VeBaMoS:

Un Código de Similaridad Molecular Basado en Vectores

Manual de Usuario

Camilo Prada Latorre
Jhon Enrique Zapata-Rivera, Ph.D
Director de Tesis
Gian Pietro Miscione, Ph.D
Co-director de Tesis

Diciembre 2020

Índice general

Capítulo 1

Introducción a VeBaMoS

1.1. Fundamentos de VeBaMoS

1.1.1. Fundamento Químico

La similaridad es un campo de estudio central en la química teórica moderna puesto que permite relacionar algunas propiedades de los compuestos químicos. Por ejemplo, la clasificación de la acidez o la basicidad de un compuesto químico en el marco de la teoría de Brønsted-Lowry requiere la comparación con otro compuesto de referencia -por definición el agua- y por lo tanto se puede asociar a la similaridad molecular.[2] Sin embargo, vale la pena notar que la similaridad es un concepto subjetivo, depende tanto de la manera en la que es definida como del tipo de comparación que se utilice. Diferentes químicos definirán la similaridad en términos distintos, de acuerdo a su interés particular, mientras que un químico orgánico podría definir la similaridad en términos de la reactividad de las moléculas, un químico cuántico utilizaría la densidad electrónica.[2] La similaridad molecular no solamente proporciona la magnitud en la que se asemejan las moléculas, también se puede asociar a mediciones de distintas técnicas propias de la química, lo que la convierte en una herramienta útil para resolver problemas químicos de interés. Por ejemplo, si nos basamos en la idea química de que *moléculas similares tienen actividades similares*, podemos usar la similaridad molecular para buscar candidatos de fármacos que tengan actividades similares a los fármacos activos o moléculas naturales activas ya conocidos. Esta misma hipótesis se puede extrapolar a otras áreas de la química, como la analítica, en la que se pueden buscar moléculas con señales similares a las de un compuesto de interés. Estos argumentos motivaron la construcción de un método de similaridad robusto, versátil y rápido que conduzca a resultados fiables en diversas aplicaciones. Nuestra apuesta es `VeBaMoS` (*Vector Based Molecular Similarity*), un código de Similaridad Molecular Basado en Vectores.

`VeBaMoS` es un programa que permite evaluar la similaridad entre dos o más moléculas basándose en la representación vectorial de algunas de sus características electrónicas y topológicas. La idea es utilizar descriptores moleculares (propiedades químicas, características geométricas, datos espectroscópicos, etc.) para construir un vector que pueda ser comparado -en términos de magnitud y dirección- con otros vectores dando como resultado información de la similaridad molecular. El procedimiento matemático se basa en álgebra lineal simple; se utilizan diferentes descriptores moleculares para definir la base del espacio vectorial y se toman como componentes de los vectores las magnitudes de cada descriptor molecular (ver sección 1.1.2). Por consiguiente, más allá de los métodos computacionales que se utilicen para su implementación, esta filosofía permite la evaluación de la similaridad molecular usando un razonamiento netamente químico. A fin de cuentas, los descriptores moleculares que se utilizan dan razón de las propiedades geométricas y electrónicas de las moléculas.

1.1.2. Fundamento Matemático

El código se basa en que se puede representar una molécula a partir de un *set* de parámetros específicos que son función de su geometría y estructura electrónica, por lo tanto, si una molécula es similar a otra, estas deberían tener *sets* similares. Para evaluar la similaridad se representa cada molécula mediante un vector (en la base de un espacio vectorial definido por el *set* de parámetros específicos) y se define un índice de similaridad que permita comparar los vectores en magnitud y dirección. Finalmente, se puede evaluar una propiedad química desconocida de alguna de las moléculas (por ejemplo, actividad biológica) teniendo en cuenta que moléculas similares, tanto electrónica como topológicamente, deben tener propiedades similares. A continuación describimos algunos detalles conceptuales.

En primer lugar es necesario definir un vector $\vec{X}$ como una combinación lineal de la forma:

$$\vec{X} = \sum_{i=1}^{n} x_i \vec{a}_i = x_1\vec{a}_1 + x_2\vec{a}_2 + \cdots + x_n\vec{a}_n \tag{1.1}$$

Es decir, que se puede expresar cualquier vector en la base de un espacio vectorial $\{\vec{a_i}\}$. Esta base corresponde a las dimensiones, esto es, el número de parámetros que se pueden utilizar para describir una molécula (el *set*). En nuestro caso, los parámetros deben incluir información estructural y electrónica. Ahora bien, si se expresan dos vectores $\vec{X}$ y $\vec{Y}$ de la forma:

$$\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \quad \vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{1.2}$$

estos se pueden comparar principalmente por sus dos propiedades, magnitud y dirección. Así, para determinar la similaridad entre los dos vectores se puede calcular: 1) la diferencia de magnitudes (α) a partir de la norma L2 de cada vector y 2) la diferencia de dirección (β) a partir el ángulo entre ambos vectores. En este orden de ideas, si se define la norma L2 de un vector $||\vec{X}||$ como:

$$||\vec{X}|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{1.3}$$

quedan definidas las magnitudes con las que se van a comparar los vectores. Por consiguiente, los índices α y β quedan de la siguiente forma:

$$\alpha = ||\vec{X}|| - ||\vec{Y}|| \tag{1.4}$$

$$\beta = \vec{X} \measuredangle \vec{Y} = arccos(\frac{\vec{X} \cdot \vec{Y}}{||\vec{X}||||\vec{Y}||}) \tag{1.5}$$

Estos dos índices por si solos pueden tomar cualquier valor en el rango $[-\infty, \infty]$ para α y en $[0, \pi]$ para β, lo que dificulta su comparación en un conjunto discreto de vectores. Sin embargo, se ha utilizado una función de similaridad $\sigma(z)$, dependiente de estos dos índices, que toma valores entre 0 y 1. Se define $\sigma(z)$ como:

$$\sigma(z) = \frac{1}{e^z} \tag{1.6}$$

en donde $z = |\alpha| + |\beta|$. De esta manera, se consigue un índice de similaridad normalizado σ que permite comparar fácilmente los vectores y realizar un análisis más sencillo. Cuando $\sigma \to 0$ significa que hay un bajo grado de similaridad (los vectores tienen direcciones opuestas y/o difieren en gran cantidad en su magnitud). En contraste, si $\sigma \to 1$ indica que hay un alto grado de similaridad (los vectores se asemejan en magnitud y dirección).

1.2. Características de VeBaMoS

El código de `VeBaMoS` fue escrito en Python utilizando las librerías `NumPy, os, SciPy, pandas y Matplotlib`. Se programó utilizando un paradigma de programación orientado a objetos en donde se calcula cada uno de los descriptores químicos que componen la base del espacio vectorial para representar las moléculas. Parte de la información usada para definir los vectores moleculares es extraída de un cálculo mecanicocuántico utilizando `ORCA`[7,8] y otra parte que se basa en información topológica tomada de las estructuras optimizadas para calcular otros descriptores químicos de importancia como el Volumen, el Área superficial o la Dureza, Suavidad y Electrofilicidad globales. Los descriptores químicos calculados por `VeBaMoS` son:

- momento dipolar
- energía de dispersión
- máxima longitud molecular
- máxima longitud perpendicular a la máxima longitud molecular
- volumen molecular
- área superficial
- momento cuadrupolar
- polarizabilidad

- energía de ionización
- afinidad electrónica
- gap HOMO-LUMO
- dureza global
- suavidad global
- potencial químico
- electrofilicidad global

`VeBaMoS` está pensado para funcionar en un ambiente de Linux; los usuarios más avanzados podrán revisar el código fuente y manipularlo para añadir o alterar funciones del programa que permitan obtener resultados de interés o compatibilidad con otros códigos.

1.3. Contenido del Manual

1.3.1. ¿Cómo está ordenado el manual?

Este manual tiene como objetivo capacitar al usuario en el uso de `VeBaMoS` con fines de investigación, es decir, que identifique el tipo de problemas que se pueden resolver haciendo uso de este programa y que pueda interpretar los resultados obtenidos de la simulación. Por consiguiente, el manual se ha separado en 6 partes que se describen a continuación:

1. **Introducción a `VeBaMoS`:** Se describe de forma general los fundamentos del método similaridad y algunas de sus características de programación e implementación. Es la sección ideal para tener un acercamiento al programa y encontrar de alguna información de interés: cómo citarlo, agradecimientos y los aspectos éticos.
2. **Guía de instalación:** Especifica las instrucciones de instalación, pre-requisitos del sistema para poder ejecutar `VeBaMoS` y la secuencia de pasos para utilizar el código de forma eficiente. Se incluye cómo adquirir el programa.
3. **Guía de uso rápida:** Describe la información necesaria para ejecutar el programa y describe la información que se plasma en el *output* del mismo. Es una sección técnica que se recomienda como referencia para usuarios que quieren empezar a utilizar el programa y requieran un conocimiento básico.
4. **Ejemplo básico:** Se realiza un estudio de la similaridad de un grupo de moléculas simples. Se muestra un paso a paso de cómo utilizar el programa y la forma en la que se deben analizar los resultados.

5. **Tutorial basado en un problema:** Se realiza la investigación de un problema real, por lo que se ahondará en la manera en la que se ha pensado la similaridad como herramienta para predecir propiedades químicas.

6. La **Documentación detallada:** Se explican los detalles técnicos del funcionamiento del código, es decir, cómo se forman los vectores y cómo se operan para obtener resultados fiables. Esta sección permite a los usuarios experimentados profundizar en detalles técnicos del programa para hacer un análisis más rigurosos o acoples con otros códigos.

1.3.2. Uso del Manual

Este manual está pensado para su uso como texto de aprendizaje de usuarios que quieran aplicar este programa de similaridad en un problema puntual. También, se han incluido apartes detallados como referencia para usuarios avanzados que quieran ajustar el código para resolver un problema complejo apoyándose en la similaridad molecular. Las primeras tres secciones son recomendadas para usuarios primerizos, pues se describen los procedimientos esenciales para utilizar el programa y familiarizarse con el método. Por otro lado, las últimas dos secciones son ideales para un usuario que desee profundizar en el funcionamiento el programa. Se recomienda que como primer acercamiento a `VeBaMoS` se lea la primera parte de introducción y luego se lean las partes del ejemplo básico y del tutorial basado en un problema. Esto le dará al lector un claro sentido de la manera en la que se debe utilizar el código y de cuál será el mejor curso de acción para resolver un problema de similaridad utilizando `VeBaMoS`. Para los usuarios más avanzados, se recomienda utilizar el índice para encontrar la información necesaria para satisfacer sus inquietudes técnicas.

1.3.3. Notaciones

Cuando sea necesario resaltar alguna información o indicar detalles a los que el lector deba prestar especial atención, se presentará de la siguiente manera:

> Dentro de este tipo de caja se presentará la información relevante para el lector, por lo que se recomienda leer con atención. Estas pueden ser útiles para entender temas específicos del manual y para centrar la atención en resultados o procedimientos que pueden ser de interés para el tema tratado.

Por otro lado, si es necesario comentar una parte específica de los controles en la terminal[1] o hacer una especificación de un archivo del input o el output se utilizará el siguiente formato:

```
Para los vectores 1 y 1, la similaridad es de 1.0.
Para los vectores 1 y 2, la similaridad es de 0.516702292939.
Para los vectores 1 y 3, la similaridad es de 0.603205265628.
Para los vectores 1 y 4, la similaridad es de 0.858634030275.
Para los vectores 1 y 5, la similaridad es de 0.2004908326359.
Para los vectores 1 y 6, la similaridad es de 0.2546090711653.
Para los vectores 1 y 7, la similaridad es de 0.3188586488581.
Para los vectores 1 y 8, la similaridad es de 0.588207246403.
Para los vectores 1 y 9, la similaridad es de 0.637349629099.
Para los vectores 1 y 10, la similaridad es de 0.549528220276.
...
```

Finalmente, para hablar de algún aspecto específico del código,[2] que no sea muy común, se hará utilizando el siguiente formato:

```
217   a = 0
218
219   for i in range(len(coordenadas_atomos)):
220       j = i
221       while (j <= len(coordenadas_atomos)-1):
222
223           sumcuadrados = np.sum((coordenadas_atomos[i]-coordenadas_atomos[j])**2)
224
225           distancia = np.sqrt(sumcuadrados)
226
227           if(a < distancia):
228               a = distancia
229               self.indice1 = i
230               self.indice2 = j
231               print(self.nombre, i, j, a)
232           j = j+1
233   self.d1 = a
```

[1] Una terminal es un intérprete de comandos que se utiliza como una interfaz entre el usuario y el sistema operativo, desde esta se puede hacer uso de funciones del sistema operativo, al igual que correr programas y utilizarlos en interfaz de texto

[2] Este es el código que se utilizó para calcular la distancia más larga entre dos átomos de la molécula. Para esto, se toma la posición de cada uno de los átomos como un vector y se encuentra la magnitud del vector resultante de la resta de las posiciones de dos átomos, esto da como resultado la distancia que hay entre esos dos átomos.

El lector debe tener en cuenta estas notaciones durante su lectura del manual. De esta manera podrá sacar un mayor provecho tanto al manual como al programa en la solución de problemas asociados a la similaridad.

1.4. Agradecimientos.

VeBaMoSes el resultado de un año de trabajo de la mano de Jhon Zapata, a quien agradezco enormemente por ser guía y ayudarme a resolver mis pensamientos cuando no lograba darles una dirección concreta, haciéndolo siempre con amabilidad y paciencia infinitas. A Gian Pietro le agradezco enormemente permitirme acercarme a la química computacional desde el principio de mi pregrado. Ha sido un honor trabajar con ustedes

Sin embargo, este trabajo de grado es el culmen de un camino que empezó mucho antes. Gracias a mi tío Roberto y a mi tío abuelo Alvaro. Por ellos estudié química, gracias por ser maestros mucho antes de entrar a las aulas.

A mis padres, por apoyarme y tratar de entenderme cuando hablo de lo que me apasiona, sé que siempre estarán para mi y espero poder retribuirles y compartir muchos momentos a su lado. Ustedes son el ejemplo de quien quiero ser durante el resto del camino.

A Alejandra, que estuvo a mi lado durante todo el pregrado y con quien quiero compartir muchos años más de amor y alegría. Eres mi apoyo más grande. Espero poder seguir creciendo a tu lado y hacerte sentir orgullosa.

A mis amigos: Antonio, Sebastian, Pablo y Emilio, por estar a mi lado desde que tengo memoria, por los buenos recuerdos y por siempre estar ahí, gracias por entender las veces que no estuve presente mientras trabajaba en este proyecto. Cuentan conmigo para lo que sea, siempre. Sergio, Natalia, Julian, Helena, Miguel, Santiago, Juan David, Juan Pablo y Alejandro, mi experiencia en la universidad hubiera sido muy distinta sin ustedes, gracias por ayudarme a resolver siempre dudas como colegas y por regalarme incontables risas a lo largo de estos años. Benjamin, Isabella, Monte, Jimmy, Maria Antonia, Maria José, Ortíz, Leco, por siempre alegrarme, por las risas y todos los momentos especiales que hemos compartido, les aprecio montones.

Este trabajo lo dedico a todxs ustedes y a quienes se quedaron en el tintero. Gracias por hacer de mi quien soy, lxs llevo en mi corazón por siempre.

1.5. Cómo Citar VeBaMoS.

En caso de utilizar VeBaMoS para una investigación, el programa debe ser citado en la bibliografía de la siguiente manera:

- Prada, C.; Miscione, G.; Zapata-Rivera, J. VeBaMos: a Vector Based Molecular Similarity code; Grupo de Estructura Electrónica Molecular, Universidad de los Andes, 2020.

1.6. Descargo de Responsabilidad.

VeBaMoS es un producto intangible que opera en el marco de la química computacional, no pretende de ninguna manera reproducir de forma exacta la realidad. Por esto mismo, el uso de este modelo por parte de terceros debe someterse a nuevas consideraciones éticas en función de su aplicación. Asimismo, es necesario tener en cuenta un manejo honesto tanto de los datos recopilados como de los resultados derivados de la investigación.

Para el uso de este código es necesario tener acceso a ORCA v.4.2.1. Debemos dejar claro que ORCA[7,8] no tiene ninguna asociación con nuestro trabajo y los créditos de este programa van a Frank Neese y todo su equipo de trabajo.

Además, se debe resaltar el apoyo del centro de cómputo de alto rendimiento de la Universidad de los Andes para llevar a cabo los cálculos cuánticos.

Capítulo 2

Guía De Instalación.

2.1. Instalación de VeBaMoS

2.1.1. Prerrequisitos

Esta versión alfa de VeBaMoS es un script de Python que se ejecuta mediante una terminal de linux o mediante una interfaz gráfica de usuario muy simple. Se deben cumplir los siguientes prerrequisitos:

- Sistema operativo linux (cualquier distribución).
- Instalación de Python 3.
- Instalar los paquetes NumPy, SciPy, Matplotlib y Pandas
- Acceso a los archivos de salida (outputs) y geometrías en formato `.xyz` de simulaciones con ORCA v. 4.2.1.[7,8]

2.1.2. Paso a paso

A continuación describimos la instalación de la versión alfa de VeBaMoS, asumiendo que no se ha instalado previamente ninguno de los paquetes necesarios.

Descarga de VeBaMoS

Actualmente VeBaMoS se ha desarrollado en una versión alfa, es decir, no es actualmente una versión para distribución. Para adquirir el programa es necesario solicitarlo vía correo electrónico a `c.prada@uniandes.edu.co` con el asunto `Adquirir VeBaMoS` y una breve reseña del objeto de su aplicación.

Instalando Python

Antes de instalar Python, se deben actualizar algunos paquetes e instalar los prerrequisitos. Para ello se corren los siguientes comandos:

```
sudo apt update
sudo apt install software-properties-common
```

luego, se añade un repositorio llamado **Deadsnake**, que contiene paquetes más actualizados que los repositorios de la mayoría de distribuciones de Linux:

```
sudo add-apt-repository ppa:deadsnakes/ppa
```

ahora si, empezamos la instalación de Python con el comando:

```
sudo apt install python3.8
```

y, finalmente, se puede verificar la versión de Python instalada usando el comando:

```
python -- version
```

Instalando los Paquetes Adicionales

Para instalar las librerías de Python necesarias para el funcionamiento de `VeBaMoS`, es suficiente utilizar el siguiente comando:

```
sudo apt-get install python-numpy python-scipy python-matplotlib
ipython ipython-notebook python-pandas python-sympy python-nose
```

Hasta aquí, Python y las librerías necesarias para correr `VeBaMoS` están instalados en el computador. Ahora falta configurar el código para su uso.

Instalando `VeBaMoS`

Para instalar `VeBaMoS` basta con guardar el archivo `VeBaMoS.py` en la carpeta de preferencia del usuario, se puede aconsejar `~/VeBaMos`, pero cualquier otra dirección es válida. Después, se debe abrir el archivo `.bashrc` ubicado en la carpeta personal y se debe pegar la siguiente linea:

```
alias vebamos='python ~/VeBaMoS/VeBaMoS.py'
```

Ahora, solo basta con ejecutar el comando `vebamos` en la terminal para que corra el programa.

Capítulo 3

Guía Rápida De VeBaMoS

3.1. Input y Output

3.1.1. Input

El input de VeBaMoS, en lugar de un archivo de texto plano convencional, es un directorio que contiene los outputs de ORCA y las coordenadas de las geometrías moleculares en formato .xyz. En la actual implementación es necesario que cada output de ORCA incluya el resultado de los cálculos de la polarizabilidad, el momento dipolar y el momento cuadrupolar (ver el uso del módulo elprop en el manual de ORCA). Según el caso de estudio las geometrías de las moléculas pueden ser el resultado de un cálculo de optimización de la geometría o las obtenidas mediante técnicas de rayos X. Los archivos de output y geometrías deben ser nombrados como moleculaX.out y moleculaX.xyz, respectivamente, donde X varía entre 1 y N, siendo N el número de moléculas bajo estudio. Por ejemplo, si se tienen 5 moléculas, es necesario que la carpeta de inputs contenga:

```
molecula1.out   molecula1.xyz   molecula2.out   molecula2.xyz
molecula3.out   molecula3.xyz   molecula4.out   molecula4.xyz
molecula5.out   molecula5.xyz
```

Tenga en cuenta la necesidad de hacer un correcto registro de la asignación de números a cada una de sus moléculas de para lograr un análisis de resultados confiable.

3.1.2. Output

Al finalizar la ejecución de VeBaMoS se generan como outputs diferentes archivos con los resultados de la simulación. El usuario solo debe especificar la ruta de la carpeta donde desea guardar los archivos (esta puede ser la misma carpeta de los inputs). Dentro de esta carpeta se guardarán, entre otros, 2 archivos de output relevantes: VectoresFull.csv y Output.txt.

VectoresFull.csv

Este archivo registra los vectores que han sido construidos por el programa.

Estos vectores tienen la información completa de la molécula, por lo que es útil para analizar cuales son los descriptores moleculares son más relevantes en cada vector (o molécula), y entender mejor las relaciones que guardan las diferentes moléculas.

Sin embargo, estos no son los vectores que se operan para calcular la similaridad, pues antes de esto hay un proceso de normalización que se describe con más detalle en la documentación detallada.

Este archivo es una herramienta fundamental para que el análisis de los resultados esté basado en un pensamiento químico. A continuación se muestra como se ven los datos en una hoja de cálculo:

Tabla 3.1: VectoresFull.csv

dipolarx	dipolary	dipolarz	magnitud dipolar	$E_{Dispersion}$	d1	⋯
0,667	1,649	-1,537	2,35	-0,031	10,048	⋯
-0,010	2,075	-0,323	2,10	-0,032	10,528	⋯
0,808	-1,077	-1,678	2,15	-0,028	9,095	⋯
0,545	1,760	-1,455	2,35	-0,038	10,910	⋯
1,608	0,960	2,287	2,96	-0,029	8,803	⋯
-0,076	1,947	1,678	2,57	-0,035	10,054	⋯
1,154	1,610	1,223	2,33	-0,032	10,071	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

La información que provee este archivo le permite al usuario tener un mayor entendimiento de los descriptores químicos que tienen injerencia en la similaridad. Por ejemplo, si este archivo indica que la similaridad depende fuertemente del volumen y el área superficial, no será un resultado útil para relacionarla con una propiedad como la acidez de las moléculas. Por el contrario, si la similaridad depende de las propiedades electrónicas, podría utilizar para clasificar y predecir la acidez.

Output.txt

Este archivo es el output principal del programa, es donde quedan registrados los valores de la similaridad entre las moléculas en un formato de texto plano. El archivo se verá algo así:

```
Para los vectores 1 y 1, la similaridad es de 1.0.
Para los vectores 1 y 2, la similaridad es de 0.5167022929392568.
Para los vectores 1 y 3, la similaridad es de 0.6032052656287629.
Para los vectores 1 y 4, la similaridad es de 0.8586340302758964.
Para los vectores 1 y 5, la similaridad es de 0.20049083263598383.
Para los vectores 1 y 6, la similaridad es de 0.25460907116534454.
Para los vectores 1 y 7, la similaridad es de 0.31885864885817883.
Para los vectores 1 y 8, la similaridad es de 0.5882072464066943.
Para los vectores 1 y 9, la similaridad es de 0.6373496290972729.
Para los vectores 1 y 10, la similaridad es de 0.5495282202762138.
Para los vectores 1 y 11, la similaridad es de 0.1251151395234791.
...
```

Los valores de similaridad estarán entre 0 y 1, siendo 0 moléculas completamente diferentes y 1 moléculas idénticas. En los capítulos 4 y 5 se puede observar las manera correcta de interpretar los resultados.

3.2. Cómo Utilizar el Programa

3.2.1. GUI

El programa es sencillo de utilizar y depende de la interfaz escogida. La versión GUI tiene este aspecto:

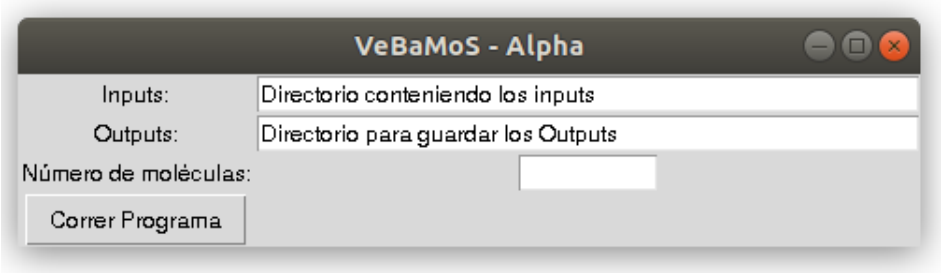


Figura 3.1: GUI de VeBaMoS

al escribir el comando `vebamos` se desplegará la GUI. En las cajas de `Inputs` y `Outputs` se deben escribir las rutas de las carpetas en donde estén guardados los inputs y en donde se quieran guardar los archivos de output. Los inputs deben tener las especificaciones antes mencionadas. Después, es necesario especificar el número de moléculas que hay en el set. El código tomará las moléculas del 1 al N, siendo N el número especificado. Si se requiere excluir algunas moléculas se aconseja crear una nueva carpeta y copiar los archivos con una numeración nueva. Solo queda presionar el botón `Correr Programa` para que el programa ejecute los cálculos.

3.2.2. Línea de Comandos en Terminal.

Para la versión de linea de comandos en una terminal de linux, se ejecuta el comando `vebamos` y el programa pedirá que escriba en la terminal la misma información que se indicó previamente en la versión GUI. Luego se ejecuta presionando `ENTER`, como se ve en la figura 3.2.

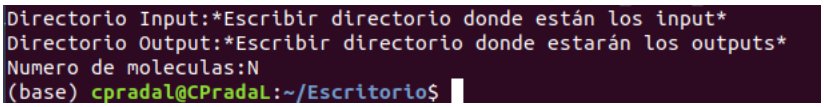


Figura 3.2: Versión terminal de VeBaMoS

Cuando se suministra la información, el programa empezará a ejecutarse y, al finalizar, generará los outputs.

El programa lee información de los archivos de input para poder efectuar los cálculos necesarios, por esto, deben cumplir con las especificaciones que se describen en el aparte 3.1. También, es necesario que además del formato, los outputs de `ORCA` tengan toda la información de los cálculos electrónicos, esto garantiza el buen funcionamiento del programa. Cuando haya un fallo en el código, y no sea evidente el porqué, se debe revisar el cumplimiento estas especificaciones.

Capítulo 4

Ejemplo Básico.

4.1. Introducción.

La intención de este capítulo es acercarse tanto a la manera en la que se realiza un estudio de similaridad, como la forma en la que podemos analizar los outputs que resultan de la simulación con `VeBaMoS`. Para esto, se tomará como ejemplo básico el estudio de la similaridad en 4 moléculas pequeñas, que resultarán conocidas para cualquier usuario con formación básica en química: H_2O, NH_2^-, NH_3 y H_2S. Aunque el uso de `VeBaMoS` está orientado al uso de la similaridad para la clasificación de una propiedad química específica, en este caso no trataremos de resolver ninguna pregunta puntual sobre estas moléculas, más bien, se hará una comparación entre las moléculas para ilustrar el funcionamiento del código.

Antes de comparar moléculas utilizando un método de similaridad, debemos tener un contexto químico que permita que nuestro análisis sea óptimo. Por consiguiente, a continuación se presenta información de interés sobre las moléculas previamente citadas:

- **H_2O:** Es una molécula polar con una alta carga parcial en el átomo de oxígeno. Este compuesto se encuentra en estado líquido a temperatura ambiente gracias a su capacidad de hacer puentes de hidrógeno y a su carácter polar que aumenta las fuerzas de Van Der Waals. Tiene una alta capacidad calorífica y alto punto de ebullición y fusión comparados con otras moléculas similares.

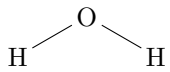


Figura 4.1: Estructura Agua.

- **NH_2^-:** El ión amino (`IUPAC: Azanide`) es estructuralmente similar al agua, pero electrónicamente distinto. No es un compuesto altamente estudiado por ser sumamente inestable, lo que se debe a la carga formal negativa que tiene sobre el átomo de nitrógeno central. Sin embargo, al igual que el oxígeno del agua, el nitrógeno es electronegativo y puede soportar la carga negativa.

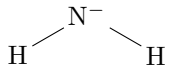


Figura 4.2: Estructura NH_2^-.

- **NH_3:** El amoniaco es una sustancia altamente estudiada y con aplicaciones relevantes en la industria. Al igual que el agua, forma puentes de hidrógeno entre sus moléculas actuando como donor de electrónes. Los puntos de fusión y ebullición no son tan altos como los del agua por lo que a condiciones estándar es un gas. Su estructura y la electronegatividad del nitrógeno hacen que sea una molécula polar.

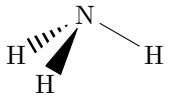


Figura 4.3: Estructura amoniaco.

- **H_2S:** El Sulfuro de hidrógeno es una molécula polar por la diferencia en electronegatividad del azufre con respecto al hidrógeno. Sin embargo, sus hidrógenos son malos aceptores de electrones y por lo tanto no forma puentes de hidrógeno, lo que explica sus bajos puntos de fusión y ebullición. En medio acuoso este compuesto libera protones por lo que tiene un carácter ácido.

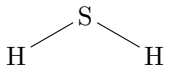


Figura 4.4: Estructura Sulfuro de hidrógeno.

4.2. Metodología del Cálculo de Similaridad.

En primer lugar, se construyeron las geometrías de las moléculas a estudiar. En nuestro caso, se utilizó la librería de Python `RDkit` para generar las estructuras desde un código SMILES, sin embargo, se puede utilizar cualquier otro método.

> Se pueden construir las geometrías de la manera que se desee, utilizando paquetes como `Avogadro`,[3] `Gaussview` o utilizando estructuras generadas por algún algoritmo, lo realmente importante es que estén en formato `.xyz` y que estén alineadas en el espacio. Por ejemplo, en este caso, debido a que todas las moléculas son polares, se alinean de tal forma que los átomos más electronegativos estén apuntando hacia el mismo lugar y que en la medida de lo posible, los hidrógenos estén en el mismo eje.

Una vez guardadas las moléculas con el nombre de archivo `moleculaX.inp.xyz` (con X del 1 al 4), se construyeron los inputs de `ORCA` para la optimización de geometría y cálculo de propiedades. Como ya se mencionó en la parte 3 es necesario calcular la polarizabilidad, el momento dipolar y el momento cuadrupolar. Para esto se utilizó el modulo `elprop` de `ORCA`. A continuación se provee un ejemplo de los inputs de `ORCA` usados en este ejemplo:

```
# Optimization Calculation
```

```
! RKS OPT PBE0 D3BJ RIJDX def2-TZVP def2/J NormalPrint TightSCF SlowConv
Grid5 NoFinalGrid CPCM(water)


 %pal nprocs 12
 end
 ! AnFreq

 %scf
 MaxIter 500
 end

 %elprop
 Dipole true
 Quadrupole True
 Polar 1 # analytic polarizability through CP-SCF
 end

 %cpcm
 epsilon 80.4
 refrac  1.33
 end

* xyzfile 0 1 molecula1.inp.xyz
```

Seguido, se efectuaron los cálculos mecanicocuánticos en `ORCA` en el cluster *Magnus* del centro HPC de la Universidad de los Andes. Se tomaron los archivos `moleculaX.xyz` y los outputs `moleculaX.out` resultantes de `ORCA` y se guardaron en la carpeta de input para `VeBaMoS`. Después de esto, se ejecutó `VeBaMoS`para obtener la información de similaridad.

4.3. Descripción del Output.

El archivo `output.txt` contiene los valores del índice de similaridad entre las moléculas. Esta es la información más relevante para efectuar el análisis de la similaridad molecular. Los resultados se presentan en la tabla 4.1. Por otro lado, el output `VectoresFull.csv` contiene los valores de los descriptores químicos utilizados para construir los vectores. Este archivo será útil en el análisis de los resultados ya que da cuenta de los descriptores que causan diferencias de similaridad.

	1 (H_2O)	2 (NH_2^-)	3 (NH_3)	4 (H_2S)
1 (H_2O)	1.0	0.6316	0.3587	0.5539
2 (NH_2^-)		1.0	0.4055	0.6856
3 (NH_3)			1.0	0.4593
4 (H_2S)				1.0

Tabla 4.1: Resultados de similaridad obtenidos para el ejemplo básico.

Los resultados presentados en la tabla 4.1 permiten ver rápidamente las relaciones de similaridad entre estas 4 moléculas. Teniendo en cuenta que es una matriz simétrica solo se presenta la mitad de esta.

4.4. Análisis de Resultados.

Para poder analizar las similaridades, se debe tomar una de las moléculas como referencia y evaluar qué tan semejante le son las demás. Es decir, no es posible comparar la similaridad entre el agua y el amoniaco contra la similaridad entre el NH_2^- y el H_2S, esto no sería consistente. Si se toma como referencia al agua, que es probablemente la molécula más familiar para el lector, se puede ver que la molécula más similar es de hecho el NH_2^-. Esto puede sorprender al lector puesto que se debe pensar que al ser una molécula inestable sería demasiado distinta electrónicamente, sin embargo, es la molécula más parecida estructuralmente al agua y comparte estrechas relaciones electrónicas. Por ejemplo, el momento dipolar del agua y del ión NH_2^- es similar (ver Tabla 4.2):

	x	y	z	magnitud
H_{20}	0,00077	-0,96081	0	0,96081
NH_2^-	0,0007	-0,91055	1E-05	0,91055
NH_3	-0,00748	0,00118	-0,83927	0,83931
H_2S	0,00039	-0,61652	0	0,61652

Tabla 4.2: Momentos Dipolares en `VectoresFull.csv`

Aunque dista del agua en otros indicadores, como las energías de los orbitales HOMO y LUMO, las diferencias no son igual de significativas.La molécula con la segunda mayor relación de similaridad con el agua es con el Sulfuro de hidrógeno. Este sigue compartiendo una estructura similar a la del agua, sin embargo, se debe recordar que el azufre es el elemento menos electronegativo de entre todos los centros moleculares. Por esto, es menos similar que el NH_2^-, no solo difiere electrónicamente sino también estructuralmente. Ahora bien, la mayor diferencia es entre el agua y el amoniaco, esto se debe en gran parte a las diferencias estructurales entre esta molécula y las demás debido a tener 3 hidrógenos en vez de 2. Esta

diferencia estructural no solamente afecta sus propiedades electrónicas y topológicas sino que dificulta el alineamiento de la molécula. Las diferencias estructurales dejan de manifiesto la necesidad de alinear las moléculas. Esas diferencias también se reflejan en la estructura electrónica, `ORCA` calcula las propiedades electrónicas dependiendo de la orientación de los átomos en los ejes del espacio cartesiano. Al tener una geometría distinta, las orientaciones son distintas, lo que acaba dando como resultado un menor índice de similaridad.

> Tenga presente que las medidas de similaridad de `VeBaMoS` son dependientes del número de moléculas que se estudien. Esto quiere decir, que el índice de similaridad obtenido es función del número de moléculas. Al aumentar el número de moléculas explorado los resultados cambian. Una consecuencia de esto es que la similaridad molecular se centrará en propiedades específicas del espacio químico estudiado.

Al utilizar 4 moléculas en las que 3 comparten una geometría similar, la cuarta molécula que se aleja de esta norma será penalizada. Es por esto que el programa espera que los investigadores utilicen un razonamiento químico al escoger las moléculas que se compararán, al igual que a la hora de analizar de los experimentos realizados. Esto se podría evitar con un grupo más homogéneo de moléculas.

Se debe resaltar que las moléculas aquí estudiadas son bastante similares entre sí, por lo que este ejercicio solo pretende proporcionar un acercamiento al uso de este código. Por esto mismo, la siguiente sección está dedicada a un tutorial en el que se estudian moléculas más complejas a la vez que se explica la manera que se deben interpretar resultados orientándose a responder preguntas químicas.

Capítulo 5

Tutorial Basado en un Ejemplo

5.1. Introducción.

La *Isocitrato deshidrogenasa* (IDH) es una enzima clave en el ciclo de Krebs, importante para la consecución de energía en la gran mayoría de organismos. Esta se encarga de catalizar la descarboxilación oxidativa del ácido isocitrico a ácido α-cetoglutárico. Las IDH se encuentran en el citoplasma (IDH1) y la mitocondria (IDH2 y 3). Sin embargo, sucede que en las células de varios tipos de cáncer, como gliomas, sarcomas e incluso algunos tipos de leucemia, las IDH1 y IDH2 tienden a estar mutadas, lo que al parecer es una complicación que está relacionada con los mecanismos del cáncer. El 90 % de las células de cáncer que tienen mutaciones mostraron la mutación R132H.[5] Esto convierte la proteína mutada en un blanco molecular de gran interés para su estudio, puesto que al ser una mutación regular en el cáncer (y que no está en las células normales) se puede buscar un inhibidor que afecte a las IDH(R132H) pero no a la IDH(Wild Type *WT*).

Se encontró que los compuestos de 1-hidroxipiridin-2-ona (Figura 5.1) son buenos inhibidores de las IDH(R132H), más no tienen un mayor efecto en la IDH(WT). En un artículo reciente se reportan estas estructuras y se hace un estudio detallado de estas mismas, su interacción con la proteína y de la actividad que estas tienen con el target molecular.[5]

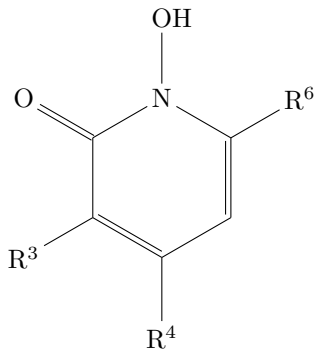


Figura 5.1: Andamiaje molecular de las moléculas estudiadas.

Ahora bien, es de interés saber como se comporta la actividad de estas moléculas en relación con la similaridad molecular para poder observar si es posible hacer uso de `VeBaMoS`como una herramienta que se pueda utilizar en *cribado virtual*. Esto permite mostrar lo robusto que es el programa para resolver problemas moleculares, además de permitir entender mejor la manera en la que se debe utilizar y como analizar resultados orientados a un problema.

En el ejemplo básico presentado en el capítulo 4 se hizo un estudio general, mostrando la forma más general de proceder en el uso de este programa. Aquí, en cambio, se detallará la manera en la que se preparó la información y se utilizó `VeBaMoS`, de manera que el lector tenga herramientas para poder hacer uso del código para resolver los problemas que sean

de su interés.

Para el estudio que se plantea, se tomó un set de 10 moléculas con actividades reportadas[5] (Tabla 5.1) y se estimaron sus índices de similaridad para después evaluar como se relacionan con su actividad.

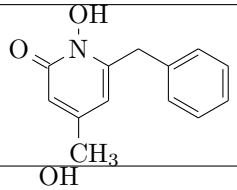
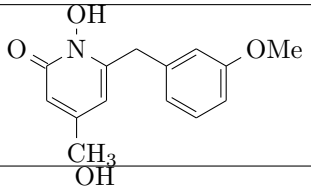
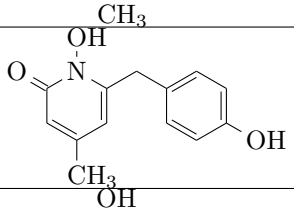
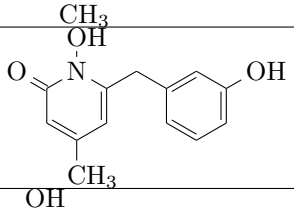
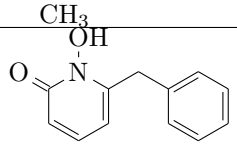
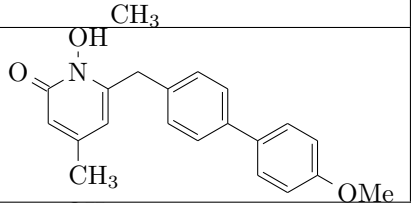
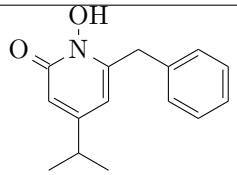
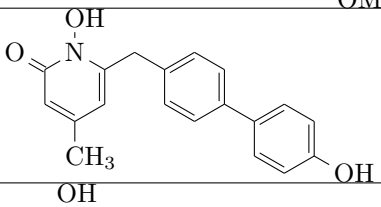
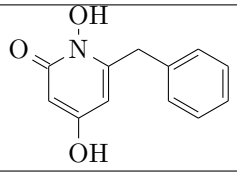
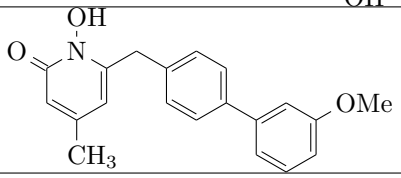
1		6	
2		7	
3		8	y 
4		9	
5		10	

Tabla 5.1: Moléculas a estudiar.

5.2. Metodología.

Para este estudio se generaron las geometrías a optimizar en `ORCA` utilizando los códigos SMILES de las moléculas presentadas en la tabla 5.1 en un archivo `.csv`. También se utilizó la librería RDKit de Python. Las estructuras resultantes en archivos `.xyz` no están alineadas en el espacio de la manera necesaria para el extraer los descriptores electrónicos. Esto se solucionó utilizando `Avogadro`[3] para alinear el grupo de compuestos derivados de la 1-hidroxipiridin-2-ona de y se guardaron con la extensión `.inp.xyz`. Después, se utilizó un

script en Python para generar los inputs de `ORCA` de manera automática. Con estos archivos, se efectuaron los cálculos QM en *Magnus*, el centro de HPC de la universidad de los Andes. De este proceso se obtuvieron los archivos de input para `VeBaMoS`, los archivos `.xyz` y `.out`. Posteriormente, se escribió un script en Python que toma los valores de similaridad de `VeBaMoS` y los valores de actividad reportados para generar unas gráficas de actividad contra similaridad, que permitirán resolver la pregunta por la relación entre similaridad y actividad.

> Esta metodología se utilizó en este estudio en específico, no es de ninguna manera una guía rigurosa para hacer uso de `VeBaMoS`. Se pueden utilizar otros métodos para generar los inputs de `ORCA`, las geometrías y demás procesos de preparación para el uso del código. De la misma manera, el tratamiento de datos mostrado en esta sección es una ilustración de cómo tratarlos para resolver un problema. Sin embargo, es fundamental hacer un análisis previo del problema para plantear una metodología correcta al momento de hacer una investigación.

5.3. Resultados.

El resultado de la simulación proporcionó los índices de similaridad en el output de `VeBaMoS`, `Output.txt` (ver Tabla 5.2). Al ser una matriz simétrica, solo se presentan los valores por encima de la diagonal. Estos se presentan con 4 cifras significativas para una mejor visualización, sin embargo, en el Output se presentan más cifras significativas de ser necesarias.

X	1	2	3	4	5	6	7	8	9	10
1	1.000	0.5988	0.8101	0.7788	0.7126	0.6418	0.4172	0.3202	0.3425	0.2694
2		1.000	0.5966	0.5928	0.5623	0.5879	0.3012	0.2562	0.2540	0.1991
3			1.000	0.6864	0.6697	0.5712	0.4198	0.3085	0.3226	0.2629
4				1.000	0.6995	0.7118	0.4589	0.3802	0.4017	0.3083
5					1.000	0.5679	0.4672	0.3793	0.3921	0.2994
6						1.000	0.3376	0.3187	0.3343	0.2448
7							1.000	0.4327	0.5095	0.5597
8								1.000	0.7520	0.5811
9									1.000	0.6678
10										1.000

Tabla 5.2: Índices de similaridad para el sistema de interés

También se obtuvieron los valores que componen los vectores que se utilizaron en el cálculo de la similaridad, sin embargo, estos datos son numerosos por lo que no se presen-

tarán. La información relevante para el análisis se expondrá a medida que sea utilizada para contextualizar al lector.

Por otro lado, se obtuvieron gráficas de actividad contra similaridad (Figura 5.2). En la dispersión se toma una de las moléculas como referencia y se muestra como se relaciona su actividad y similaridad con las demás moléculas. Este tipo de gráfica podría ser útil en trabajos de ingeniería molecular y diseño de fármacos. Cabe añadir que los resultados de similaridad dependen de la cantidad de moléculas que se estén estudiado, por lo que para este estudio se simuló de manera artificial el uso de 62 moléculas para ajustar los resultados de similaridad obtenidos con estas 10 moléculas a un valor similar al que tendrían si se hubiesen estudiado las 62 moléculas presentadas en el artículo original. Esta no es una aproximación rigurosa pero es suficientemente buena para dar una mejor idea del fenómeno a estudiar y mostrar un análisis más riguroso basado en mejores resultados.

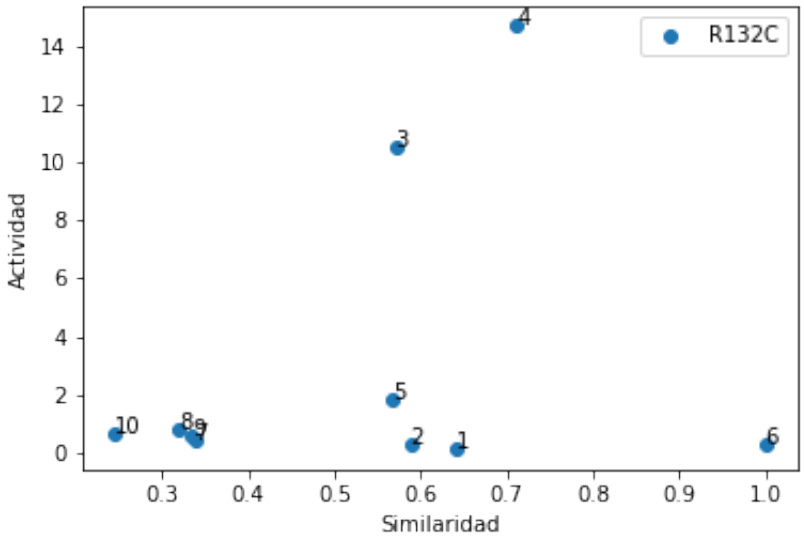


Figura 5.2: Gráfica de Actividad contra Similaridad con la molécula 6 como referencia.

Estas gráficas se generaron con cada molécula como referencia, sin embargo, solo se analizará la gráfica que tiene la molécula 6 como referencia, pues es la gráfica que aporta la información más relevante al análisis.

5.4. Análisis.

En primer lugar, se evaluarán las relaciones de similaridad que guardan las moléculas entre sí. Para esto, lo más sensato podría ser tomar como referencia la molécula más sencilla

de todas, la molécula 3, con las menores modificaciones al andamio molecular de la 1-hidroxipiridin-2-ona (Figura 5.1). De esta manera, podemos ver cual es el efecto que tienen las diferentes modificaciones hechas a la molécula en las relaciones de similaridad. Para esto, se puede hacer una gráfica de linea como la figura 5.3, en la que se muestra la relación de similaridad de cada molécula con respecto a la molécula 3.

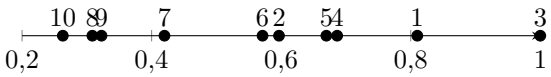


Figura 5.3: Similaridades de las moléculas con respecto a 3.

Basándonos en la información del output, podemos revisar una a una las diferencias que tienen estas moléculas con la molécula 3 y veremos cual es el efecto que tienen estas en el índice de similaridad. Para comenzar, centraremos la atención en la molécula 1, la más similar y avanzaremos hasta la 10, la más diferente.

Para la molécula 1, solo se evidencia una diferencia con la molécula 3 en un metilo en la posición 4 del anillo de la 1-hidroxipiridin-2-ona, por esto, es la molécula más similar, pues el metilo es un grupo relativamente pequeño, que no afecta mucho las propiedades electrónicas. La siguiente molécula es la 4, la segunda más similar, que tiene un isopropilo en la misma posición que está sustituida en la molécula 1. Este es un grupo más voluminoso, por lo que la topología se ve bastante más afectada y la similaridad igual. Algo similar ocurre con la siguiente molécula, la número 5, que tiene un grupo hidroxilo en la misma posición. En este caso, no se trata tanto del volumen de la molécula sino de la electronegatividad del grupo sustituyente que va a generar un cambio en la estructura electrónica de la molécula, afectando la polarizabilidad y los momentos dipolares y cuadrupolares. Estas son las moléculas con sustituciones en la 1-hidroxipiridin-2-ona, las siguientes tienen modificaciones que cambian más respecto al andamio molecular.

Sigue, entonces, en similaridad la molécula número 2, que tiene un grupo hidroxilo en la posición *para* del anillo aromático adyacente a la 1-hidroxipiridin-2-ona. Esto le da una característica polar esta parte de la molécula, lo que en definitiva interfiere con su estructura electrónica y, en términos de actividad, afecta la manera en la que interactúa con los residuos de la proteína. La siguiente molécula en similaridad es la 6, esta tiene un grupo metoxi en la posición meta del anillo aromático. Este sustituyente tiene la característica de ser voluminoso, pero menos reactivo que el grupo hidroxilo. Es por esto, que la siguiente molécula, la número 7, tiene un aún menor índice de similaridad, pues tiene un grupo hidroxilo en la posición *meta* del anillo aromático. De estas últimas tres moléculas, es interesante resaltar la manera en la que la posición del sustituyente en este anillo (y no solamente su

naturaleza) afecta el índice de similaridad.

Las últimas tres moléculas son la 8, la 9 y la 10, moléculas que tienen un anillo aromático extra. Estas moléculas reproducen los resultados de las moléculas 2, 6 y 7, en el sentido que un grupo metoxi es menos perjudicial a la similaridad que un grupo hidróxi y que un sustituyente en la posición *para* es menos perjudicial que un sustituyente en la posicion *meta*.

Estos resultados son muy interesantes, pues muestran que `VeBaMoS` tiene la capacidad de reconocer las diferencias que tienen las moléculas y a partir de estas cuantificar la similaridad que tienen. Además, muestran que los datos numéricos y cuantificables que se utilizan para el cálculo de estos índices se pueden traducir a características cualitativas que pueden ser fácilmente interpretadas usando un pensamiento químico más clásico.

Ahora bien, para revisar la relación entre la actividad y la similaridad, hay que prestar atención a la figura 5.2, en la cual se muestra una gráfica de puntos que relaciona la actividad inhibidora de las moléculas, actuando sobre la proteína mutada R132C, con la similaridad que tienen con respecto a la molécula 6. Se escogió la molécula 6 como estándar pues esta tiene la mejor actividad y por tanto es una buena referencia para este análisis. Esta comparación, sin embargo, es complicada dado que todas las moléculas tienen una buena actividad. los peores valores son de $k_i = 8$ y 9 μM, por lo que todas las moléculas tienen una buena actividad.

Además de unos puntos con valores atípicos (moléculas 3 y 4), se observa una ligera tendencia a que la actividad disminuya (su valor aumente) a medida que disminuye la similaridad de 6. Esto mostraría que la similaridad a la molécula más activa también significa una mejor actividad. Sin embargo, la molécula más similar a la molécula 6 (la molécula 4) tiene la peor actividad de todas las moléculas estudiadas. Esto es problemático por que no corresponde con la hipótesis de que hay una relación directa entre ambas variables.

Sin embargo, es importante remitirse a la naturaleza de la similaridad para entender completamente este fenómeno. Un indicador de similaridad nos habla de la manera en la que las moléculas tienen relaciones estructurales y electrónicas. Entonces, dado que las moléculas activas dependen de una estructura electrónica y topológica específicas, una diferencia alta de una molécula activa no implica necesariamente una menor actividad, pues puede ser diferente en varias partes más tener las características necesarias para tener una actividad, de manera inversa. una molécula con una similaridad alta puede tener una estructura similar a la de la molécula activa pero solo diferenciarse en la posiciones importantes para la actividad.

Teniendo esto en cuenta, `VeBaMoS` puede ser una herramienta importante para hacer estudios de actividad, sin embargo, es imperante hacer un análisis crítico no solo de los índices de similaridad, sino también de las propiedades electrónicas y topológicas descritas en los vectores (`VectoresFull.csv`) para tener un panorama más completo de los resultados y así llegar a conclusiones más robustas e interesantes. Esto significa que a pesar de ser una herramienta útil para el *cribado virtual*, podría generar algunos falsos positivos y falsos negativos.

Capítulo 6

Documentación Detallada.

6.1. Introducción

En esta parte del manual, se hace una descripción detallada de la manera en la que `VeBaMoS` calcula los descriptores químicos que componen el espacio vectorial en el que están representadas las moléculas. También, aborda la manera en la que el programa trata estos mismos vectores para conseguir un índice de similaridad normalizado. La sección está pensada especialmente para los usuarios avanzados que esperan hacer uso del software de manera más educada, profundizando en la manera en la que funciona el código de similaridad molecular. Para esto, se dividirá este trabajo en dos partes, la primera se centrará en la manera que cada uno de los descriptores químicos son calculados u obtenidos, acercando al lector a la manera en la que está escrito el código y brindando un poco de información teórica sobre estos descriptores. La segunda parte describe la manera en la que se obtienen los índices de similaridad a partir de los vectores. Puede pensarse como una profundización de la sección 1.1.2, explicando los ajustes que se hacen en la implementación y como afectan los resultados.

6.2. Descriptores Químicos

6.2.1. Descriptores Topológicos

Distancia Entre Puntos Más Distantes $D1$

El primer descriptor topológico que se calcula es la distancia mayor entre dos átomos. Esta distancia muestra el eje más largo de la molécula por lo que da información sobre el tamaño de la molécula. Esto es relevante especialmente en aplicaciones que impliquen interacciones intermoleculares. Para calcular el valor de esta distancia (desde ahora D_1) se toman las coordenadas de cada átomo como un vector tridimensional desde el origen al átomo por lo que la distancia entre dos átomos es igual a la magnitud del vector resultante de la resta entre dos vectores. Así, teniendo dos vectores $\vec{v_i}$ y $\vec{v_j}$ definidos como:

$$\vec{v_i} = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix}; \quad \vec{v_j} = \begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} \tag{6.1}$$

En donde el vector resultante de la resta de los vectores $\vec{v_i}$ y $\vec{v_j}$ sería:

$$\vec{v_i} - \vec{v_j} = \begin{bmatrix} x_i - x_j \\ y_i - y_j \\ z_i - z_j \end{bmatrix} \tag{6.2}$$

Por tanto, para encontrar la distancia entre puntos más distantes se utiliza la ecuación

6.3 para todos los pares de átomos i, j en la molécula. Sin embargo, dado que la pareja i, j tiene el mismo valor de D1 que j, i solo se calcula la primera para optimizar el desempeño del código.

$$D_1 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{6.3}$$

Esta medida es la medida en la que el resto de descriptores topológicos se basan, por lo que es una primera medida que tiene implicaciones en el resto de descriptores.

Distancia máxima Perpendicular a D1

Esta nueva distancia traza una linea entre los átomos que tienen la distancia D_1 y se calculan las distancias perpendicular entre los demás átomos en la molécula a la linea trazada en D_1. La distancia máxima calculada en esta manera se denomina D_2.

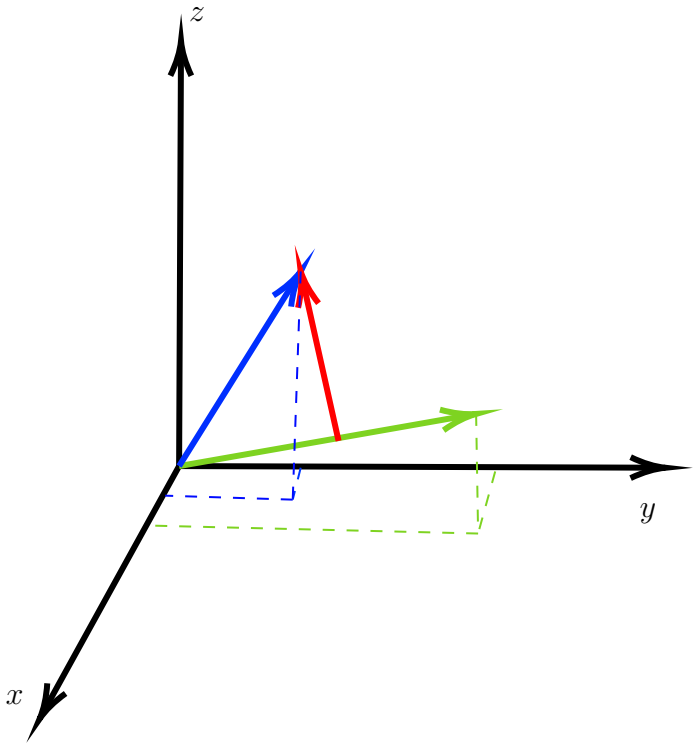


Figura 6.1: Caption

Para calcular esto, se define el vector que va entre el átomo A y el átomo B donde estos átomos son los que están sobre D_1 como $v\vec{_{AB}}$. Además de esto, se debe definir un vector $v\vec{_{Ai}}$ que va del átomo A en D_1 al átomo i al cual se le quiere calcular la distancia D_2. Teniendo estos vectores definidos, se puede encontrar la distancia perpendicular a D_1 con la magnitud del vector resultante de la ecuación 6.4.

$$\vec{v_{D_2}} = \vec{v_{Ai}} - proj_{\vec{v_{AB}}}\vec{v_{Ai}} \tag{6.4}$$

Volumen

El volumen de la molécula no se evalúa de una forma exacta, esto debido a su geometría compleja que depende tanto de la estructura que dan los núcleos como de la geometría de los orbitales electrónicos. Por lo tanto, se utiliza una aproximación basándonos en los valores calculados en D_1 y D_2. Esta aproximación consiste en tomar los valores de D_2 y disponerlos de manera equidistante en un eje, para después hacer un ajuste de una curva a estos puntos (figura 6.2). teniendo la función de esta curva, digamos $f(x)$, se hace uso de los sólidos de rotación para encontrar el volumen de la molécula.

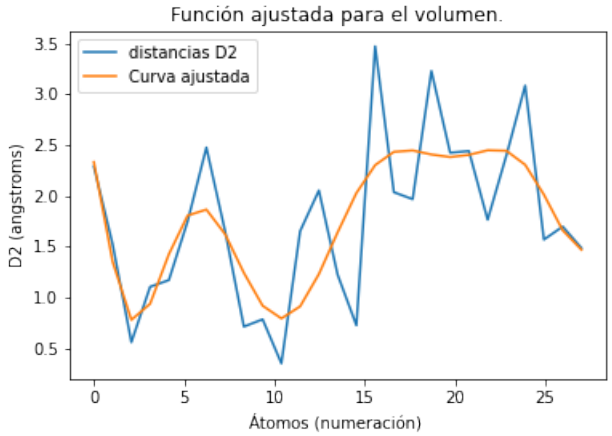


Figura 6.2: Ajuste de una curva a los puntos obtenidos con D_2

Suponiendo que ya se tiene $f(x)$, para obtener el volumen utilizamos la ecuación 6.5.

$$V = \pi \cdot \int_0^N f_{(x)}^2 dx \tag{6.5}$$

en donde el eje sobre el cual se rota es la linea trazada por D_1 y se revisa el volumen que se genera en la curva de rotación. Si bien esta aproximación puede sobre estimar un poco el valor real de volumen de la molécula, es una solución simple para encontrarlo haciendo uso de la información estructural que ya se tiene disponible.

Área superficial.

El área superficial aprovecha también la curva ajustada a los puntos de D_2 que se utilizó para encontrar el volumen, encontrando una superficie de revolución con la ecuación 6.6,

este método es consistente con la manera en la que se encuentra el volumen, encontrando el área superficial de este mismo. Por lo tanto, el área superficial que se encuentra con este método obedece a la misma aproximación que tiene el volumen y no es el área superficial real de la molécula, sin embargo, es una opción simple y útil para dar cuenta de esto. El área superficial es importante para dar cuenta en la similaridad de interacciones que puede tener la molécula, pues se puede pensar esta como el área de posibles puntos de interacción que tiene una molécula al ser expuesta a una situación en donde esto sea parte del problema de estudio. Es por esto que se tiene en cuenta como descriptor topológico de importancia para el programa.

$$A = 2\pi \cdot \int_0^N f_{(x)} \sqrt{1 + \left(\frac{df}{dx}\right)^2} dx \tag{6.6}$$

6.2.2. Descriptores Electrónicos

Polarizabilidad[4]

La polarizabilidad es la tendencia que tiene una distribución de cargas a ser deformada por un campo eléctrico externo. En la química, la distribución de cargas de interés es la nube electrónica de la molécula a estudiar. Esta es una propiedad que influencia la manera en la que una carga o un campo eléctrico externo va a interactuar con la molécula, por ejemplo, un dipolo molecular. Entonces, es una propiedad que da razón de la manera en la que se relacionan las moléculas y se ve influenciada por los átomos que componen la molécula y la posición relativa que ocupan entre ellos. Esta propiedad se toma directamente del output de `ORCA` y no se manipula de ninguna manera. Sin embargo, es importante resaltar que esta propiedad está definida de manera escalar como

$$\alpha = \frac{\delta p}{\delta E} \tag{6.7}$$

Donde α es la polarizabilidad, p es el momento dipolar inducido por el campo y E es el campo eléctrico externo.

Debido a que la polarizabilidad se define como una magnitud unidimensional, esta debe calcularse en las dimensiones del espacio cartesiano para encontrar la polarizabilidad total. Es por esto que en el vector están consignadas las polarizabilidades en las direcciones x, y y z, además de la polarizabilidad isotrópica.

Momento Dipolar[4]

El momento dipolar es una medida de la separación entre las cargas positivas y negativas en la molécula en una configuración de partículas cargadas. Esta muestra las interacciones

entre pares atómicos enlazados y resulta en un vector de momento dipolar total que es una sumatoria de los vectores de momento dipolar individuales de cada enlace. Esta propiedad da cuenta de la polaridad de la molécula y de la manera en la que esta puede interactuar electrónicamente. Dado que es una magnitud vectorial, se incluyen las componentes x, y, z y la magnitud del vector. Este es otro indicador que se lee directamente del output de `ORCA`.

Momento Cuadrupolar[4]

Es el momento asociado a otra de las configuraciones posibles que pueden tener partículas ligadas por fuerzas, en este caso eléctricas. Esta surge normalmente de una expansión del desarrollo multipolar, en donde el primer término (que está asociado al momento dipolar) es muy pequeño o igual a cero. Por lo tanto, esta propiedad habla de interacciones interatómicas lejanas, no enlazantes, que pueden ayudar a describir la manera en la que la molécula va interactuar con otras debido a su densidad electrónica. Este es un descriptor químico que se toma directamente del output de `ORCA`. Al ser un tensor, se toman solamente las componentes del vector isotrópico y su magnitud.

Energía de Dispersión[4]

La energía de dispersión es la energía asociada a las fuerzas de dispersión de London. Estas fuerzas son la causa de interacciones intermoleculares débiles debidas a dipolos momentáneos. Así, este descriptor químico es una medida de las interacciones de la molécula a larga distancia. Este es el último descriptor electrónico que es leído directamente del output de `ORCA`, es una magnitud escalar, por lo que solo se incluye este único término en el vector de representación molecular.

Gap Homo-Lumo

Es la brecha energética entre los orbitales HOMO y LUMO, de gran importancia, pues un movimiento electrónico en este intervalo es la manera más posible de que suceda transición electrónica, una excitación. Esta brecha depende de las energías de los orbitales moleculares, lo que la convierte en un descriptor químico interesante para tener en cuenta en la similaridad molecular. Esta se obtiene al restar la energía del HOMO a la energía del LUMO. Las energías de cada orbital molecular está registrada en el output de `ORCA`, por lo que para obtener la brecha energética solamente se debe hacer la resta de estas magnitud.

$$\Delta E_{HOMO-LUMO} = E_{LUMO} - E_{HOMO} \tag{6.8}$$

Teorema de Koopmans:[1,9] A pesar de ser un teorema establecido para la teoría de Hartree-Fock, tiene un teorema análogo en la teoría DFT, llamado teorema DFT-Koopmans, este es el teorema relevante para los siguientes descriptores químicos. Este señala una relación entre las energías de Kohn-Sham de los orbitales HOMO y LUMO con la primer energía de ionización y la afinidad electrónica de la siguiente manera:

$$I = -\varepsilon_H; AE = -\varepsilon_L \tag{6.10}$$

Esto permite el cálculo de diversos descriptores químicos en función de estos parámetros.

Energía de Ionización[1,4]

La energía de ionización es la energía necesaria para separar un electrón en su estado basal del átomo (o molécula). La primera energía de ionización se refiere específicamente a la energía para separar un electrón del orbital HOMO de la molécula. Esta propiedad electrónica da información de la manera en la que la molécula se comportará en un entorno cargado de manera positiva, por lo que es de interés para interacciones con receptores moleculares como proteínas. Se registra la energía negativa del HOMO en el vector, de acuerdo al teorema DFT-Koopmans.

Afinidad Electrónica[1,4]

La afinidad electrónica está definida como la energía liberada cuando un átomo neutro captura un electrón y forma un ion negativo, Esto corresponde a tomar un electrón y ubicarlo en el orbital disponible de menor energía, esto normalmente será el orbital LUMO, aunque algunas veces será el orbital HOMO. Al contrario que la energía de ionización, permite dar una idea de la manera en la que la molécula se comporta en un ambiente con cargas negativas. Se registra entonces el negativo de la energía del LUMO en el vector, de acuerdo al teorema DFT-Koopmans.

Potencial Químico[1,6]

El potencial químico se ha definido como el negativo de la electronegatividad de Pauling. Por ende, podemos calcular el potencial químico con:

$$\mu = \frac{(\epsilon_H + \epsilon_L)}{2} \tag{6.11}$$

Recordando que el potencial químico en la teoría DFT se puede definir como:

$$\mu = \left(\frac{\delta E}{\delta N}\right)_{v(r)} \tag{6.12}$$

Este descriptor molecular da cuenta de la manera en la que la energía cambia con el número de electrones presentes, por lo que indica la reactividad de la molécula, específicamente, habla de la capacidad de la molécula para donar electrones. Al vector se le consigna el valor obtenido por la ecuación 6.11.

Dureza Global[1,6]

Es una medida de la cercanía entre electrones y núcleos, lo que indica la resistencia del sistema para transferir cargas. Este descriptor nos da razón de la reactividad de la molécula y se describe en el marco de la química cuántica como:

$$\eta = \frac{(\epsilon_H - \epsilon_L)}{2} \tag{6.13}$$

Por lo que se consigna el valor de η en el vector después de calcularla con la ecuación anterior.

Suavidad Global[1,6]

La suavidad es el contrario de la dureza. Este descriptor es un indicativo la lejanía entre los electrónes y los nucleos de los átomos con poca electronegatividad relativa, por lo que da indicio de la capacidad que tiene la molécula para cambiar su densidad electrónica, lo que es, a su vez, evidencia de la reactividad química. Se calcula de la siguiente manera:

$$S = \frac{1}{2\eta} \tag{6.14}$$

Y se usa el valor de S en los vectores moleculares.

Electrofilicidad Global[1,6]

La electrofilicidad global indica la estabilidad de la molécula a la hora de aceptar electrones, por lo que puede ser pensado como el contrario al potencial químico, dando razón de la capacidad reactiva en presencia de una molécula cargada. Esta se calcula usando dos descriptores químicos anteriormente discutidos, la dureza global y el potencial químico, de la siguiente manera:

$$W = \frac{\mu^2}{2\eta} \tag{6.15}$$

El valor de W también es incluido en el vector molecular.

6.3. Manipulación de los Vectores Moleculares.

Podemos definir el vector molecular del siguiente modo:

$$\vec{X} = \sum_{i=1}^{n} x_i \vec{a}_i = x_1\vec{a}_1 + x_2\vec{a}_2 + \cdots + x_n\vec{a}_n \tag{6.16}$$

Es decir, se puede expresar este vector vector en la base del espacio vectorial $\{\vec{a_i}\}$ - los descriptores moleculares-. Ahora bien, si se expresan dos vectores $\vec{X}$ y $\vec{Y}$, correspondientes a dos moléculas X y Y, de la forma:

$$\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \quad \vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{6.17}$$

Estos vectores tienen un problema para comparar de manera sensible moléculas extremas (muy grandes, cargadas, muy reactivas) con otras moléculas cuyos atributos sean contrarios (por ejemplo, el agua con el taxol). Para resolver este problema, se normalizan los descriptores químicos construyendo los vectores con la fracción correspondiente al total de cada descriptor químico. Esto es, para cada descriptor químico, se consigna en los vectores el valor:

$$x_{i,n} = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{6.18}$$

obteniendo así los vectores:

$$\vec{X_n} = \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ \vdots \\ x_{n,n} \end{bmatrix}; \quad \vec{Y_n} = \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ \vdots \\ y_{n,n} \end{bmatrix} \tag{6.19}$$

Estos son los vectores que inicialmente se van a comparar para encontrar el índice de similaridad, comparando sus dos propiedades, magnitud y dirección.

Así, para determinar la similaridad entre los dos vectores se calculam: 1) la diferencia de magnitudes (α) a partir de la norma de cada vector y 2) la diferencia de dirección (β) a partir el ángulo entre ambos vectores. En este orden de ideas, si se define la norma de un vector $||\vec{X}||$ como:

$$||\vec{X}|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{6.20}$$

quedan definidas las magnitudes con las que se van a comparar los vectores. Por consiguiente,

los índices α y β se definen la siguiente forma:

$$\alpha = ||\vec{X}|| - ||\vec{Y}|| \tag{6.21}$$

$$\beta = \vec{X} \measuredangle \vec{Y} = arccos(\frac{\vec{X} \cdot \vec{Y}}{||\vec{X}||||\vec{Y}||}) \tag{6.22}$$

Los índices tomar cualquier valor en el rango $[-\infty, \infty]$ para α y en $[0, \pi]$ para β, lo que dificulta su comparación en un conjunto discreto de vectores. Por esto mismo, se ha utilizado una función de similaridad $\sigma(z)$, dependiente de estos dos índices, que toma valores entre 0 y 1. Se define $\sigma(z)$ como:

$$\sigma(z) = \frac{1}{e^z} \tag{6.23}$$

en donde $z = |\alpha| + |\beta|$. Cuando $\sigma \to 0$ significa que hay un bajo grado de similaridad (los vectores tienen direcciones opuestas y/o difieren grandemente en su magnitud). En contraste, si $\sigma \to 1$ indica que hay un alto grado de similaridad (los vectores se asemejan en magnitud y dirección).

Sin embargo, los valores de sigma son dependientes de la cantidad de moléculas utilizadas en el set a estudiar por causa de la normalización de los vectores. Por esta razón, es necesario hacer un ajuste a la función σ que permita ajustar los valores de sigma, para que no todos se acerquen a 1 en el caso de tener muchas moléculas ni a 0 en el caso de tener pocas. Se ajusta entonces con:

$$\sigma(z) = \frac{1}{e^{z/Log_{10}(N)}} \tag{6.24}$$

En donde N es el número de moléculas del set estudiado. Estos son los resultados que se reportan en `Output.txt`

Bibliografía

[1] H. Chermette. Chemical reactivity indexes in density functional theory. *Journal of Computational Chemistry*, 20(1):129–154, 1999.

[2] Peter Gedeck. *Reviews in Computational Chemistry, Volume 10*, volume 3. 1997.

[3] Marcus D. Hanwell, Donald E. Curtis, David C. Lonie, Tim Vandermeerschd, Eva Zurek, and Geoffrey R. Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(8):17, aug 2012.

[4] W.M. Haynes, David R. Lide, and Bruno Thomas J., editors. *CRC Handbook of Chemistry and Physics*. CRC Press, New York, 97th edition, 2017.

[5] Zhen Liu, Yuan Yao, Mari Kogiso, Baisong Zheng, Lisheng Deng, Jihui J. Qiu, Shuo Dong, Hua Lv, James M. Gallo, Xiao Nan Li, and Yongcheng Song. Inhibition of cancer-associated mutant isocitrate dehydrogenases: Synthesis, structure-activity relationship, and selective antitumor activity. *Journal of Medicinal Chemistry*, 57(20):8307–8318, 2014.

[6] Jesús M. López, Adolfo E. Ensuncho, and Juana R. Robles. Descriptores globales y locales de la reactividad para el diseño de nuevos fármacos anticancerosos basados en cisplatino(II). *Quimica Nova*, 36(9):1308–1317, 2013.

[7] Frank Neese. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1):73–78, jan 2012.

[8] Frank Neese. Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1), jan 2018.

[9] Gang Zhang and Charles B. Musgrave. Comparison of DFT methods for molecular orbital eigenvalue calculations. *Journal of Physical Chemistry A*, 111(8):1554–1561, mar 2007.