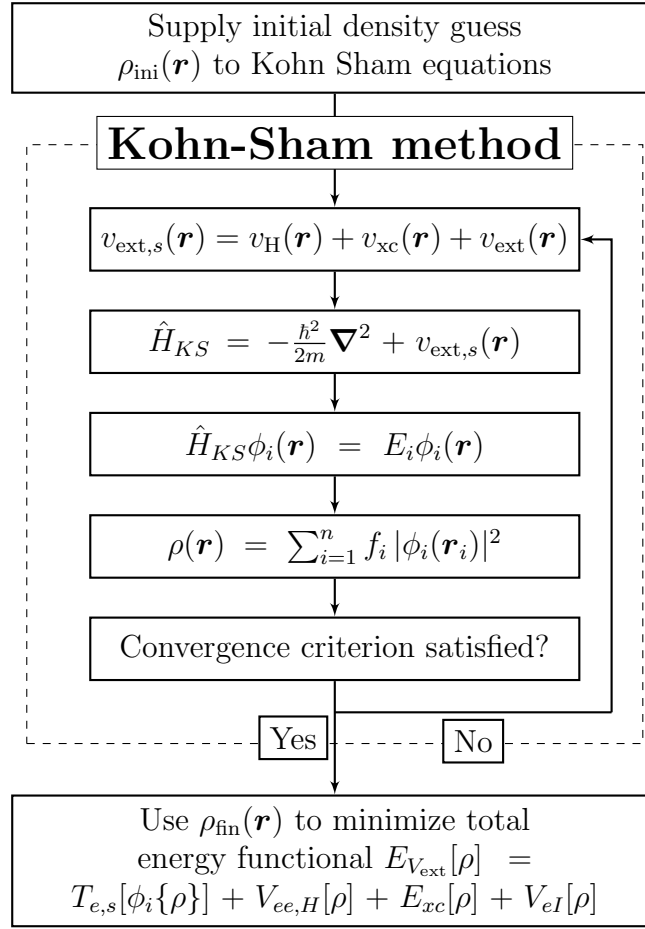


Figura B.6: Diagrama de flujo del método de Kohn-Sham.

Apéndice C

Calculos por medio de VASP

VASP ofrece soporte para una gran variedad de plataformas, como Pentium, Athlon, IBM, HP, Cray y SUN(-). Se encuentra disponible computación de alto rendimiento (computadoras en paralelo). Todas estas características hacen que el paquete sea muy adecuado para cálculos de sistemas grandes extendidos. En principio, en la computación paralela, cuantos más procesadores se utilizan, más rápido es el cálculo. Sin embargo, para los clústeres de computadoras, dependiendo del tipo de procesador, siempre hay una cantidad crítica de CPU, mayor que la cual la velocidad de cálculo no aumenta linealmente con la cantidad de CPU utilizadas; por ejemplo, la computación en cuatro nodos puede ser dos veces mayor. más rápido que computar en dos nodos, pero computar en ocho nodos no necesariamente mejora la eficiencia hasta cuatro veces. Un cálculo VASP típico requiere al menos cuatro archivos de entrada básicos: INCAR, POTCAR, POSCAR y KPOINTS.

C.0.1. input file

POSCAR

Información sobre las posiciones y dela geometría del sistema

```

1  vasp
2  1.00000000
3  8.213286    0.000000    0.000000
4  0.000000    8.213286    0.000000
5  0.000000    0.000000    8.213286
6  Ta Ti
7  3  29
8  Direct
9  0.000000    0.000000    0.000000
10 0.000000    0.250000    0.250000
11 0.250000    0.000000    0.750000
12 0.000000    0.000000    0.500000
13 0.000000    0.500000    0.000000
14 0.000000    0.500000    0.500000
15 0.500000    0.000000    0.000000
16 0.500000    0.000000    0.500000
17 0.500000    0.500000    0.000000
18 0.500000    0.500000    0.500000
19 0.000000    0.250000    0.750000
20 0.000000    0.750000    0.250000

```

Listing C.1: 2x2-TaTi_03/CALCS/poscar/c00012.vasp lineas
0-20

C.0.2. output file

Archivos de salida estándar:

OUTCAR

El archivo de salida principal incluye información sobre todo el proceso de cálculo.

CONTCAR

Datos reales de la estructura una vez completado el cálculo.

OSZICAR

Información sobre pasos electrónicos e iónicos - itteraciones electrónicas

C.0.3. E0 Efree

Corresponde a la energía calculada para cada configuración de impurezas. Este resultado es obtenido por el programa [vasp](#) y se encuentra en el archivo OUTCAR.

C.0.4. Energías de la iteración

Aquí la idea es hablar un poco de la información del output

EO, EWOE, EWOE.sigma_to_0

```

1
2
3 ----- Iteration      1( 14)
4
5
6 POTLOK:  cpu time    0.0435: real time    0.0437
7 SETDIJ:  cpu time    0.2591: real time    0.2593
8 EDDIAG:  cpu time    5.9218: real time    5.9468
9 RMM-DIIS: cpu time   34.8733: real time   35.2812
10 ORTHCH:  cpu time    1.0413: real time    1.0467
11 DOS:     cpu time    0.0183: real time    0.0183
12 CHARGE:  cpu time    1.5363: real time    1.5449
13 MIXING:  cpu time    0.0045: real time    0.0045
14
15 LOOP:    cpu time   43.6982: real time   44.1454
16
17 eigenvalue-minimisations : 30168
18 total energy-change (2. order) :-0.2854391E-02 (-0.7878084E-02)
19 number of electron      323.0000013 magnetization
20 augmentation part      146.8051483 magnetization
21
22 Broyden mixing:
23 rms(total) = 0.36172E-01      rms(broyden)= 0.36167E-01
24 rms(prec ) = 0.56196E-01
25 weight for this iteration      100.00
26
27 eigenvalues of (default mixing * dielectric matrix)
28 average eigenvalue GAMMA= 0.8915
29 2.1478 1.4374 1.4374 0.8711 0.5603 0.5603 0.4070 0.3226 0.2792
30
31 Free energy of the ion-electron system (eV)
32 -----
33 alpha Z      PSCENC =      2208.17025628
34 Ewald energy TEWEN =     -26212.48557488
35 -Hartree energy DENC =     -6003.89355032
36 -exchange    EXHF  =           0.00000000
37 -V(xc)+E(xc) XCENC =      1073.96785912
38 PAW double counting =      34582.85394504 -34625.36851179
39 entropy T*S   EENTRO =       -0.02773931
40 eigenvalues   EBANDS =     -4659.44316874
41 atomic energy EATOM =      33376.87011986
42 Solvation     Ediel_sol =           0.00000000
43 -----
44 free energy   TOTEN  =       -259.35636473 eV
45
46 energy without entropy =     -259.32862543 energy(sigma->0) =
-259.34249508

```

Listing C.2: 2x2-TaTi_03/CALCS/outcar/OUTCAR.c00012 líneas 1585-1632 donde podemos ver las energías de la iteración.

iteration	Efree
0	4263.247652090
1	80.446540950
2	-264.044923580
3	-293.252952110
4	-294.212210870
5	-262.369933150
6	-262.156710950
7	-259.703544540
8	-259.402176330
9	-259.408697490
10	-259.329049170
11	-259.360232940
12	-259.325998350
13	-259.328625430
14	-259.328941050
15	-259.328747630
16	-259.329031620
17	-259.329050720
18	-259.329119510
19	-259.329118600
20	-259.329130210
21	-259.329135290
22	-259.329138500
23	-259.329138290
24	-259.329140330
25	-259.329140320

Cuadro C.1: Energía para en cada iteración (26 iteraciones) para la configuración c00012 de 3 impurezas donde EDIFF = 1.0e-06 en el archivo OUTCAR

Iteraciones vs Efree para c00012 (3 impurezas)

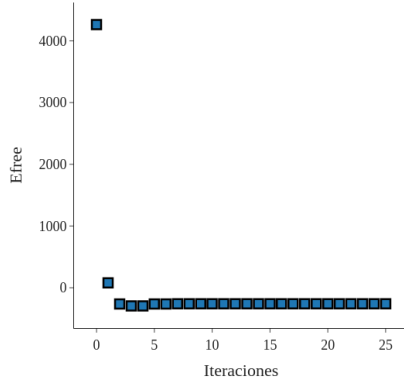


Figura C.1: E0 vs número de iteraciones (0 a 26) para una configuración dada

Iteraciones vs Efree para c00012 (3 impurezas)

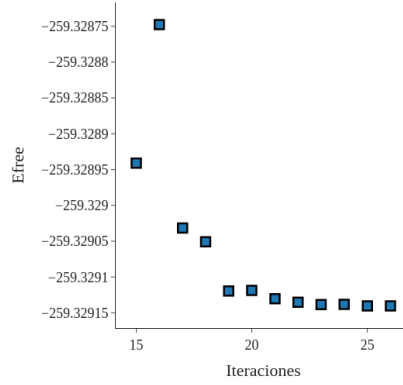


Figura C.2: E0 vs número de iteraciones (15 a 26) para una configuración dada

Iteraciones vs Efree para c00012 (3 impurezas)

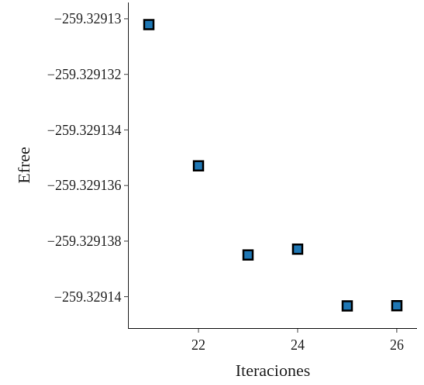


Figura C.3: E0 vs número de iteraciones (21 a 26) para una configuración dada

El criterio para escoger los k - points se encuentra en [18]

Apéndice D

Simetría de las configuraciones

D.1. Simetrías

La simetría de las configuraciones se analizó principalmente mediante el uso de ASE[11]. En las siguientes subsecciones se utilizan los descriptores `nsymop` y `nsymop_rot`, los cuales son obtenidos por medio de ASE las funciones `nsymop` y `rotations` actuando sobre la clase `Spacegroup()` a la que pertenece cada configuración.

D.1.1. Número de peraciones de simetría

Se consideran número de simetrías de **rotación**, **translación** y **subtranslación**.

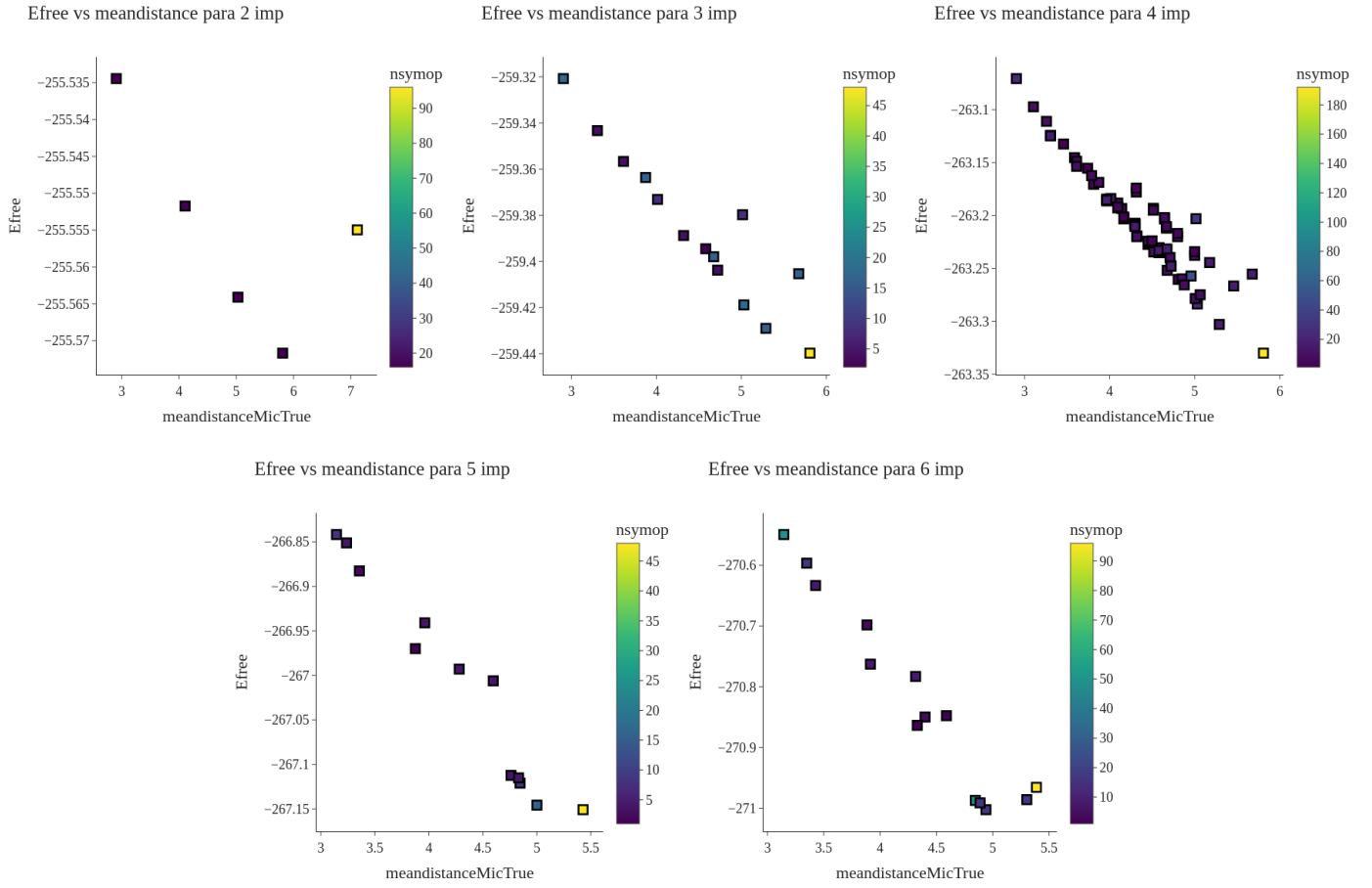


Figura D.1: E_0 vs $meandistance$, color por número de operaciones de simetría ($nsymop$)

D.1.2. Número de simetrías de rotación

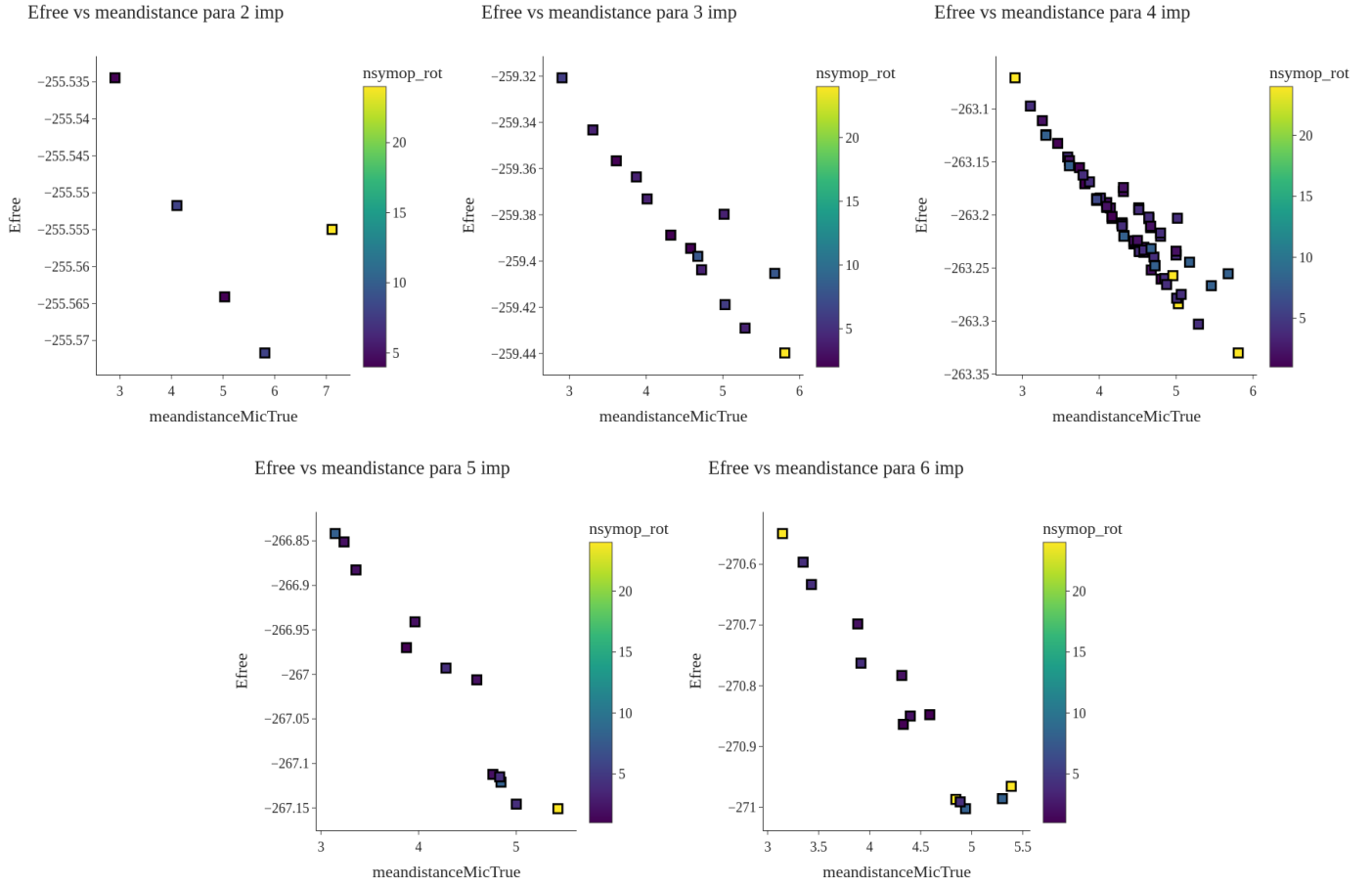


Figura D.2: E_0 vs meandistance , color por número de rotaciones de simetría (nsymop_rot)

D.1.3. Desviación estándar del descriptor meandistance

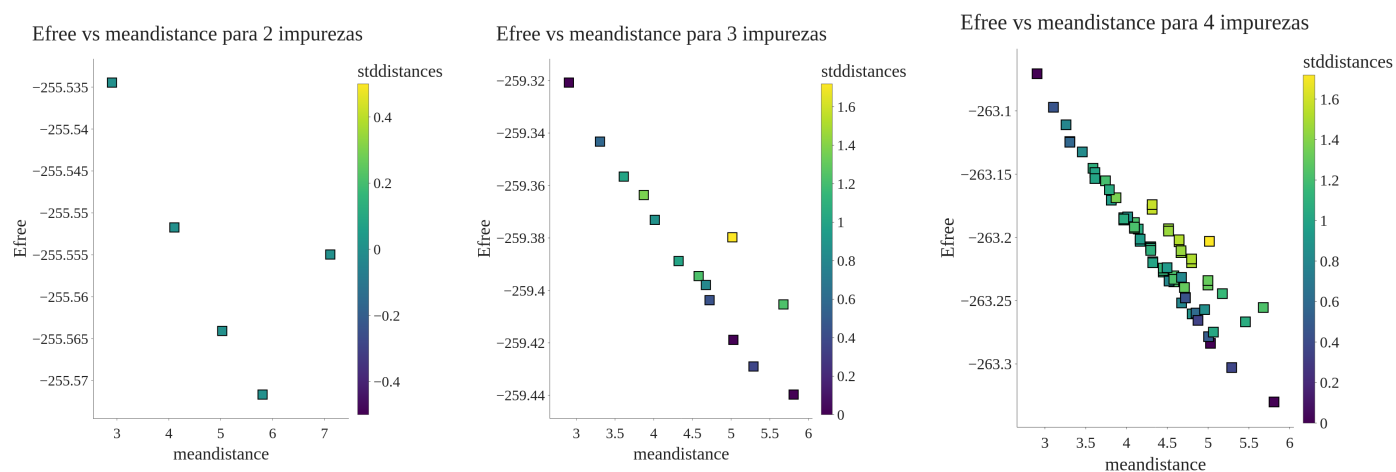


Figura D.3: E_{free} vs $\text{meandistance}_{\text{MicTrue}}$ para 2, 3 y impurezas con color para la desviación estándar

D.1.4. Tipo

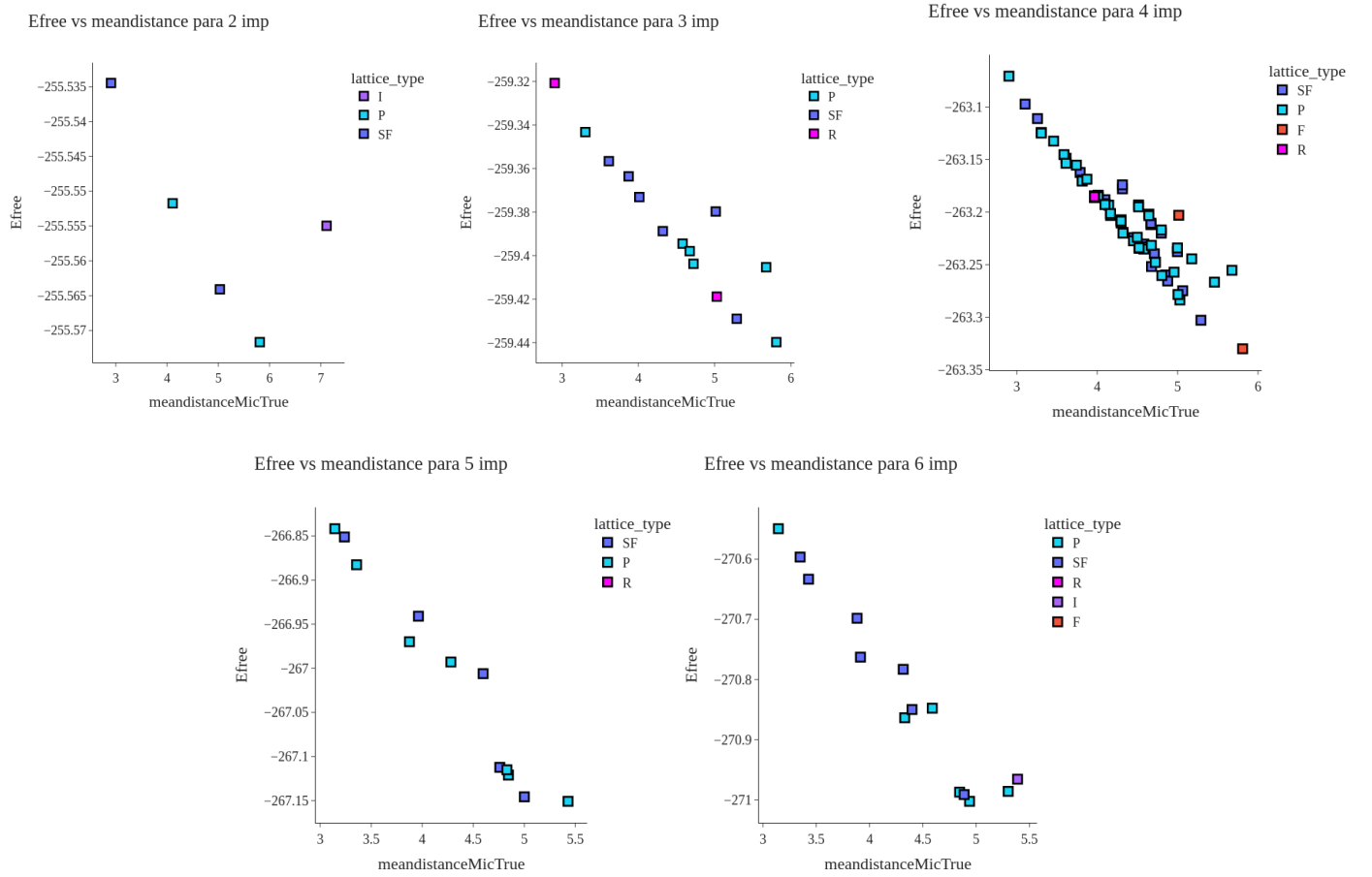


Figura D.4: ‘P’ primitive, ‘I’ body centering, ‘F’ face centering, ‘SF’ single face centering y ‘R’ rhombohedral centering

D.1.5. Matriz de Correlación

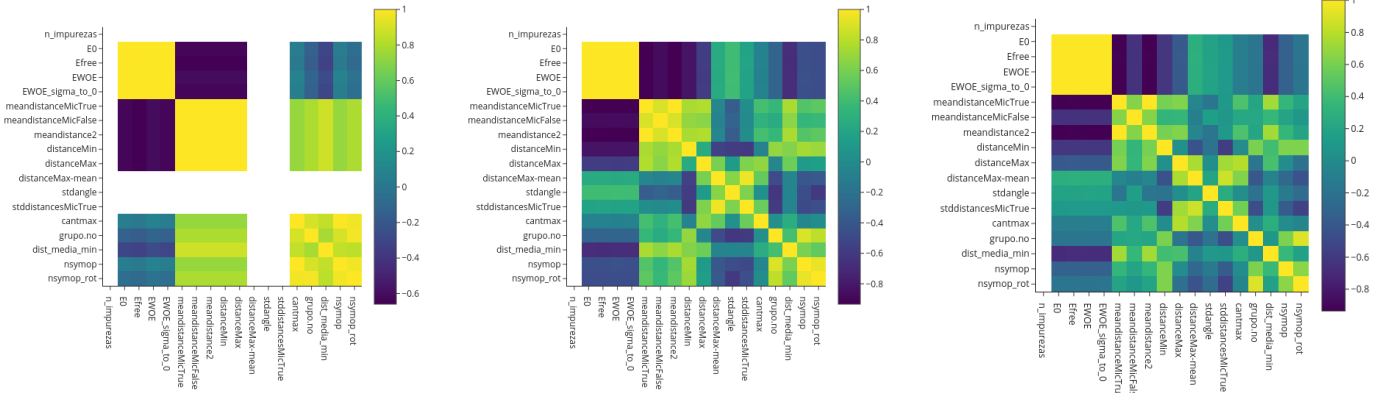


Figura D.5: Matriz de correlación para 2, 3 y 4 impurezas

Bibliografía

- [1] Ankit Agrawal y Alok Choudhary. «Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science». En: *APL Materials* 4 (mayo de 2016), pág. 053208. DOI: [10.1063/1.4946894](https://doi.org/10.1063/1.4946894).
- [2] C. Aguilar y col. «Evolution of synthesis of FCC nanocrystalline solid solution and amorphous phase in the Ti-Ta based alloy by high milling energy». En: *Journal of Alloys and Compounds* 854 (2021), pág. 155980. ISSN: 0925-8388. DOI: <https://doi.org/10.1016/j.jallcom.2020.155980>. URL: <https://www.sciencedirect.com/science/article/pii/S0925838820323446>.
- [3] C. Aguilar y col. «Improving mechanical properties and antibacterial response of alpha-beta ternary Ti-Ta alloy foams for biomedical uses». En: *Journal of Materials Research and Technology* 24 (2023), págs. 8735-8753. ISSN: 2238-7854. DOI: <https://doi.org/10.1016/j.jmrt.2023.05.115>. URL: <https://www.sciencedirect.com/science/article/pii/S2238785423010797>.
- [4] DeepMind. *DeepMind*. Available online. [Online; accessed 21-January-2024]. 2024. URL: <https://www.deepmind.com>.
- [5] Jurgen Hafner. «Ab-initio simulations of materials using VASP: Density-functional theory and beyond». En: *Journal of computational chemistry* 29.13 (2008), págs. 2044-2078.
- [6] Charles R. Harris y col. «Array programming with NumPy». En: *Nature* 585.7825 (sep. de 2020), págs. 357-362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL: <https://doi.org/10.1038/s41586-020-2649-2>.
- [7] Douglas Rayner Hartree. «The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods». En: *Mathematical Proceedings of the Cambridge Philosophical Society* 24 (1928), págs. 89-110. URL: <https://api.semanticscholar.org/CorpusID:122077124>.

-
- [8] P. Hohenberg y W. Kohn. «Inhomogeneous Electron Gas». En: *Phys. Rev.* 136 (3B nov. de 1964), B864-B871. DOI: [10.1103/PhysRev.136.B864](https://doi.org/10.1103/PhysRev.136.B864). URL: <https://link.aps.org/doi/10.1103/PhysRev.136.B864>.
- [9] W. Kohn y L. J. Sham. «Self-Consistent Equations Including Exchange and Correlation Effects». En: *Phys. Rev.* 140 (4A nov. de 1965), A1133-A1138. DOI: [10.1103/PhysRev.140.A1133](https://doi.org/10.1103/PhysRev.140.A1133). URL: <https://link.aps.org/doi/10.1103/PhysRev.140.A1133>.
- [10] G. Kresse y Jürgen Furthmüller. «Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set». En: *Physical Review B* 54 (oct. de 1996), pág. 11169. DOI: [10.1103/PhysRevB.54.11169](https://doi.org/10.1103/PhysRevB.54.11169).
- [11] Ask Hjorth Larsen y col. «The atomic simulation environment—a Python library for working with atoms». En: *Journal of Physics: Condensed Matter* 29.27 (2017), pág. 273002. URL: <http://stacks.iop.org/0953-8984/29/i=27/a=273002>.
- [12] Ji-Chun Lian y col. «Algorithm for generating irreducible site-occupancy configurations». En: *Phys. Rev. B* 102 (13 oct. de 2020), pág. 134209. DOI: [10.1103/PhysRevB.102.134209](https://doi.org/10.1103/PhysRevB.102.134209). URL: <https://link.aps.org/doi/10.1103/PhysRevB.102.134209>.
- [13] Ji-Chun Lian y col. «Algorithm for generating irreducible site-occupancy configurations». En: *Phys. Rev. B* 102 (13 oct. de 2020), pág. 134209. DOI: [10.1103/PhysRevB.102.134209](https://doi.org/10.1103/PhysRevB.102.134209). URL: <https://link.aps.org/doi/10.1103/PhysRevB.102.134209>.
- [14] Xiujuan Liu y col. «Material machine learning for alloys: Applications, challenges and perspectives». En: *Journal of Alloys and Compounds* 921 (2022), pág. 165984. ISSN: 0925-8388. DOI: <https://doi.org/10.1016/j.jallcom.2022.165984>. URL: <https://www.sciencedirect.com/science/article/pii/S0925838822023751>.
- [15] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004. DOI: [10.1017/CB09780511805769](https://doi.org/10.1017/CB09780511805769).
- [16] Amil Merchant y col. «Scaling deep learning for materials discovery». En: *Nature* 624.7990 (2023), págs. 80-85. ISSN: 1476-4687. DOI: [10.1038/s41586-023-06735-9](https://doi.org/10.1038/s41586-023-06735-9). URL: <https://doi.org/10.1038/s41586-023-06735-9>.

- [17] Ulrich Müller. *Symmetry Relationships between Crystal Structures: Applications of Crystallographic Group Theory in Crystal Chemistry*. Oxford University Press, abr. de 2013. ISBN: 9780199669950. DOI: [10.1093/acprof:oso/9780199669950.001.0001](https://doi.org/10.1093/acprof:oso/9780199669950.001.0001). URL: <https://doi.org/10.1093/acprof:oso/9780199669950.001.0001>.
- [18] Kaoru Ohno, Keivan Esfarjani y Yoshiyuki Kawazoe. «Computational Materials Science: From Ab Initio to Monte Carlo Methods». En: (ene. de 1999).
- [19] Kirill Okhotnikov, Thibault Charpentier y Sylvian Cadars. «Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals». En: *Journal of Cheminformatics* 8.17 (mar. de 2016). DOI: [10.1186/s13321-016-0129-3](https://doi.org/10.1186/s13321-016-0129-3).
- [20] John Perdew y col. «Prescription for the Design and Selection of Density Functional Approximations: More Constraint Satisfaction with Fewer Fits». En: *The Journal of chemical physics* 123 (sep. de 2005), pág. 62201. DOI: [10.1063/1.1904565](https://doi.org/10.1063/1.1904565).
- [21] Gabriel R. Schleder y col. «From DFT to Machine Learning: recent approaches to Materials Science – a review». En: *J. Phys. Mater.* (feb. de 2019). ISSN: 2515-7639. DOI: [10.1088/2515-7639/ab084b](https://doi.org/10.1088/2515-7639/ab084b). URL: <http://iopscience.iop.org/article/10.1088/2515-7639/ab084b>.
- [22] J. C. Slater. «The Theory of Complex Spectra». En: *Phys. Rev.* 34 (10 nov. de 1929), págs. 1293-1322. DOI: [10.1103/PhysRev.34.1293](https://doi.org/10.1103/PhysRev.34.1293). URL: <https://link.aps.org/doi/10.1103/PhysRev.34.1293>.
- [23] SOD. 2023. URL: <https://github.com/gcmt-group/sod>.
- [24] The pandas development team. *pandas-dev/pandas: Pandas*. Ver. latest. Feb. de 2020. DOI: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134). URL: <https://doi.org/10.5281/zenodo.3509134>.
- [25] I T Todorov y col. «Simulation of mineral solid solutions at zero and high pressure using lattice statics, lattice dynamics and Monte Carlo methods». En: *Journal of Physics: Condensed Matter* 16.27 (jun. de 2004), S2751. DOI: [10.1088/0953-8984/16/27/011](https://doi.org/10.1088/0953-8984/16/27/011). URL: <https://dx.doi.org/10.1088/0953-8984/16/27/011>.
- [26] Tomasz Adam Wesolowski y Jacques Weber. «Kohn-Sham equations with constrained electron density: an iterative evaluation of the ground-state electron density of interacting molecules». En: *Chemical Physics Letters* 248.1 (1996), págs. 71-76. ISSN: 0009-2614. DOI: [https://doi.org/10.1016/0009-2614\(96\)00071-1](https://doi.org/10.1016/0009-2614(96)00071-1).

1016/0009-2614(95)01281-8. URL: <https://www.sciencedirect.com/science/article/pii/S0009261495012818>.