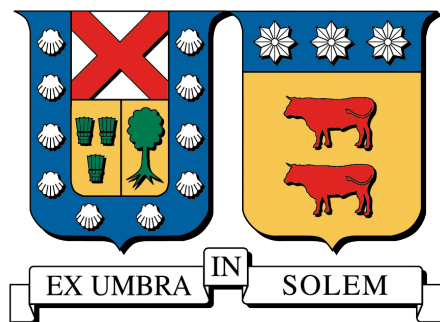


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INGENIERÍA MECÁNICA
VALPARAÍSO - CHILE



**CUANTIFICACIÓN DE INCERTIDUMBRE EN ENERGÍA DE
SOLVATACIÓN DE MOLÉCULAS DEBIDO A POSICIÓN DE
ÁTOMOS**

MIGUEL ANDRÉS GODOY DÍAZ

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MECÁNICO

PROFESOR GUÍA : Ph.D. CHRISTOPHER COOPER V.
PROFESOR CORREFERENTE : Dr. MARCOS VALDEBENITO C.
PROFESOR CORREFERENTE : Ph.D. ALEJANDRO PACHECO S.

ABRIL 2021

A mis padres, por su incondicional amor, su esfuerzo permanente y su presencia excepcional a lo largo de toda mi vida.

A mis hermanos, por ser mis modelos a seguir durante mi vida académica.

A mis amigos de la universidad, por ser los autores de mis mejores recuerdos de esta hermosa etapa.

A mis colegas de CIAC, por apoyarme en mis ideas y hacer del trabajo un lugar entretenido.

A TSL por convertir la pandemia y la vida online en momentos llenos de alegría y compañerismo.

Al profesor Cooper, por su paciencia y disposición durante este trabajo.

A todos ustedes agradezco y dedico esta memoria.

RESUMEN

Las ecuaciones de Poisson y Poisson Boltzmann permiten describir las interacciones electrostáticas de una molécula inserta en una solución iónica. A través del modelo de solvente implícito y el método de elementos de frontera es posible estimar parámetros de interés como la energía de solvatación o energía libre electrostática de una molécula al tener información de la posición y carga de los átomos del soluto. La posición de los átomos es representada empíricamente y tiene un grado de incertidumbre debido a la medición y las vibraciones moleculares que dan origen a propiedades térmicas.

Este trabajo tiene como fin cuantificar la incertidumbre generada por la posición de los átomos del soluto sobre la energía de solvatación de las moléculas. Para esto se implementaron varios modelos de generación aleatoria de posiciones de átomos (agitación) y se calcularon parámetros estadísticos usando el método Monte Carlo. Se reportan los resultados obtenidos por tres modelos de agitación: átomos se mueven aleatoriamente dentro de esferas de radio constante, según su conectividad y dentro de esferas de radio en función de su masa. Los datos muestran que esta metodología funciona para cuantificar incertidumbre y encontrar relaciones entre variables que pueden tener relación con los objetivos de este trabajo.

ABSTRACT

The Poisson and Poisson-Boltzmann equation allow us to describe electrostatic interactions of a molecule inserted in an ionic solution. Through the implicit solvent model and boundary element method it is possible to estimate parameters of interest such as the solvation energy or electrostatic free energy of a molecule by having information of the position and charge of the solute atoms. The position of these atoms is empirically represented and has a degree of uncertainty due to measurement and molecular vibrations that cause thermal properties.

The purpose of this work is to quantify uncertainty generated by the position of the solute atoms in the solvation energy of molecules. For this, several models of random generation of atom positions (shake) were implemented and statistical parameters were calculated using the Monte Carlo method. The results obtained by three shake models are reported: atoms move randomly within spheres of constant radius, according their connectivity, and within spheres of radius according to their mass. The data show that this methodology works to quantify uncertainty and find relationships between variables that may be related to the objectives of this work.

Índice de Contenidos

1. Introducción	1
2. Marco Teórico	3
2.1. Modelo de solvente implícito	3
2.2. Formulación integral de frontera	5
2.3. Método de elementos de frontera	8
2.4. Energía de solvatación	9
3. Metodología	11
3.1. Creación de mallas	12
3.2. Cálculo de energía de solvatación	12
3.3. Método Monte Carlo	14
3.4. Base de datos de FreeSolv	15
4. Agitación esférica radio constante	16
4.1. Metodología y algoritmos	16
4.2. Análisis de resultados	19
5. Agitación con conectividad	21
5.1. Metodología	21
5.2. Análisis de resultados	26
6. Agitación térmica	29
6.1. Longitud térmica promedio y tiempo característico	29
6.2. Algoritmo de agitación térmica	30
6.3. Relación exponencial entre tiempo característico y desviación estándar	31
6.4. Comparación modelos implícito y explícito para energía de solvatación	34
6.5. Comparación de incertidumbre en energía de solvatación	35
6.6. Búsqueda de relaciones y clusters	36
6.6.1. Comparación métodos anteriores	36
6.6.2. Tabla de correlación	37
7. Conclusiones	42
Bibliografía	44

Índice de Tablas

4.1. Valores estadísticos simulación Monte Carlo para energía de solvatación de biomolécula 1PGB	19
5.1. Constante dieléctrica (adimensional) para grupos funcionales presentes dentro de las moléculas candidatas al experimento	26
6.1. Tabla de correlación para simulación con 75 moléculas candidatas	38

Índice de Figuras

2.1. Comparación entre modelos de solvente (extraído de Ramm 2019 [13])	3
2.2. Representación del sondeo de una molécula (extraído de Cooper 2015 y Fernández 2019 [4, 7])	4
2.3. Definiciones de superficies (extraído de Cooper 2015 [4])	4
2.4. Ciclo termodinámico para la energía de solvatación (extraído de Cooper 2020 [3]) . .	10
3.1. Esquema de funcionamiento del programa (extraído de Fernández 2019 [7])	13
4.1. Ejemplo en dos dimensiones de áreas en las que puede ser reasignada la posición de cada átomo de la molécula etano.	17
4.2. Visualización de la incertidumbre a través de cuartiles a distintos radios de agitación para la energía de solvatación de molécula 1PGB	19
4.3. Relación lineal entre la desviación estándar y el radio de agitación para cálculos con método Monte Carlo de la energía de solvatación de molécula 1PGB	20
5.1. Diagrama de modelo de agitación entre átomos	23
5.2. Comparación de valor promedio de energía de solvatación calculada (BEM) con respecto a DM (Moblely)	27
5.3. Promedio de energía de solvatación con BEM vs Mobley, para variación dist. 1 % y ángulo $\pi/9$	28
6.1. Ejemplo en dos dimensiones de las áreas en que pueden ser reubicados los átomos de una molécula de etano	30
6.2. Regresión lineal para tetrafluorometano	31
6.3. Regresión lineal para 1,1,1,2 tetracloroetano	31
6.4. Cambios en μ y σ según la cantidad de pruebas para tetrafluorometano	33
6.5. Cambios en μ y σ según la cantidad de pruebas para 1112_tetracloroetano	33
6.6. Comparación de promedios de energía de solvatación con valores de Mobley 2009 . .	34
6.7. Comparación de incertidumbre con valores de Mobley 2009	35
6.8. Comparación de promedios de energía de solvatación con Mobley 2009, en colores se indican grupos funcionales de moléculas	36
6.9. Comparación de desviación estándar DM y PB. En colores se muestran grupos funcionales	37
6.10. Relación entre α y β con 75 moléculas candidatas. Grupos funcionales resaltados en color.	39
6.11. Relación entre α y β con 75 moléculas candidatas. Promedio de energía de solvatación usando BEM se resalta en barra de color	40
6.12. Dispersión de tiempo característico vs desviación estándar del área de las superficies accesibles por el solvente (SASA). Grupos funcionales resaltados en barra de colores	40
6.13. Dispersión de datos de α y promedio de energía de solvatación μ_{BEM} . Grupos funcionales resaltados en color	41

1 | Introducción

La biofísica computacional permite mediante modelos físicos y matemáticos simular las interacciones electrostáticas de moléculas insertas en un medio salino con gran precisión siendo objeto de estudio actual.

Una técnica de simulación es la dinámica molecular (DM) que permite visualizar el movimiento de las partículas por cortos períodos de tiempo con aplicaciones en biofísica y ciencia de materiales. Un inconveniente de esta técnica es la gran cantidad de cálculos complejos a realizar iterativamente que hace que el modelar sistemas sencillos sea inviable para grandes moléculas en poco tiempo.

En dinámica molecular es posible calcular parámetros de interés como energías libres, estas consideran estructuras estáticas a nivel macro y su valor final estima un promedio de los estados más probables que pueden tomar las configuraciones moleculares. Uno de estos parámetros a tratar en este informe es la energía de solvatación, cuyo sentido físico es el trabajo necesario de traer una molécula desde el infinito hacia una cavidad de solvente.

El problema de la dinámica molecular en biofísica es el costo de modelar todos los átomos asociados, por ello una gran ventaja es considerar el solvente como un medio dieléctrico continuo en lugar de un conjunto de partículas (solvente implícito y explícito respectivamente). Esta técnica entrega valores de energías libres para moléculas insertas (soluto) con una buena aproximación y será la base para nuestro estudio.

Esta simplificación permite presentar la superficie del soluto como superficie de exclusión de solvente (SES). Así como también evitar discretizar todo el volumen del soluto para trabajar con esta superficie o frontera. Mediante el método de elementos de frontera (BEM por sus siglas en inglés) y a través de formulaciones integrales del problema es posible calcular el potencial electrostático en la superficie de exclusión y teóricamente calcular el potencial en cualquier punto de la región interna y externa de la molécula (soluto y solvente propiamente tal).

Para el cálculo de la energía de solvatación utilizando el modelo de solvente implícito es necesario conocer la posición, carga y radio de los átomos. Estos dos últimos datos si bien usan

distintos modelos de campos de fuerza, son más bien empíricos y su incertidumbre ha sido objeto de estudio previamente [18].

El objetivo general de este trabajo es cuantificar la incertidumbre en la energía de solvatación debido a las posiciones de los átomos del soluto, esto lo hacemos con el método Monte Carlo al generar varias configuraciones moleculares con átomos agitados aleatoriamente y calculando parámetros de interés como el promedio y desviación estándar de las energías obtenidas.

Objetivos específicos de este trabajo son la implementación de diferentes modelos de generación aleatoria de posiciones de átomos, la optimización de parámetros de cada modelo usando el método Monte Carlo y por último, comparar los resultados con respecto a la dinámica molecular.

2 | Marco Teórico

2.1. Modelo de solvente implícito

El caso de estudio se centra en cuantificar la incertidumbre al calcular la energía libre electrostática de una molécula inmersa en un medio acuoso (con o sin presencia de sal). En los modelos de representación del medio acuoso, dos de estos serán de nuestro interés, el de solvente explícito y solvente implícito. El primero se encarga de modelar cada uno de los átomos e iones de agua y sal del solvente, computacionalmente implica un gran costo en el cálculo de operaciones complejas que modelan las interacciones y fuerzas atómicas. Por otra parte, el modelo de solvente implícito considera el medio acuoso como un dieléctrico continuo, disminuyendo de gran manera los tiempos de computación. El modelo utilizado es de dos regiones de soluto y solvente interconectadas por una superficie de exclusión Γ .

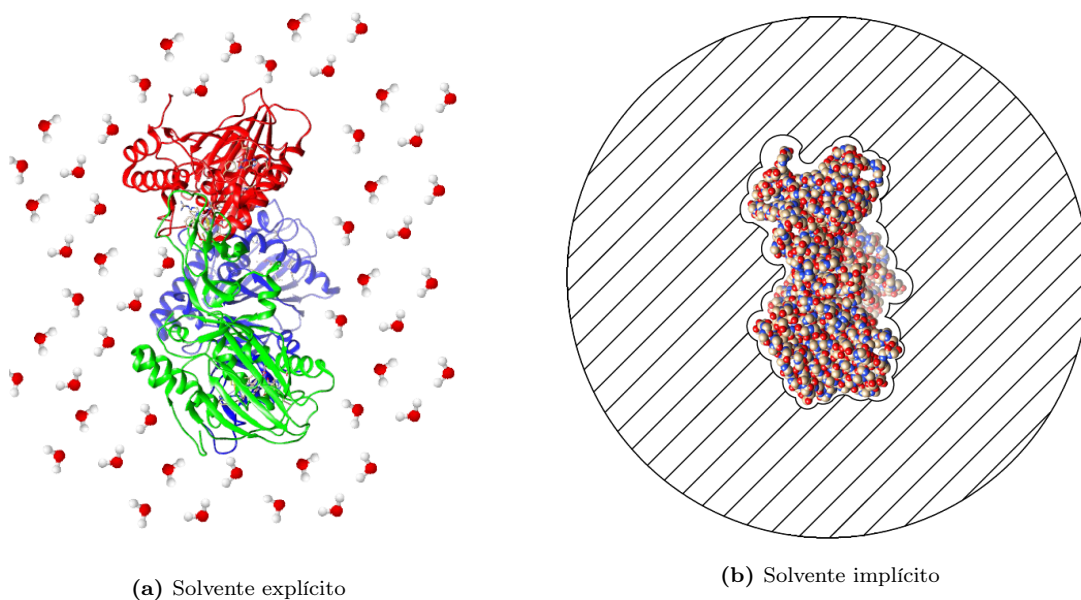


Figura 2.1: Comparación entre modelos de solvente (extraído de Ramm 2019 [13])

Para definir la región de exclusión se utiliza una probeta esférica para sondear el contorno de los radios de van der Waals de los átomos que componen la molécula. En muchos casos se utilizará una esfera de radio 1.4 Å para simular el radio del átomo de oxígeno del agua.

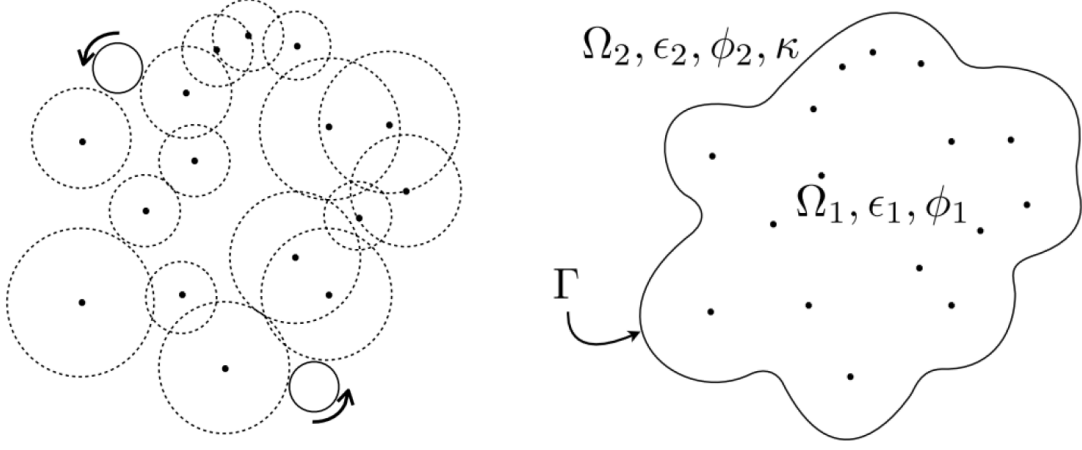


Figura 2.2: Representación del sondeo de una molécula (extraído de Cooper 2015 y Fernández 2019 [4, 7])

Las condiciones de contorno de la interfaz soluto-solvente (Γ) son

$$\phi_1 = \phi_2 \quad (2.1)$$

$$\epsilon_1 \frac{\partial \phi_1}{\partial \mathbf{n}} = \epsilon_2 \frac{\partial \phi_2}{\partial \mathbf{n}} \quad (2.2)$$

Siguiendo esta idea de sondeo con una molécula de agua se desglosan dos superficies de interés: superficie accesible por el solvente (SAS: Solvent Accessible Surface) y la superficie de exclusión de solvente (SES: Solvent Excluded Surface).

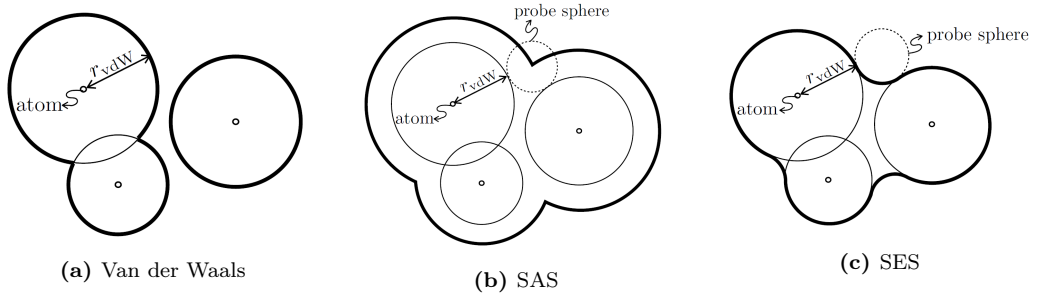


Figura 2.3: Definiciones de superficies (extraído de Cooper 2015 [4])

En la región del soluto Ω_1 se encuentran los átomos que componen la molécula de interés, en donde no pueden entrar moléculas de agua o iones de sal. Se modela como un volumen vacío con cargas puntuales a través de la ecuación de Poisson (2.3), donde ϕ_1 es el potencial electrostático dentro de la región del soluto, ϵ_1 es la constante dieléctrica (o permitividad eléctrica) del medio y $\vec{x}_k$ es la posición del k-ésimo átomo de la molécula.

$$\nabla^2 \phi_1 = - \sum_k \frac{q_k}{\epsilon_1} \delta(\vec{x}_k) \quad (2.3)$$

Para el caso de la región del solvente Ω_2 , modelado como un medio dieléctrico continuo, tiene una distribución de partículas, así la densidad de carga ρ se puede representar como una suma de cargas puntuales en ecuación 2.4.

$$\nabla^2 \phi_2 = -\frac{\rho}{\epsilon} = -\frac{1}{\epsilon} \sum_k q_k c_k(r) = -\frac{1}{\epsilon} q_k c_0 \exp\left(\frac{-\phi q_k}{k_B T}\right) \quad (2.4)$$

con $k_B T$ como el producto entre la constante de Boltzmann y la temperatura absoluta del medio, además de c_0 como el número promedio de especies. Simplificando la ecuación mediante una expansión de Taylor y al ser $\frac{\phi q}{k_B T} \approx 0$ (ecuación 2.5)

$$\exp\left(\frac{-\phi q_k}{k_B T}\right) \approx 1 + \frac{-\phi q_k}{k_B T} \quad (2.5)$$

Dando origen a la ecuación de Poisson-Boltzmann linealizada (2.6), donde κ es el inverso de la longitud de Debye y simboliza el apantallamiento producido por las cargas del solvente.

$$\nabla^2 \phi_2 - \kappa^2 \phi_2 = 0 \quad (2.6)$$

2.2. Formulación integral de frontera

La formulación integral para el problema de Laplace base de nuestro estudio parte de la consideración de un cuerpo de volumen Ω y fronteras Γ , con un potencial u con condiciones de contorno de Dirichlet o Neumann

$$\nabla^2 u = 0 \text{ en } \Omega \quad (2.7)$$

$$u = u_0 \text{ o bien } \frac{\partial u}{\partial n} = \frac{\partial u_0}{\partial n} \text{ en } \Gamma \quad (2.8)$$

Integrando la ecuación 2.7 con una función peso ω se tiene

$$\int_{\Omega} \nabla^2 u \omega \, d\Omega = 0 \quad (2.9)$$

Estas expresiones dependen respectivamente de $u = u(\mathbf{x})$, $\omega = \omega(\mathbf{r}) = \omega(\mathbf{x}, \mathbf{y})$, donde $\mathbf{x}$ e $\mathbf{y}$ son vectores presentes en Ω y Γ respectivamente, además se define $\mathbf{r}$ como $\mathbf{r} = \mathbf{x} - \mathbf{y}$.

Se tienen dos expresiones convenientes al operar utilizando la regla de la cadena

$$\nabla \cdot (\nabla u \omega) = \nabla^2 u \omega + \nabla u \nabla \omega \quad (2.10)$$

$$\nabla \cdot (u \nabla \omega) = u \nabla^2 \omega + \nabla u \nabla \omega \quad (2.11)$$

Reescribiendo (2.9) con la igualdad en (2.10).

$$\int_{\Omega} \nabla \cdot (\nabla u \omega) d\Omega - \int_{\Omega} \nabla u \nabla \omega d\Omega = 0 \quad (2.12)$$

Reescribiendo (2.12) reemplazando según (2.11)

$$\int_{\Omega} \nabla \cdot (\nabla u \omega) d\Omega - \int_{\Omega} \nabla \cdot (u \nabla \omega) d\Omega + \int_{\Omega} u \nabla^2 \omega \, d\Omega = 0 \quad (2.13)$$

Donde ω es función peso que cumple con

$$\nabla^2 \omega = -\delta(\mathbf{x})$$

$$\omega(r) = \begin{cases} \frac{1}{2\pi} \ln(r) & \text{para dos dimensiones} \\ \frac{1}{4\pi r} & \text{para tres dimensiones} \end{cases} \quad (2.14)$$

Usando el teorema de la divergencia, para un volumen Ω cuya frontera es Γ , se tiene que

$$\int_{\Omega} \nabla \cdot \vec{F} \, d\Omega = \int_{\Gamma} \vec{F} \cdot \vec{n} \, d\Gamma \quad (2.15)$$

donde $\vec{n}$ es un vector normal en Γ que apunta hacia afuera del volumen Ω Reescribiendo (2.13) teniendo en cuenta el teorema (2.15)

$$\int_{\Gamma} \nabla u \omega \cdot \vec{n} \, d\Gamma - \int_{\Gamma} u \nabla \omega \cdot \vec{n} \, d\Gamma = - \int_{\Omega} u \nabla^2 \omega \, d\Omega \quad (2.16)$$

Que a su vez equivale a

$$\int_{\Gamma} \nabla u \cdot \vec{n} \, d\Gamma - \int_{\Gamma} u \nabla \omega \cdot \vec{n} \, d\Gamma = \int_{\Omega} u \delta(\mathbf{x}) \, d\Omega$$

Y este a su vez en

$$\int_{\Gamma} \nabla u \cdot \vec{n} \, d\Gamma - \int_{\Gamma} u \nabla \omega \cdot \vec{n} \, d\Gamma = u(\mathbf{x}) \quad (2.17)$$

Esta formulación permite conocer el potencial u y su flujo ∇u en cualquier punto interno del dominio Ω . Para el método numérico, esto implica en primer lugar conocer todos los valores en la frontera Γ del potencial u y su flujo ∇u dependiendo del tipo de condición de borde. Para esto es necesario entonces definir una ecuación para calcular u y ∇u en Γ .

Con el fin de calcular el valor restante (de potencial o flujo) de un punto en particular o placa discretizada no es posible decir $\mathbf{x} = \mathbf{y}$. Cuando $\mathbf{x} \rightarrow \mathbf{y}$ ocurre una indeterminación interesante para nuestro análisis, por lo que se integrará sobre una semiesfera de radio $\epsilon \rightarrow 0$. (Para el caso 2-D el tratamiento es similar, en donde se integra alrededor de una semicircunferencia de radio ϵ)

En el caso del primer término del lado izquierdo de (2.17), al decir $\mathbf{r} = \epsilon = \mathbf{x} - \mathbf{y}$ tiende a 0

$$\lim_{\epsilon \rightarrow 0} \left[\int_{\Gamma_{\epsilon}} \nabla u \cdot \vec{n} \, d\Gamma \right] = \lim_{\epsilon \rightarrow 0} \left[\int_{\Gamma_{\epsilon}} \nabla u \cdot \frac{1}{4\pi\epsilon} d\Gamma \right] = \nabla u \lim_{\epsilon \rightarrow 0} \frac{2\pi\epsilon^2}{4\pi\epsilon} = 0 \quad (2.18)$$

Para el caso de la segunda integral del lado izquierdo de (2.17)

$$\lim_{\epsilon \rightarrow 0} \left[\int_{\Gamma_{\epsilon}} u \nabla \omega \cdot \vec{n} \, d\Gamma \right] = \lim_{\epsilon \rightarrow 0} \left[\int_{\Gamma_{\epsilon}} u \frac{\partial}{\partial \epsilon} \frac{1}{4\pi\epsilon} d\Gamma \right] = \lim_{\epsilon \rightarrow 0} \int_{\Gamma_{\epsilon}} u \cdot \left(-\frac{1}{4\pi\epsilon^2} \right) d\Gamma = -u \lim_{\epsilon \rightarrow 0} \frac{2\pi\epsilon^2}{4\pi\epsilon^2} = -\frac{u}{2} \quad (2.19)$$

Luego, es posible escribir la ecuación general (2.17) cuando $x \rightarrow y$ como

$$\int_{\Gamma} \nabla u \cdot \vec{n} \, d\Gamma - \int_{\Gamma} u \nabla \omega \cdot \vec{n} \, d\Gamma = \frac{u}{2} \quad (2.20)$$

y reescribirla como

$$\frac{u}{2} + \int_{\Gamma} u \nabla \omega \cdot \vec{n} \, d\Gamma - \int_{\Gamma} \nabla u \cdot \vec{n} \, d\Gamma = 0 \quad (2.21)$$

2.3. Método de elementos de frontera

Para resolver la formulación anterior aplicado a nuestro caso de estudio se usan las ecuaciones (2.17) y (2.21) modificando el potencial u por $\phi(\mathbf{x})$ y la función peso ω por la función de Green $G(x, y)$, donde $\mathbf{x}$ es un vector dentro del dominio Ω , así como $\mathbf{y}$ es un vector de la frontera Γ . Las ecuaciones (2.22) y (2.23) son la formulación integral de las ecuaciones (2.3) y (2.6).

$$\phi_1(\mathbf{x}) + \int_{\Gamma} \phi_1(\mathbf{y}) \frac{\partial G_L}{\partial \mathbf{n}}(\mathbf{x}, \mathbf{y}) dy - \int_{\Gamma} \frac{\partial \phi_1}{\partial \mathbf{n}}(\mathbf{y}) G_L(\mathbf{x}, \mathbf{y}) dy = \sum_k \frac{q_k}{4\pi\epsilon_1 |\mathbf{x} - \mathbf{y}|} \quad (2.22)$$

$$\phi_2(\mathbf{x}) - \int_{\Gamma} \phi_2(\mathbf{y}) \frac{\partial G_H}{\partial \mathbf{n}}(\mathbf{x}, \mathbf{y}) dy + \int_{\Gamma} \frac{\partial \phi_2}{\partial \mathbf{n}}(\mathbf{y}) G_H(\mathbf{x}, \mathbf{y}) dy = 0 \quad (2.23)$$

Los signos de las integrales son opuestos debido a la dirección del vector normal al momento de integrar sobre la superficie interior o exterior de la frontera.

Las funciones de Green a utilizar serán dos dependiendo de la ecuación a tratar, G_L para la ecuación de Poisson (o Laplace) y G_H para la ecuación de Poisson-Boltzmann (también conocida como ecuación de Helmholtz o potencial de Yukawa)

$$G_L = \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \quad (2.24)$$

$$G_H = \frac{e^{-\kappa |\mathbf{x} - \mathbf{y}|}}{4\pi |\mathbf{x} - \mathbf{y}|} \quad (2.25)$$

Así, el potencial en la frontera y su derivada pueden ser calculados con el sistema de ecuaciones integrales de frontera que se presenta en (2.26) y (2.27), usando los cambios de variable

$$\begin{aligned} \phi &= \phi_2 = \phi_1 \\ \frac{\partial \phi}{\partial \mathbf{n}} &= \frac{\partial \phi_2}{\partial \mathbf{n}} = \frac{\epsilon_1}{\epsilon_2} \frac{\partial \phi_1}{\partial \mathbf{n}} \\ \epsilon &= \frac{\epsilon_1}{\epsilon_2} \end{aligned}$$

$$\frac{1}{2}\phi(\mathbf{x}) + \int_{\Gamma} \phi(\mathbf{y}) \frac{\partial G_L}{\partial \mathbf{n}}(\mathbf{x}, \mathbf{y}) dy - \int_{\Gamma} \frac{\partial \phi}{\partial \mathbf{n}}(\mathbf{y}) G_L(\mathbf{x}, \mathbf{y}) dy = \sum_k \frac{q_k}{4\pi\epsilon_1 |\mathbf{x} - \mathbf{x}_k|} \quad (2.26)$$

$$\frac{1}{2}\phi(\mathbf{x}) - \int_{\Gamma} \phi(\mathbf{y}) \frac{\partial G_H}{\partial \mathbf{n}}(\mathbf{x}, \mathbf{y}) dy + \epsilon \int_{\Gamma} \frac{\partial \phi}{\partial \mathbf{n}}(\mathbf{y}) G_H(\mathbf{x}, \mathbf{y}) dy = 0 \quad (2.27)$$

Queda así definido el potencial de todo el dominio de la frontera Γ , siendo posible utilizar las ecuaciones (2.22) y (2.23) para calcular el potencial en el dominio del soluto Ω_1 o del solvente Ω_2 .

2.4. Energía de solvatación

El proceso de solvatación o disolución se entiende como la interacción entre el soluto y el solvente, donde se estabilizan las partículas del medio acuoso. Se define como el cambio en la energía libre de Gibbs cuando un ión o molécula es transferida del vacío al solvente [10].

Siguiendo el trabajo de Levy [9], la energía de solvatación puede ser computada siguiendo el ciclo termodinámico de la figura 2.4. Aquí, la energía de solvatación ΔG_{solv} se descompone en dos partes: polar ΔG_{es} y no polar ΔG_{np} .

$$\Delta G_{es} = \Delta G_{charge}^{solv} - \Delta G_{charge}^{vac} \quad (2.28)$$

Por su parte la componente no polar es la suma de la energía requerida para formar la cavidad en el solvente y la interacción dispersiva del soluto y solvente.

$$\Delta G_{np} = \Delta G_{cav} + \Delta G_{disp} \quad (2.29)$$

La energía potencial electrostática de un cuerpo se puede calcular con la ecuación (2.30), para una densidad de carga $\rho(x)$ y una distribución de potencial $\phi(x)$ es posible obtener la energía de solvatación de una molécula integrando sobre el dominio Ω_1 , correspondiente a la región interna de la superficie de exclusión.

$$\Delta G = \frac{1}{2} \int_{\Omega} \phi(x) \rho(x) dx \quad (2.30)$$

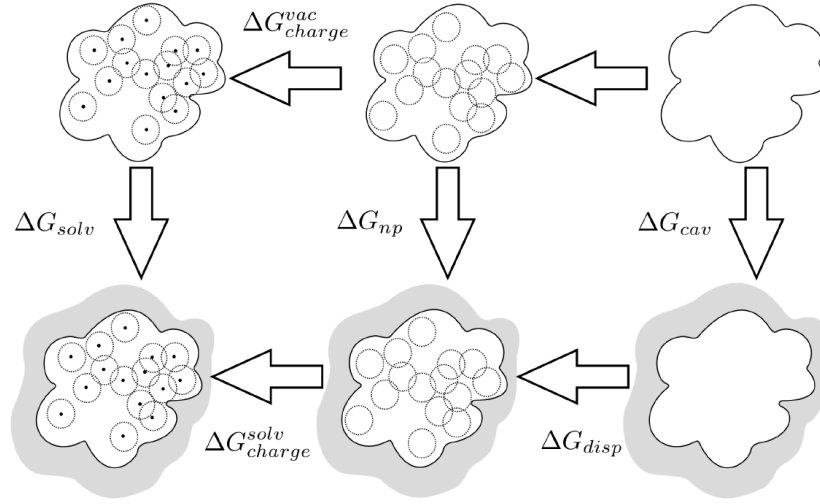


Figura 2.4: Ciclo termodinámico para la energía de solvatación (extraído de Cooper 2020 [3])

La distribución de potencial utilizada es la de reacción, este potencial considera la diferencia entre el potencial de las partículas en el vacío y el potencial de las partículas en el solvente.

$$\phi_{\text{reac}} = \phi_{\text{total}} - \phi_{\text{Coulomb}} \quad (2.31)$$

$$\phi_{\text{Coulomb}} = \frac{1}{4\pi\epsilon_1|\mathbf{x} - \mathbf{y}|} \quad (2.32)$$

Como la carga se encuentra concentrada en puntos del espacio, la distribución de carga queda definida como cargas puntuales, en otras palabras, la suma de las cargas individuales por deltas de Dirac que indican la posición, de esta forma

$$\begin{aligned} \Delta G &= \frac{1}{2} \int_{\Omega} \phi(\mathbf{x}) \sum_k q_k \delta(\mathbf{x}_{\mathbf{k}}) = \frac{1}{2} \sum_k q_k \int_{\Omega} \phi(\mathbf{x}) \delta(\mathbf{x}_{\mathbf{k}}) \\ \Delta G &= \frac{1}{2} \sum_k q_k \phi(\mathbf{x}_{\mathbf{k}}) \end{aligned} \quad (2.33)$$

3 | Metodología

En este capítulo se analizarán los métodos para obtener la energía de solvatación a partir de un archivo con la información de una proteína o molécula.

Para el caso de las proteínas o macromoléculas, sus estructuras son fácilmente descargables del banco de datos de proteínas [1] en formato `.pdb`. Esta base de datos contiene a libre disposición información de estructuras moleculares de interés obtenidas a través de métodos populares. Se destacan la cristalografía por rayos X y la resonancia magnética nuclear. Ambos métodos presentan una resolución del orden de los Angstroms que es interesante de estudiar para efectos de la incertidumbre.

Los archivos de extensión `.pdb` entregan información de la estructura interna a una escala mayor de la requerida para los modelos matemáticos usados. Por esta razón es necesario generar una estructura con la información relevante de la posición de los átomos con sus cargas y radios. Para esto se usan métodos de campos de fuerza (forcefields) para crear archivos de extensión `.pqr`¹ que sí contienen datos relevantes para el cálculo. Estos métodos de campos de fuerza son una mezcla de modelos teóricos y resultados experimentales. En este caso se usará el método *amber* [14].

El software utilizado es `pdb2pqr` [6] que hace una relajación de la estructura y agrega los átomos de hidrógeno que el archivo `.pdb` no considera. Un ejemplo de un archivo `pqr` es el siguiente

¹Información detallada del formato en https://docs.mdanalysis.org/1.0.0/documentation_pages/coordinates/PQR.html

1112_tetrachloroethane.pqr

ATOM	1	C1	TMP	1	1.085000	0.120000	-0.098000	0.010200	1.908000
ATOM	2	C2	TMP	1	1.460000	0.623000	-1.494000	0.243800	1.908000
ATOM	3	C11	TMP	1	1.597000	1.215000	1.218000	-0.135500	1.948000
ATOM	4	C12	TMP	1	3.232000	0.818000	-1.662000	-0.104700	1.948000
ATOM	5	C13	TMP	1	0.916000	-0.585000	-2.706000	-0.104700	1.948000
ATOM	6	C14	TMP	1	0.656000	2.179000	-1.863000	-0.104700	1.948000
ATOM	7	H1	TMP	1	1.544000	-0.853000	0.108000	0.097800	1.387000
ATOM	8	H2	TMP	1	-0.000000	0.000000	0.000000	0.097800	1.387000

3.1. Creación de mallas

Al tener archivos con la información de los átomos del soluto corresponde continuar el cálculo generando una malla que represente la superficie de exclusión. Para esto es conveniente filtrar la información de los átomos del archivo `.pqr` (que contiene posición, carga y radio de van der Waals de los átomos) para obtener sólo la posición y radio atómico en un archivo de extensión `.xyzr`.

Para la creación de las mallas existen diversos softwares externos, en nuestro caso usaremos NanoShaper [5] o MSMS [15] dependiendo del contexto. Empíricamente, se obtienen mejores resultados para mallas de moléculas pequeñas con MSMS y en el caso de proteínas o macromoléculas es preferible usar NanoShaper.

Ambos softwares pueden tratar las mallas en formatos convenientes que guardan la información de los vértices y caras (`.vert` y `.face`) para ser importados en `bempp` (en archivos de extensión `.off`).

`bempp` (Boundary Elements Method Python Package) es una librería de código abierto para el cálculo computacional mediante elementos de frontera [2]. Inicialmente llamada BEM++ por usar algunas funcionalidades de C++ ha ido gradualmente migrando de plataforma a python únicamente. Esta librería permite definir operadores de frontera, utilizar métodos numéricos en computación paralela y visualizar geometrías con facilidad.

3.2. Cálculo de energía de solvatación

La metodología de cálculo de energía de solvatación usando la librería `bempp` se basa en el trabajo Fernández [7], posteriormente recopilado por Search [17] en la librería `bem_electrostatics` [16].

El diagrama del algoritmo de cálculo de energía de solvatación se muestra en la figura 3.1, se basa en la definición de operadores integrales, tanto de frontera como de potencial (para el cálculo en la superficie de exclusión y dentro del dominio respectivamente para nuestro caso de estudio). Los usados son

$$V_n^a[\phi(x)] = \int_{\Gamma_a} G_n(x, y) \phi(y) dy$$

$$K_n^a[\phi(x)] = \int_{\Gamma_a} \frac{\partial G_n(x, y)}{\partial n(y)} \phi(y) dy$$

donde X es la coordenada donde se desea conocer el potencial, y la coordenada de la superficie de integración, subíndice n indica la función de Green que contiene el operador y el superíndice a la superficie sobre la que se está integrando.

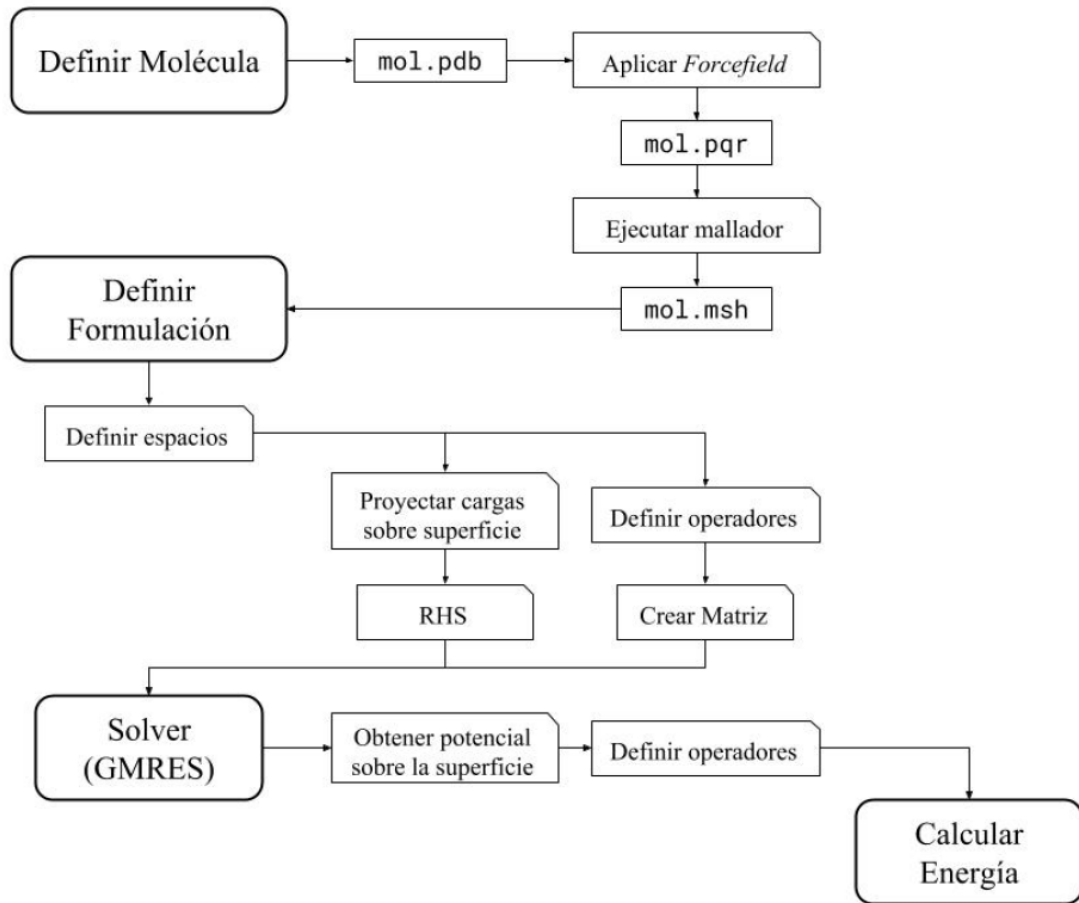


Figura 3.1: Esquema de funcionamiento del programa (extraído de Fernández 2019 [7])

Usando los operadores integrales de frontera se resuelve el sistema matricial 3.1 (Id es matriz identidad) para computar los valores en la frontera o superficie de exclusión, note que es otra forma de expresar el sistema de ecuaciones (2.26) y (2.27).

$$\begin{bmatrix} \frac{1}{2}\text{Id} + \mathbf{K}_L & -\mathbf{V}_L \\ \frac{1}{2}\text{Id} - \mathbf{K}_H & \frac{\epsilon_1}{\epsilon_2}\mathbf{V}_H \end{bmatrix} \begin{Bmatrix} \phi \\ \frac{\partial \phi}{\partial \nu} \end{Bmatrix} = \begin{Bmatrix} \sum_k q_k \\ 0 \end{Bmatrix} \quad (3.1)$$

Para el cálculo de potenciales internos, específicamente en los puntos donde se encuentran las cargas puntuales de los átomos se usan técnicamente los mismos operadores con distinta notación ($\mathcal{V}_L \mathcal{K}_L$), de esta forma el potencial de reacción ϕ_{reac} se expresa en la ecuación 3.2, que es otra forma de expresar 2.17 con operadores integrales.

$$\phi_{\text{reac}} = \mathcal{K}_L \left[\frac{\partial \phi}{\partial \mathbf{n}} \right] - \mathcal{V}_L [\phi] \quad (3.2)$$

3.3. Método Monte Carlo

Nuestro principal objetivo es analizar la incertidumbre de la posición de los átomos en la energía de solvatación causada por la resolución de la medición y las vibraciones moleculares que dan origen a las propiedades térmicas de un cuerpo.

La forma de estudiar el movimiento de los átomos será suponer que la posición de cada átomo se puede representar mediante variables aleatorias dependiendo del modelo de agitación, así se podrán generar aleatoriamente configuraciones nuevas que tengan en esencia la misma forma de la molécula a estudiar pero con los átomos en distintas posiciones. Cada modelo de agitación se estudiará en detalle en los capítulos siguientes.

El método Monte Carlo en este caso consiste en realizar repetidamente N pruebas de cálculo de energía de solvatación agitando la molécula. Para nuestro caso particular, se estudiará la desviación estándar σ como indicador para cuantificar la incertidumbre y su promedio μ .

Algoritmo 1 Simulación Monte Carlo para calcular parámetros estadísticos

- 1: Definir molécula y cantidad de pruebas N
 - 2: Inicializar lista vacía $\mathbf{G}$
 - 3: **Para** $i = 1$ hasta N
 - 4: Agitar molécula y crear archivo `.pqr`
 - 5: Generar malla de superficie de exclusión
 - 6: Calcular energía de solvatación
 - 7: Agregar energía calculada a lista $\mathbf{G}$
 - 8: Calcular promedio y desviación estándar en base a cálculos almacenados en lista $\mathbf{G}$
-

3.4. Base de datos de FreeSolv

Con la motivación de verificar los resultados obtenidos con BEM computacionalmente, compararemos la energía de solvatación de moléculas pequeñas entregadas en la base de datos FreeSolv [12] así como la presentada en [11].

Es interesante destacar que esta base de datos entrega información sobre grupos funcionales de varias moléculas que han sido objeto de estudio para comparaciones entre dinámica molecular con solvente explícito y método de elementos de frontera con solvente implícito [8], logrando obtener valores ajustados para la constante dieléctrica del soluto como se analizará en capítulos siguientes.

4 | Agitación esférica radio constante

Tomando como principal motivación la incertidumbre en la posición de los átomos del experimento con el que se obtienen los datos experimentales en la generación de una estructura de una macromolécula del banco de dato de proteínas [1], buscamos obtener la desviación estándar en la energía de solvatación para grandes biomoléculas usando el método Monte Carlo, en este caso hicimos cálculos para una inmunoglobulina de 460 átomos identificada como 1PGB.

Para buscar relaciones probamos variar el radio de agitación encontrando un aumento en la desviación estándar y magnitud de la energía de solvatación.

4.1. Metodología y algoritmos

El experimento consta de tomar la información de la estructura de una proteína y calcular su energía de solvatación sin ningún tipo de agitación, para esto se usa la herramienta `pdb2pqr` que prepara la estructura en un archivo `.pqr` reconstruyendo los átomos faltantes, agregando hidrógenos y asignando cargas atómicas y radios usando un campo de fuerza específico (en este caso AMBER [14]).

Teniendo un valor de referencia, se procede a agitar la estructura, asignando una nueva posición a cada átomo dentro de una esfera de radio constante y generar un archivo `.pqr` con la información de la estructura a calcular su energía. Ejemplo gráfico en figura 4.1, las rectas indican enlaces entre átomos y el área demarcada la región probable donde puede ser reubicado el átomo luego de la agitación molecular.

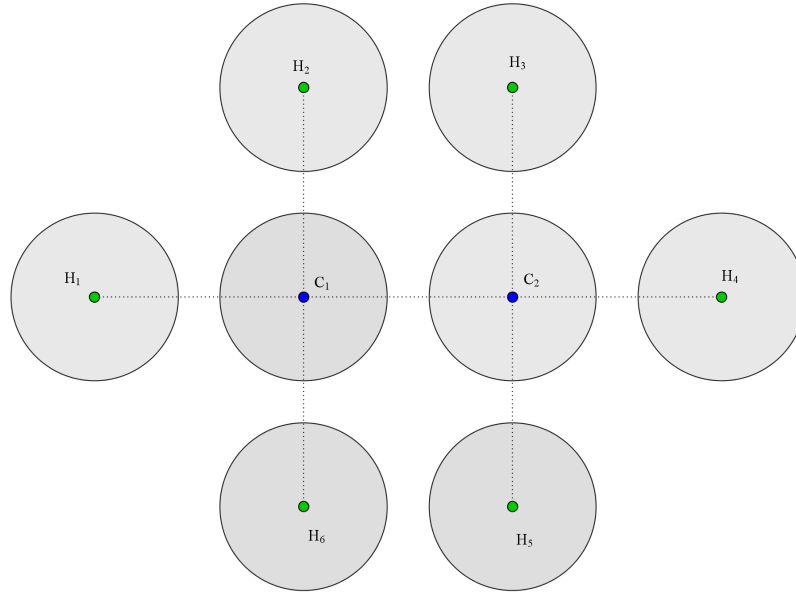


Figura 4.1: Ejemplo en dos dimensiones de áreas en las que puede ser reasignada la posición de cada átomo de la molécula etano.

Sea r , θ y φ el radio, ángulo azimutal y polar en coordenadas esféricas que hacen referencia al movimiento de un átomo en una esfera de agitación, su valor se modela como variables aleatorias de distribución de probabilidad uniforme (o meseta) tal que

$$0 \leq r \leq R$$

$$0 \leq \theta \leq 2\pi$$

$$0 \leq \varphi \leq \pi$$

Donde R es conocido como el radio de la esfera de agitación, donde su valor es constante y estudiado. Así en coordenadas cartesianas, la nueva posición del átomo agitado con respecto a su posición inicial (con subíndice 0) es

$$x = x_0 + r \cdot \cos \theta \cdot \sin \varphi$$

$$y = y_0 + r \cdot \sin \theta \cdot \sin \varphi$$

$$z = z_0 + r \cdot \cos \varphi$$

Este paso se repite según el número de muestras del método Monte Carlo y según la cantidad de radios de agitación a considerar (en primera instancia hicimos 100 pruebas aleatorias para cada uno de los 5 radios considerados). En resumen, los algoritmos explicados en pseudocódigo son

Algoritmo 2 Simulación Monte Carlo con agitación esférica a radio constante

- 1: Definir radios de prueba
 - 2: Definir molécula
 - 3: Para cada radio de prueba R
 - 4: Repetir N veces
 - 5: Agitar molécula y crear archivo `.pdb` nuevo (Algoritmo 3)
 - 6: Usando `pdb2pqr` crear archivo `.pqr`
 - 7: Calcular energía de solvatación (Figura 3.1)
 - 8: Calcular promedio y desviación estándar
-

Algoritmo 3 Agitación esférica a radio constante

Entrada: radio de prueba R , archivo `pqr` o `pdb` molécula original

- 1: Filtrar información atómica y guardar en lista
 - 2: Para cada átomo de la molécula guardado en lista
 - 3: Asignar variables de posición en x_0, y_0, z_0
 - 4: Generar variables aleatorias uniformes r, θ, φ
 - 5: Asignar variables de nueva posición x, y, z
 - 6: Modificar posición en lista
 - 7: Guardar nuevo archivo con la agitación realizada
-

4.2. Análisis de resultados

El experimento calculó la energía de solvatación con 100 configuraciones moleculares distintas para cinco radios de agitación de 0.1, 0.5, 1, 2 y 5 (Å) respectivamente. Cada prueba fue controlada con una tolerancia para el método GMRES de 10^{-5} con un máximo de 150 iteraciones.

Tabla 4.1: Valores estadísticos simulación Monte Carlo para energía de solvatación de biomolécula 1PGB

Radio (Å)	Promedio (kcal/mol)	Desv. Estandar (kcal/mol)
0.1	-255.09	2.10
0.5	-260.75	6.35
1.0	-279.39	12.13
2.0	-338.17	24.48
5.0	-547.29	55.33

Los cálculos tomaron cerca de 9 horas en terminar, donde los valores del último radio de prueba de 5 Å no son muy confiables por estar en el tope de iteraciones de GMRES, no obstante, el resto de pruebas entrega resultados interesantes mostrados en la figura 4.2

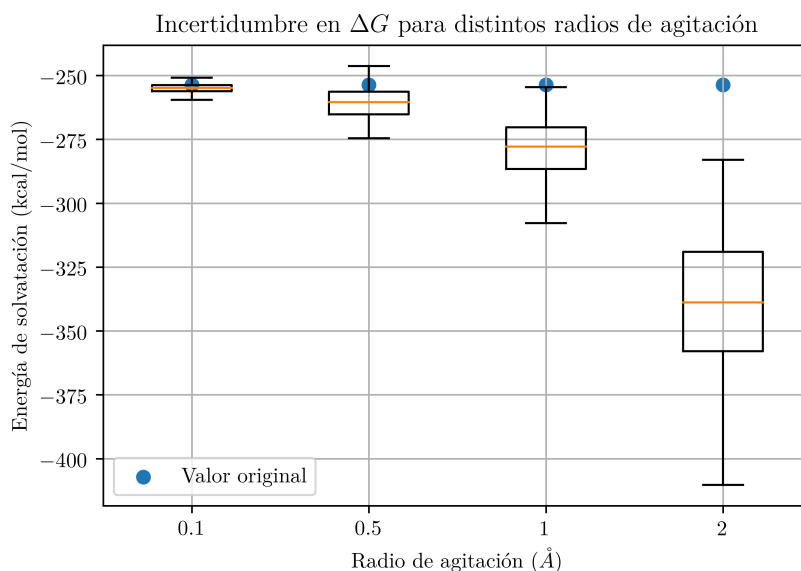


Figura 4.2: Visualización de la incertidumbre a través de cuartiles a distintos radios de agitación para la energía de solvatación de molécula 1PGB

De igual manera es posible notar una relación lineal entre la desviación estándar de la simulación Monte Carlo para cálculo de la energía de solvatación y el radio de agitación (figura 4.3)

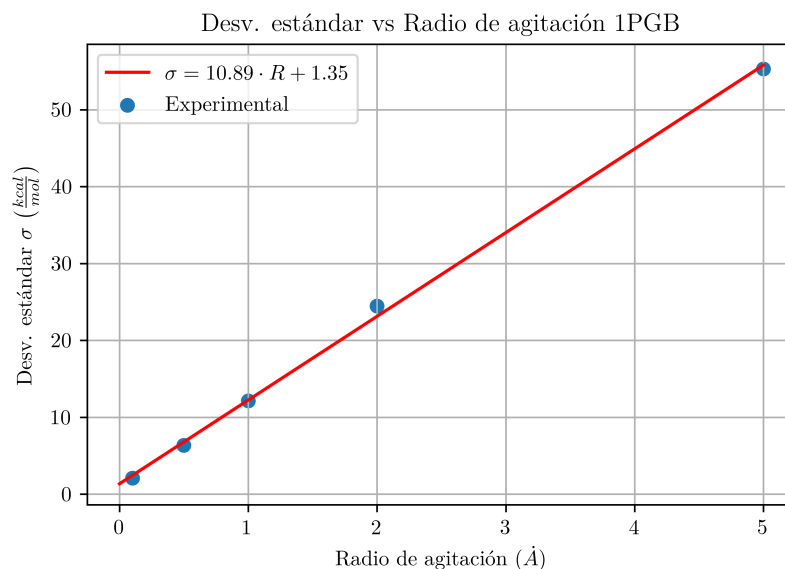


Figura 4.3: Relación lineal entre la desviación estándar y el radio de agitación para cálculos con método Monte Carlo de la energía de solvatación de molécula 1PGB

Se observa como al aumentar el radio de agitación la incertidumbre en la energía de solvatación no sólo aumenta sino que también el promedio disminuye. Actualmente no se tiene una explicación física para este fenómeno, es posible decir que la proteína se vuelve más estable al ser agitada (entiéndase que mientras mayor magnitud absoluta tenga la energía de solvatación, más estable es).

Otra explicación pendiente de probar es que algunas nuevas configuraciones aleatorias de esta macromolécula presentan cúmulos o agrupaciones de átomos con baja distancia de separación, lo que provocaría una polarización mayor en el solvente. Al tener esta posibilidad decidimos que las conexiones naturales mediante enlaces covalentes de los distintos átomos de la estructura se deben mantener para evitar estas agrupaciones sin orden ni sentido físico. En el capítulo siguiente abordaremos una nueva metodología para agitar las posiciones de los átomos de una molécula manteniendo la conectividad inicial.

5 | Agitación con conectividad

Con la principal motivación de evitar modificar mayormente los enlaces naturales de los átomos de una molécula y darle un sentido más físico a la agitación aleatoria de posiciones, creamos un método que pretende mantener la forma de la molécula.

Para probar este modelo usamos moléculas de la base de datos FreeSolv [12] y comparamos los resultados con los experimentos de Mobley et. al. 2009 [11] quien usó dinámica molecular y solvente explícito en un conjunto de aproximadamente 500 moléculas.

La relación de la incertidumbre entre los métodos de solvente explícito y solvente implícito aún no ha sido estudiada con los parámetros y supuestos que hacemos en este trabajo. Así, un primer acercamiento es comparar la incertidumbre entregada por dinámica molecular como intervalos de confianza con la desviación estándar de la simulación Monte Carlo.

5.1. Metodología

El formato de archivo .mol2 guarda la estructura de una molécula indicando la conectividad de los distintos átomos (en la sección BOND) ², esto con un número identificador del enlace, identificador del átomo de origen, identificador del átomo objetivo y tipo de enlace.

²Más información del formato de archivo se encuentra en <http://chemyang.ccnu.edu.cn/ccb/server/AIMMS/mol2.pdf>

1112_tetrachloroethane.mol2

```

@<TRIPOS>MOLECULE
1112_tetrachloroethane
      8      7      1      0      0
SMALL
USER_CHARGES
@<TRIPOS>ATOM
  1 C1      1.0848      0.1201      -0.0975 C.3      1 TMP 0.0102
  2 C2      1.4605      0.6234      -1.4942 C.3      1 TMP 0.2438
  3 C11     1.5971      1.2148      1.2175 C1      1 TMP -0.1355
  4 C12     3.2317      0.8176      -1.6617 C1      1 TMP -0.1047
  5 C13     0.9165     -0.5849     -2.7058 C1      1 TMP -0.1047
  6 C14     0.6563      2.1791     -1.8630 C1      1 TMP -0.1047
  7 H1      1.5438     -0.8532      0.1083 H      1 TMP 0.0978
  8 H2     -0.0000      0.0000      0.0000 H      1 TMP 0.0978
@<TRIPOS>BOND
  1      1      2 1
  2      1      3 1
  3      2      4 1
  4      2      5 1
  5      2      6 1
  6      1      7 1
  7      1      8 1
@<TRIPOS>SUBSTRUCTURE
  1 TMP 1
@<TRIPOS>COMMENT
SMALL
USER_CHARGES

```

Para simplificar el experimento decidimos trabajar con un conjunto reducido de moléculas, este grupo tiene las siguientes características:

- Los enlaces entre átomos son solo de tipo simple.
- No presentan estructuras cíclicas, es decir, la cantidad de enlaces es uno menos que la cantidad de átomos.
- Se puede generar la estructura en base a un único átomo inicial, es decir, sólo un átomo puede ser considerado únicamente como origen y no estar en el conjunto de átomos objetivo en la información de conectividad.

De esta forma, se pueden ubicar las cargas aleatoriamente a partir de un único átomo sin correr el riesgo de que sea agitado más adelante.

Este conjunto de moléculas candidatas puede ser caracterizado posteriormente por tres grandes grupos denominados

- Alcoholes, por contener el grupo hidroxilo $-OH$
- Aminas, por contener el grupo amino $-NH_2$
- Alcanos (y otros), por ser en su mayoría hidrocarburos, contener el grupo alquilo o no presentar un grupo funcional determinado pero cumplir con las condiciones de enlaces simples y no tener ciclos en su estructura.

Se busca acercar los resultados obtenidos con método Monte Carlo y modelo de solvente implícito al promedio y desviación estándar de cada molécula calculada a los valores indicados por Mobley [11] usando dinámica molecular y solvente explícito. Para esto probamos distintos valores de dos parámetros de interés:

- Porcentaje de variación de distancia de enlace entre átomos.
- Ángulo de variación (polar y azimutal) con respecto a la orientación inicial.

Estos modelos de agitación entregan un volumen de ubicación posible para un átomo objetivo de forma de casquete esférico. En la figura 5.1 se muestra en rojo el área donde puede ser reubicado el átomo objetivo.

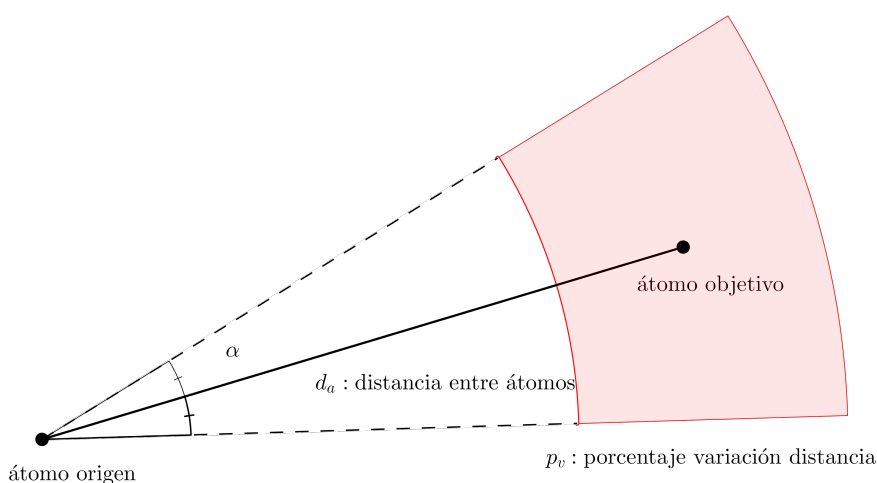


Figura 5.1: Diagrama de modelo de agitación entre átomos

Sea d_a la distancia de enlace entre átomos, p_v el porcentaje de variación de la distancia entre átomos y α el ángulo de variación tanto polar como azimutal con respecto a la orientación inicial. La nueva posición del átomo objetivo se generará aleatoriamente en base a su posición relativa con el átomo de origen.

Sea $\mathbf{x}_{orig} = (x_1, y_1, z_1)$ la posición del átomo de origen y $\mathbf{x}_{obj} = (x_2, y_2, z_2)$ la posición del átomo objetivo (o átomo a agitar), así como

$$\mathbf{x}_{obj} - \mathbf{x}_{orig} = (X, Y, Z)$$

coordenadas cartesianas relativas, se definen las variables R_0, θ_0, ϕ_0 como coordenadas esféricas relativas del átomo objetivo con respecto al de origen (sin agitar)

$$R_0 = \sqrt{X^2 + Y^2 + Z^2}$$

$$\theta_0 = \begin{cases} \arctan(Y/X) & \text{para } X > 0 \\ \frac{\pi}{2} \cdot \frac{Y}{|Y|} & \text{para } X = 0 \wedge Y \neq 0 \\ \pi + \arctan(Y/X) & \text{para } X < 0 \\ \text{V.A. uniforme entre } 0 \text{ y } 2\pi & \text{para } X = 0 \wedge Y = 0 \end{cases}$$

$$\varphi_0 = \arccos(Z/R)$$

Así, la nueva posición relativa R, θ, φ del átomo de origen en función de los parámetros p_v y α se generará como una variable aleatoria uniforme, tal que

$$R_0(1 - p_v) \leq R \leq R_0(1 + p_v)$$

$$\theta_0 - \alpha \leq \theta \leq \theta_0 + \alpha$$

$$\varphi_0 - \alpha \leq \varphi \leq \varphi_0 + \alpha$$

Con esto las nuevas coordenadas cartesianas relativas del átomo objetivo con respecto al átomo de origen serán X', Y', Z' definidas como

$$X' = X + R \cdot \cos(\theta) \cdot \sin(\varphi)$$

$$Y' = Y + R \cdot \sin(\theta) \cdot \sin(\varphi)$$

$$Z' = Z + R \cdot \cos(\varphi)$$

Los algoritmos explicados en pseudocódigo son

Algoritmo 4 Simulación Monte Carlo de agitación con conectividad

- 1: Definir moléculas y cantidad de pruebas N
 - 2: Definir lista de porcentajes de variación de distancia
 - 3: Definir lista de ángulos de variación
 - 4: Para cada porcentaje de variación p_v
 - 5: Para cada ángulo de variación α
 - 6: Para cada molécula
 - 7: Repetir N veces
 - 8: Agitar molécula y crear archivo `.pqr` nuevo (Algoritmo 5)
 - 9: Calcular energía de solvatación (Figura 3.1)
 - 10: Calcular promedio y desviación estándar
-

Algoritmo 5 Agitación con conectividad

Entrada: Porcentaje de variación p_v , ángulo de variación α , archivos `.pqr` y `.mol2` inicial

Salida: Lista de posiciones de átomos agitados

- 1: Filtrar información y generar lista de posiciones de átomos
 - 2: Generar lista con conectividad de átomos ³
 - 3: Para cada ID_atomo_origen y ID_atomo_objetivo
 - 4: Obtener $\mathbf{x}_{orig}$ y $\mathbf{x}_{obj}$
 - 5: Definir x_1, y_1, z_1 así como x_2, y_2, z_2
 - 6: Definir coordenadas R_0, θ_0, φ_0
 - 7: Obtener variables aleatorias R, θ, φ
 - 8: Definir coordenadas X', Y', Z'
 - 9: Modificar posición de átomo objetivo en lista de posiciones
-

³Tiene la forma (ID_enlace, ID_atomo_origen, ID_atomo_objetivo, tipo_enlace), en el paso siguiente se recorren sólo los ID de átomos de origen y objetivo

5.2. Análisis de resultados

Probamos 9 combinaciones distintas de parámetros de distancia entre átomos y ángulo, las opciones fueron

- Para la variación de distancia de enlace 5 %, 3 % y 1 %.
- Para el ángulo con respecto a orientación inicial $\frac{\pi}{4}, \frac{\pi}{6}, \frac{\pi}{9}$.

La constante dieléctrica del soluto fue seleccionada con base en la mejor aproximación mediante grupos funcionales a los resultados obtenidos por Mobley [11] en su base de datos FreeSolv [12] según comparaciones hechas con la ecuación de Poisson-Boltzmann por Huencho [8].

Tabla 5.1: Constante dieléctrica (adimensional) para grupos funcionales presentes dentro de las moléculas candidatas al experimento

Grupo funcional	Constante dieléctrica
Alcanos	1.40
Alcoholes	1.35
Aminas	1.15

Analizaremos en base a tres indicadores: Coeficiente de correlación de Pearson R^4 , raíz del error cuadrático medio RMSD 5.1 y error absoluto medio MUE 5.2. Los subíndices i y j hacen referencia al valor de BEM o DM según sea el caso

$$\text{RMSD} = \sqrt{\frac{\sum_{k=1}^N (X_i - X_j)^2}{N}} \quad (5.1)$$

$$\text{MUE} = \frac{\sum_{k=1}^N |X_i - X_j|}{N} \quad (5.2)$$

En la figura 5.2 se observa para los parámetros de distancia y ángulo más bajos (1 % y $\pi/9$) existe una buena correlación ($R = 0,92$). Los indicadores de error (RMSD=17.8 (kcal/mol) y MUE=10.1 (kcal/mol)) no aportan conclusiones más que las vistas gráficamente. La recta $y = x$ indica los puntos donde ambos métodos entregan el mismo valor. Es interesante observar que el método Monte Carlo aparentemente aumenta la magnitud en el promedio de la energía de solvatación, tal como el capítulo anterior sobre la simulación con una proteína.

En la figura 5.3 se puede ver que si bien los promedios de energía de solvatación tienen una buena correlación ($R_\mu = 0,92$), la desviación estándar se dispara por sobre lo esperado y no presenta una correlación aceptable ($R_\sigma = 0,39$).

⁴Calculado con la librería `numpy` <https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html>

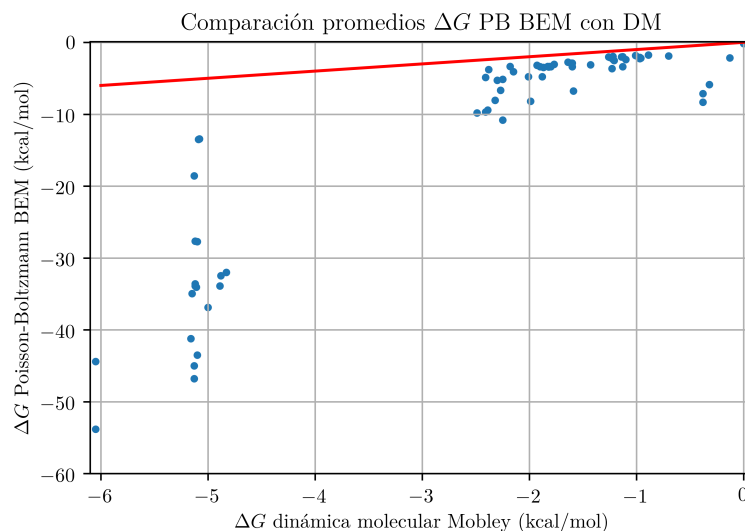


Figura 5.2: Comparación de valor promedio de energía de solvatación calculada (BEM) con respecto a DM (Mobley)

Para valores con baja magnitud de energía de solvatación ($\Delta G_{Mobley} > -4(kcal/mol)$) se tienen resultados aceptables, estas moléculas fueron clasificadas mayormente como alcanos. En cambio, para las moléculas restantes, clasificadas como alcoholes y aminas, el método no presenta una buena correlación con dinámica molecular y solvente explícito.

En el capítulo siguiente analizaremos un método que pretende modelar la agitación térmica de los átomos a través del concepto de velocidad térmica derivado de la distribución de Maxwell-Boltzmann.

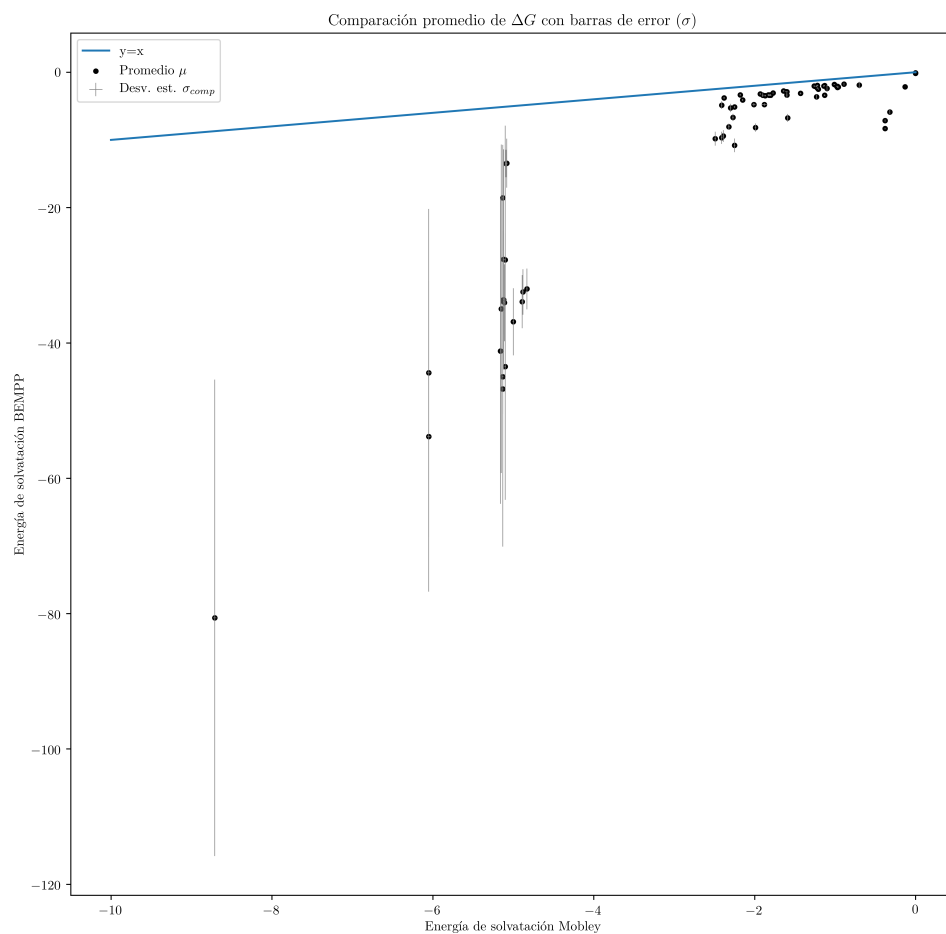


Figura 5.3: Promedio de energía de solvatación con BEM vs Mobley, para variación dist. 1% y ángulo $\pi/9$

6 | Agitación térmica

Con la idea de modelar las posiciones de los átomos tomando en cuenta las vibraciones moleculares, nos basaremos en el concepto de velocidad térmica para crear un modelo de agitación aleatoria con esferas de radio no constante. Este trabajo se comparará con los resultados de dinámica molecular y solvente explícito de Mobley 2009 [11]. También intentaremos probar la hipótesis de que somos capaces de obtener la incertidumbre explícita a partir de la implícita.

6.1. Longitud térmica promedio y tiempo característico

Derivado de la distribución de Maxwell-Boltzmann sobre la probabilidad de velocidad de un gas, se desglosa el concepto de velocidad térmica v_t

$$v_t = \sqrt{\frac{k_B \cdot T}{m}} \quad (6.1)$$

donde k_B y T se refieren a la constante de Boltzmann y la temperatura absoluta. Al expresar la velocidad térmica como un cociente con forma distancia sobre tiempo, presentamos el concepto de **longitud térmica promedio** (por sus siglas ATL: Average Thermal Length) y un tiempo característico t_c que buscaremos en este capítulo. De esta forma la expresión anterior puede modificarse para expresar ATL como

$$\text{ATL} = t_c \cdot \sqrt{\frac{k_B \cdot T}{m}} \quad (6.2)$$

6.2. Algoritmo de agitación térmica

Con respecto a la ecuación (6.2), en primera instancia buscaremos variar el término del tiempo característico t_c para encontrar valores similares de desviación estándar con los intervalos presentados por Mobley usando dinámica molecular y solvente explícito [11]. El modelo de agitación será aleatorio para cada átomo dentro de una esfera de radio igual a la longitud térmica promedio en función de su masa a una temperatura de $T = 298[K]$.

Realizamos una simulación Monte Carlo agitando cada átomo en una esfera de radio igual a la longitud térmica promedio variando los valores de tiempo característico. Note que los valores de masa de cada átomo son obtenidos por tabla y divididos por el número de Avogadro.

En la figura 6.1 se puede observar como los átomos de hidrógeno al tener menos masa presentan radio de agitación mucho mayor que el carbono. Note que este método de agitación sigue la lógica vista en el capítulo 4 cambiando el radio de la esfera donde se reubica cada átomo.

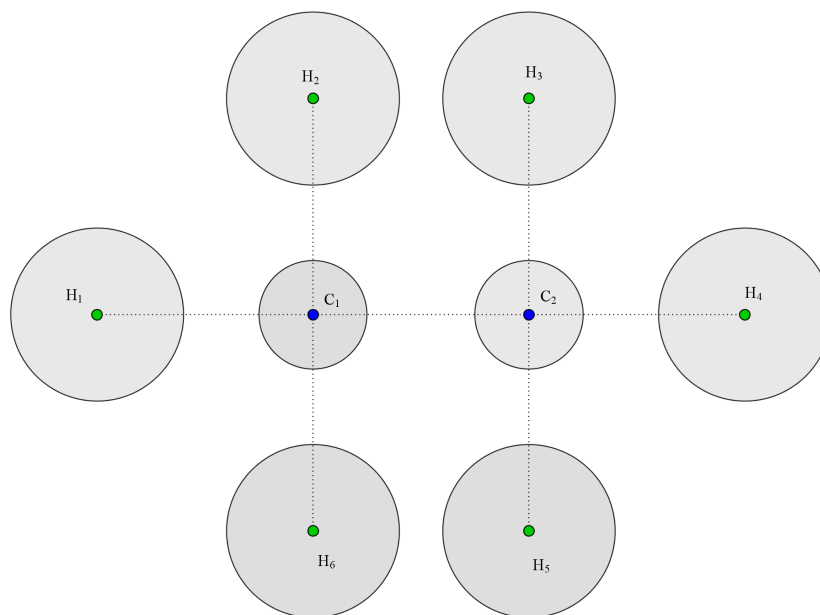


Figura 6.1: Ejemplo en dos dimensiones de las áreas en que pueden ser reubicados los átomos de una molécula de etano

6.3. Relación exponencial entre tiempo característico y desviación estándar

Usando el algoritmo descrito anteriormente, probamos para dos moléculas (tetrafluorometano y 1,1,1,2 - tetracloroetano) varios órdenes de magnitud de tiempos característicos. Al visualizar los puntos en una escala logarítmica para ambos ejes, encontramos una relación lineal entre $\ln(\sigma)$ y $\ln(t_c)$ (figuras 6.2 y 6.3), que tendrá como coeficientes α y β

$$\ln(\sigma) = \alpha \cdot \ln(t_c) + \beta \quad (6.3)$$

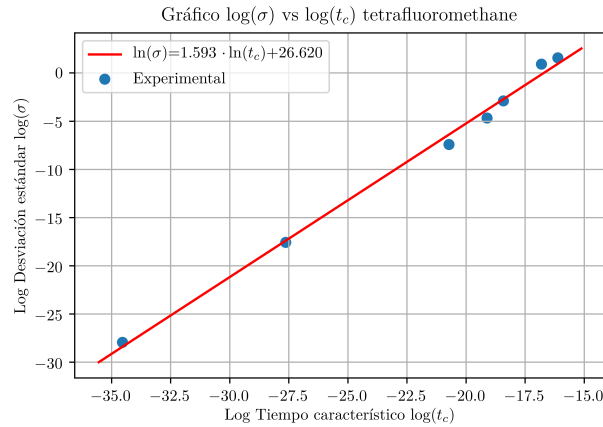


Figura 6.2: Regresión lineal para tetrafluorometano

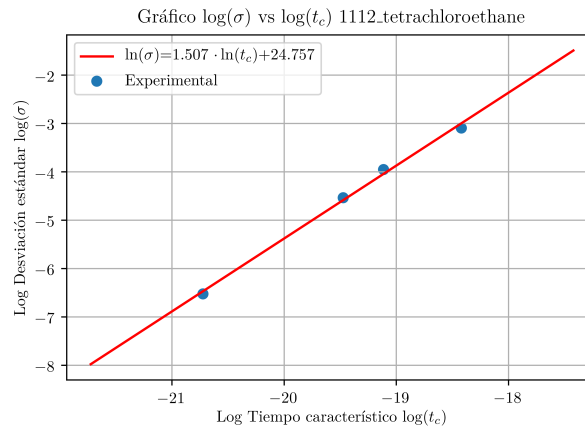


Figura 6.3: Regresión lineal para 1,1,1,2 tetracloroetano

Al ser uno de los objetivos de estudio validar mediante dinámica molecular y solvente explícito la incertidumbre generada por las posiciones, creamos un algoritmo para obtener los valores de tiempo característico para cada molécula que consiste en

1. Cuantificar incertidumbre mediante el método Monte Carlo para dos (o más) valores de tiempo característico.
2. Usando los valores anteriores, se ajusta una recta con la ecuación (6.3) y se obtiene un tiempo característico t_c^* usando la incertidumbre indicada según dinámica molecular con solvente explícito σ_M .

$$t_c^* = \exp\left(\frac{\ln(\sigma_M) - \beta}{\alpha}\right) \quad (6.4)$$

3. Usando el método de Monte Carlo se calcula una nueva desviación estándar σ y promedio de cada molécula.

Finalmente, la relación propuesta entre tiempo característico y desviación estándar del método se expresa con la ecuación

$$\sigma = e^{\beta} \cdot t_c^{\alpha} \quad (6.5)$$

Una idea es intentar encontrar una relación para los exponentes α y β de modo que se asocie con alguna característica física de la molécula. Algunas metodologías de búsqueda se analizarán más adelante.

Cabe destacar que aún está en estudio encontrar una forma eficiente de calcular los parámetros de promedio y desviación estándar usando la simulación Monte Carlo o un método alternativo. En este caso podemos concluir con las figuras 6.4 y 6.5 que los valores buscados se estancaban a las 100 pruebas.

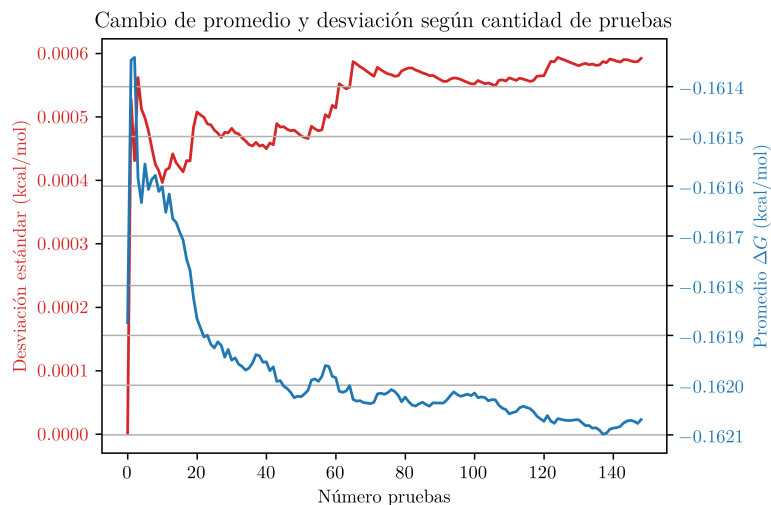


Figura 6.4: Cambios en μ y σ según la cantidad de pruebas para tetrafluorometano

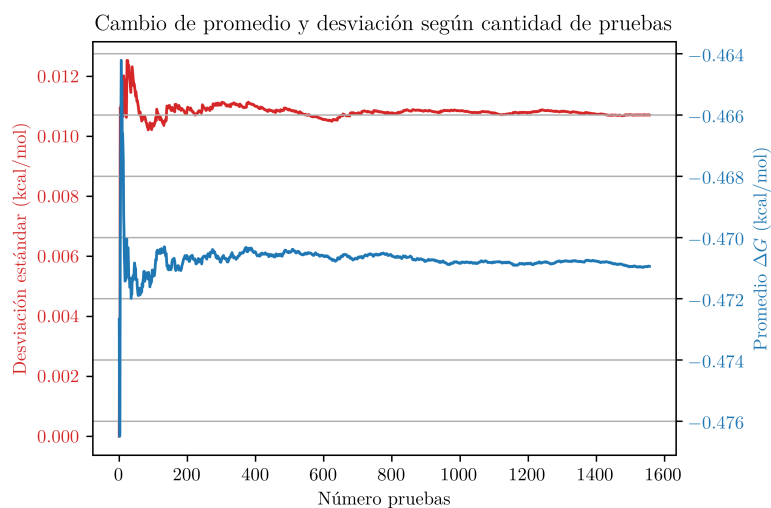


Figura 6.5: Cambios en μ y σ según la cantidad de pruebas para 1112_tetrachloroetano

6.4. Comparación modelos implícito y explícito para energía de solvatación

Comparamos los promedios de energía de solvatación entregados por Mobley usando dinámica molecular y solvente explícito con los calculados usando Poisson-Boltzmann y solvente implícito junto con el método Monte Carlo expuesto anteriormente.

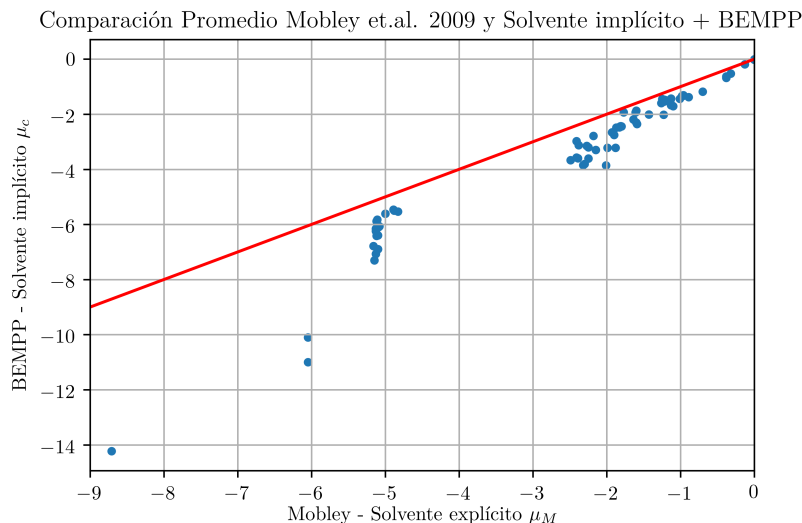


Figura 6.6: Comparación de promedios de energía de solvatación con valores de Mobley 2009

Para el grupo de 75 moléculas candidatas del capítulo anterior encontramos una buena correlación para el promedio ($R = 0,97$), los valores de energía de solvatación si bien decrecen no lo hacen con la misma magnitud que en el capítulo anterior que consideraba un modelo de agitación con conectividad.

6.5. Comparación de incertidumbre en energía de solvatación

Siguiendo la hipótesis de que es posible obtener la incertidumbre indicada con solvente explícito a partir del implícito, comparamos en primera instancia la desviación estándar obtenida con la simulación Monte Carlo con la incertidumbre de la base de datos de Mobley, se puede observar una mejor correlación que la del capítulo anterior ($R = 0,52$).

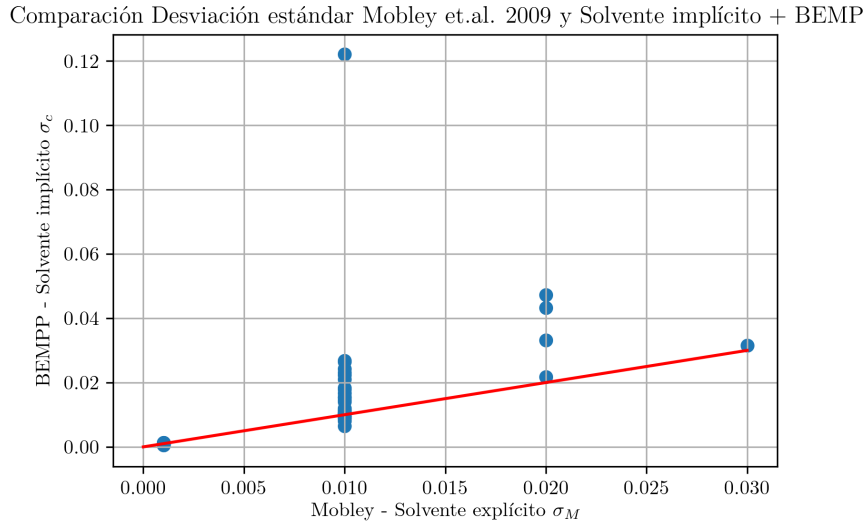


Figura 6.7: Comparación de incertidumbre con valores de Mobley 2009

Cabe destacar que la base de datos de Mobley indica de forma gruesa la incertidumbre, se desconoce la metodología de representación del intervalo de confianza y por eso debe considerarse este análisis como un primer acercamiento a una metodología de comparación de incertidumbre entre dos modelos de solvente distintos.

Si bien la simulación Monte Carlo disminuye el intervalo de confianza de un indicador promedio a mayor cantidad de pruebas (con factor de $\frac{1}{\sqrt{N}}$), creemos que el análisis pueda tener ruido al sumar una nueva variable al estudio y consideramos prudente estudiar los intervalos de confianza bajo el supuesto de que la energía de solvatación puede expresarse como una función de densidad de probabilidad normal.

6.6. Búsqueda de relaciones y clusters

En este punto, al tener una gran cantidad de datos y variables se torna interesante buscar relaciones entre lo obtenidos computacionalmente. En este caso hicimos dos análisis: manualmente viendo gráficos de dispersión con código de colores y usando una tabla de correlación.

6.6.1. Comparación métodos anteriores

Para el primer experimento con 75 moléculas pequeñas, la figura 6.8 muestra la correlación entre el promedio obtenido con dinámica molecular y Poisson-Boltzmann con solvente implícito. En colores se muestran los grupos funcionales del capítulo anterior donde los colores corresponden a:

1. Alcanos
2. Alcoholes
3. Aminas

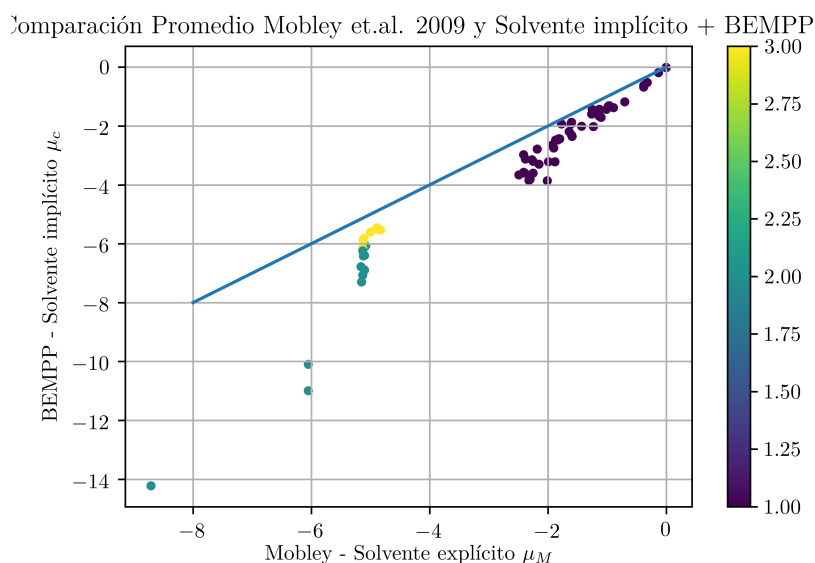


Figura 6.8: Comparación de promedios de energía de solvatación con Mobley 2009, en colores se indican grupos funcionales de moléculas

Para el caso de la comparación de la desviación estándar con la incertidumbre presentada con dinámica molecular (figura 6.9), no encontramos la linealidad esperada, esto es atribuible a los pocos pasos utilizados en la interpolación al buscar el tiempo característico t_c^* que acerca más los datos de solvente implícito al explícito. Con respecto a la clave de colores que muestra el grupo funcional sólo se puede concluir que la desviación estándar de los alcoholes se aleja más de los datos de Mobley 2009.

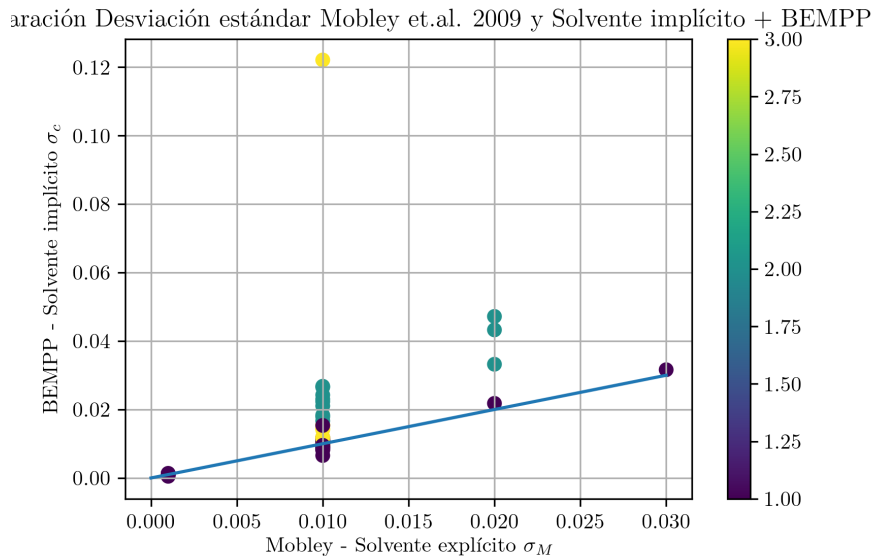


Figura 6.9: Comparación de desviación estándar DM y PB. En colores se muestran grupos funcionales

6.6.2. Tabla de correlación

En la simulación realizada obtuvimos para las 75 moléculas 12 variables de interés que se enumeran a continuación

1. N_{atomos} : Cantidad de átomos de cada molécula
2. α : Exponente de tiempo característico en ecuación (6.5)
3. β : Componente de coeficiente del tiempo característico en ecuación (6.5)
4. t_c : Tiempo característico
5. σ_{Mobley} : Incertidumbre en la energía de solvatación usando dinámica molecular y solvente explícito.
6. σ_{BEM} : Desviación estándar obtenida computacionalmente usando Poisson-Boltzmann y solvente implícito.

7. μ_{SASA} : Promedio del área de la superficie accesible por el solvente (SAS) de las moléculas generadas de forma aleatoria.
8. σ_{SASA} : Desviación estándar del áreas de las superficies accesibles por el solvente (SAS).
9. μ_{SESV} : Promedio del volumen interno a la superficie de exclusión del solvente (SES) de las moléculas generadas aleatoriamente.
10. σ_{SESV} : Desviación estándar de los volúmenes formados con SES.
11. μ_{Mobley} : Promedio de la energía libre electrostática (o de solvatación) obtenida por Mobley (dinámica molecular y solvente explícito).
12. μ_{BEM} : Promedio de la energía de solvatación obtenido con Poisson-Boltzmann y solvente implícito.

Utilizando estos arreglos de datos para las 75 moléculas, ordenamos en una tabla (o matriz triangular superior) el coeficiente de correlación de Pearson.

Tabla 6.1: Tabla de correlación para simulación con 75 moléculas candidatas

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	0.01	0.12	-0.32	0.01	0.24	0.93	-0.15	0.92	0.04	-0.39	-0.29
2		1.00	0.97	-0.08	0.17	0.25	-0.13	-0.04	-0.12	-0.15	-0.46	-0.47
3			1.00	-0.25	0.28	0.36	-0.06	-0.19	-0.06	-0.26	-0.65	-0.65
4				1.00	0.25	-0.13	-0.20	0.87	-0.20	0.80	0.48	0.45
5					1.00	0.52	-0.03	0.12	-0.02	0.16	-0.55	-0.62
6						1.00	0.14	0.06	0.14	0.21	-0.59	-0.58
7							1.00	-0.14	1.00	0.08	-0.18	-0.11
8								1.00	-0.14	0.93	0.39	0.39
9									1.00	0.08	-0.17	-0.10
10										1.00	0.32	0.34
11											1.00	0.97
12												1.00

Para nuestro interés en particular, analizaremos valores cercanos a 1 y -1, que indican que existe una correlación ya sea positiva o negativa. En este caso serán de nuestro interés las relaciones de los grupos de variables con un coeficiente mayor a 0.5 o menor a -0.5 y que sean relevantes al estudio (por ejemplo, no considerar correlaciones entre número de átomos y volumen por ser físicamente lógico):

- 2-3 α vs β (0.97)
- 4-8 t_c^* vs σ_{SASA} (0.87)
- 3-12 β vs μ_{BEM} (-0.65)
- 6-12 σ_{BEM} vs μ_{BEM} (-0.58)

Para el primer par de variables presentes en la ecuación (6.5), se observa una clara correlación lineal positiva. En la figura 6.10 se muestran los grupos funcionales en la clave de colores utilizada anteriormente, es tentador hacer la diferencia entre alcanos y alcoholes con aminas por resaltar la forma de dos rectas con igual pendiente.

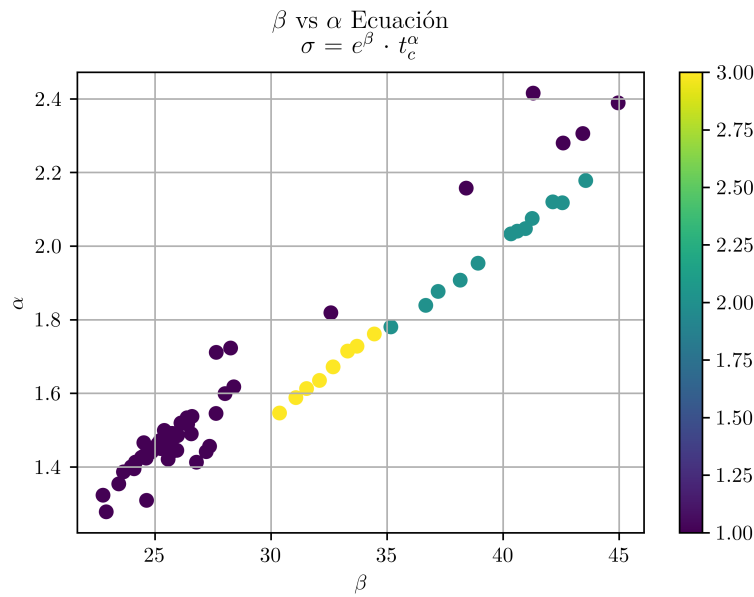


Figura 6.10: Relación entre α y β con 75 moléculas candidatas. Grupos funcionales resaltados en color.

De esta relación es posible desglosar que las variables pueden ser escritas en función de la otra, esto es $\alpha = \alpha(\beta)$ o bien $\beta = \beta(\alpha)$:

$$\alpha = 0,042 \cdot \beta + 0,381$$

así como

$$\beta = 22,262 \cdot \alpha - 6,585$$

Por otra parte y guiado por el resultado del coeficiente de correlación entre el promedio de energía μ_{BEM} y β , el mismo gráfico con barra de colores indicando los valores de energía de solvatación promedio se muestra en la figura 6.11

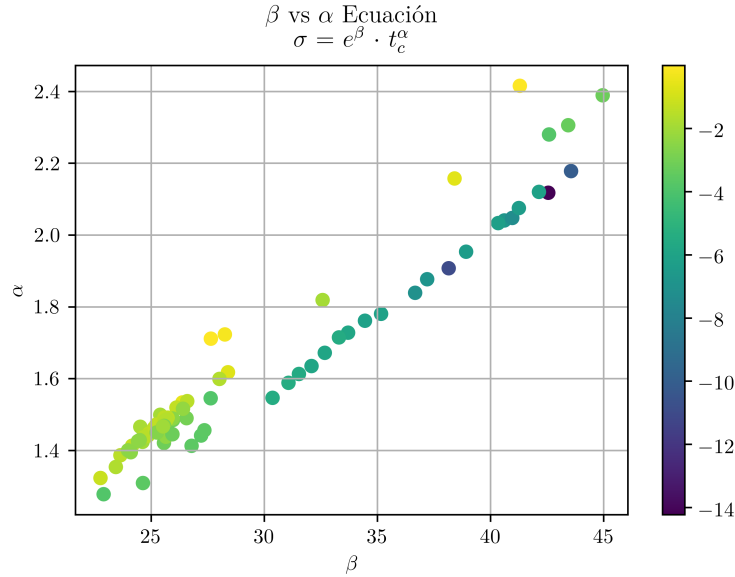


Figura 6.11: Relación entre α y β con 75 moléculas candidatas. Promedio de energía de solvatación usando BEM se resalta en barra de color

En este caso resulta tentador visualizar este gráfico con una cantidad mayor de moléculas, ya que se observan conjuntos de datos agrupados por el promedio de la energía libre.

Siguiendo con el análisis de par de variables, para el caso de t_c^* vs σ_{SASA} mostrado en la figura 6.12, tiene lógica que al ser el tiempo característico mayor, los átomos tendrán un volumen de probabilidad de reubicación mayor, modificando en gran parte el área de la cavidad formada en el solvente.

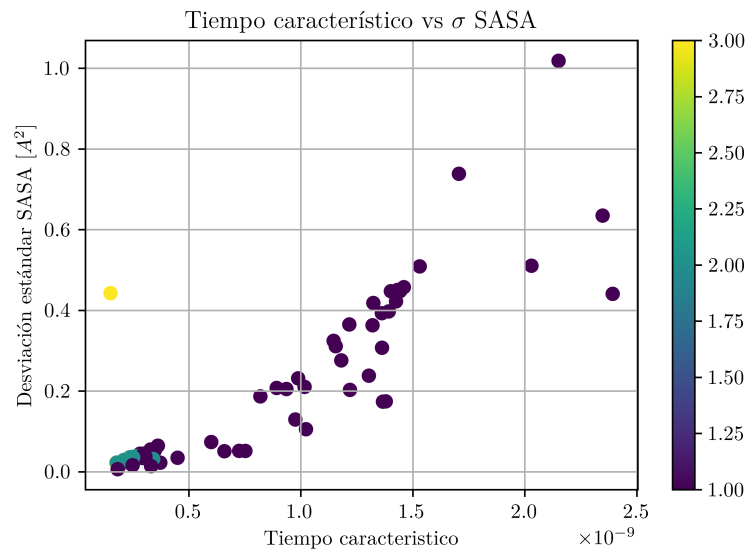


Figura 6.12: Dispersión de tiempo característico vs desviación estándar del área de las superficies accesibles por el solvente (SASA). Grupos funcionales resaltados en barra de colores

En este caso, llama la atención considerar esta relación para poder definir una aproximación del tiempo característico en función de la desviación estándar del área de la cavidad $t_c = t_c(\sigma_{SASA})$, considerando que el costo computacional de calcular σ_{SASA} es mucho menor que para el cálculo de t_c^* .

En el caso de la relación 3-12, modificamos β por α y su relación se muestra en la figura 6.13. Es interesante destacar el clúster formado por el grupo funcional de alcanos y sus derivados. Es posible replantear la función objetivo de la tabla de correlación por algún tipo de métrica que indique distancias promedios intra-clusters o cantidad de conjuntos encontrados, así como aumentar el número de moléculas probadas para dar más densidad a los conjuntos.

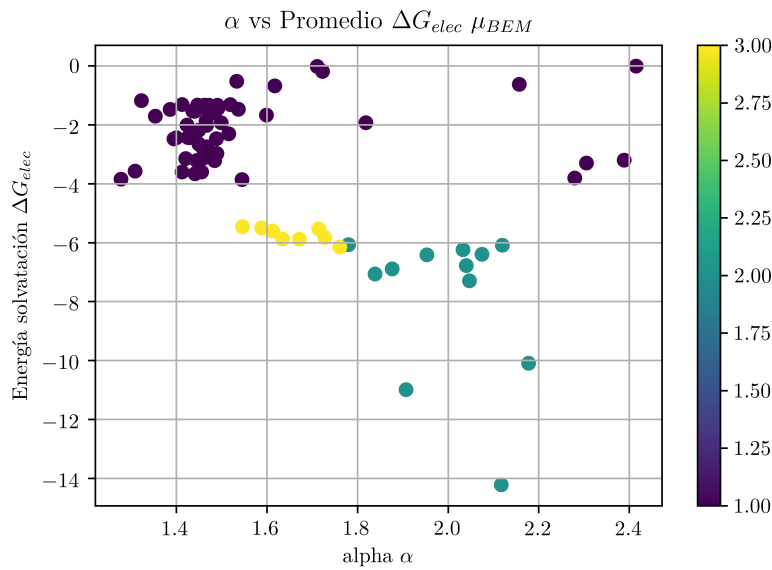


Figura 6.13: Dispersión de datos de α y promedio de energía de solvatación μ_{BEM} . Grupos funcionales resaltados en color

Por último, la tabla de correlación indica que existe una relación negativa entre el promedio de la energía de solvatación y su desviación estándar. En este caso es bueno definir el parámetro CV como *coeficiente de variación* de la forma

$$CV = \frac{\sigma}{|\mu|} \cdot 100 \% \quad (6.6)$$

De este modo al calcular el coeficiente de variación para las 75 moléculas se llega a un promedio de 0.5 %.

7 | Conclusiones

Es posible cuantificar la incertidumbre y obtener una desviación estándar usando el método Monte Carlo para modelar aleatoriamente la posición de los átomos de una molécula. Es sabido que a mayor cantidad de pruebas de esta simulación la energía de solvatación tiende a un valor y su incertidumbre decrece al mismo tiempo que el costo de cálculo aumenta, observándose que para este caso en particular 100 pruebas es un número aceptable para el propósito.

La física a nivel molecular es aún una incógnita, por esto las distribuciones de probabilidad que modelan el movimiento de átomos deben ser revisadas con mayor detenimiento. Alternativas a este dilema es considerar una distribución normal o alimentar el modelo implícito usando dinámica molecular.

Para la agitación esférica a radio constante observamos buenos resultados para el cálculo de desviación estándar, además de observar que la energía de solvatación muestra tendencias de distribución normal por lo que es posible crear una metodología de cálculo de intervalos de confianza utilizando como por ejemplo t-student. De igual manera vemos que la desviación estándar aumenta al incrementar los radios de agitación, siendo concluyente que la resolución en la medición de las posiciones de los átomos repercute en la incertidumbre de la energía calculada.

En el caso de la agitación que considera la conectividad entre átomos, verificamos su funcionamiento y su uso puede extrapolarse a moléculas con mayor cantidad de átomos mientras los parámetros sean pequeños (sobre todo para el caso del ángulo). Por otra parte, su validación con la dinámica molecular no fue posible, no significando que esta metodología deba desecharse. También es posible seguir trabajando este modelo cambiando algunos de sus parámetros en base a los mapas de densidad electrónica.⁵

La agitación esférica con radios en función de la masa y velocidad térmica de cada átomo presenta los resultados más prometedores al haber sido verificados y validados con dinámica molecular. Ideas para un trabajo a futuro son considerar distribuciones de probabilidad basados en

⁵Resultados directos de la cristalografía por rayos X https://proteopedia.org/wiki/index.php/Electron_density_maps

la distribución de probabilidad de velocidades de Maxwell-Boltzmann. En este caso la posición de un átomo estaría sujeta a una probabilidad normal teniendo un poco más de sentido físico.

Los radios de las esferas donde se agitaron los átomos son menores a la distancia promedio entre átomos, la longitud térmica promedio de mayor magnitud encontrada (para el átomo de hidrógeno) es aproximadamente un cuarto de la distancia entre átomos o distancia entre enlaces. Con esto podemos afirmar que la conectividad no es tan importante en el objetivo de comparar incertidumbres de modelos explícito e implícito.

La búsqueda de correlaciones se verifica como un método manual rápido para trabajar una gran cantidad de datos que consideran múltiples variables. Bajo este contexto proponemos seguir experimentando funciones objetivo distintas al coeficiente de correlación de Pearson, además de filtrar la búsqueda por grupos funcionales o otro tipo de categoría.

Para continuar entregando un valor de incertidumbre proponemos trabajar a futuro en intervalos de confianza. En este caso creemos que no basta con suponer que la energía de solvatación tiene una distribución normal, por lo que sería correcto crear una métrica y una metodología que permita filtrar y clasificar gran cantidad de resultados según pruebas de normalidad, curtosis y oblicuidad.

Bibliografía

- [1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. 3, 4
- [2] T. Betcke and M. W. Scroggs. Bempp-cl: A fast python based just-in-time compiling boundary element library. *Journal of Open Source Software*, 6(59):2879, 2021. 3.1
- [3] C. D. Cooper and J. P. Bardhan. A simple electrostatic model for the hard-sphere solute component of nonpolar solvation. *arXiv preprint arXiv:2005.13019*, 2020. (document), 2.4
- [4] C. D. Cooper Villagran. Biomolecular electrostatics with continuum models: a boundary integral implementation and applications to biosensors, 2015. (document), 2.2, 2.3
- [5] S. Decherchi and W. Rocchia. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PloS one*, 8(4):e59744, 2013. 3.1
- [6] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(suppl_2):W665–W667, 07 2004. 3
- [7] P. A. Fernández Martínez. Implementación y comparación de formulaciones integrales de frontera para el cálculo de energías de solvatación en macromoléculas disueltas en un medio salino. 2019. (document), 2.2, 3.2, 3.1
- [8] R. A. Huencho Mora. Comparación de energías de solvatación calculadas con poisson-boltzmann y dinámica molecular para moléculas pequeñas. 2021. 3.4, 5.2
- [9] R. M. Levy and E. Gallicchio. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annual review of physical chemistry*, 49(1):531–567, 1998. 2.4
- [10] A. D. McNaught, A. Wilkinson, et al. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997. 2.4
- [11] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *Journal of chemical theory and computation*, 5(2):350–358, 2009. 3.4, 5, 5.1, 5.2, 6, 6.2
- [12] D. L. Mobley and J. P. Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014. 3.4, 5, 5.2
- [13] V. N. Ramm Montiel. Estimación de error en la malla superficial para la ecuación de poisson boltzmann aplicado al método de elementos de borde. 2019. (document), 2.1

- [14] R. Salomon-Ferrer, D. A. Case, and R. C. Walker. An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013. 3, 4.1
- [15] M. F. Sanner, A. J. Olson, and J.-C. Spehner. Fast and robust computation of molecular surfaces. In *Proceedings of the eleventh annual symposium on Computational geometry*, pages 406–407, 1995. 3.1
- [16] S. Search. bem_electrostatics python library. https://github.com/SDSearch/bem_electrostatics, 2020. 3.2
- [17] S. D. Search Toloza. Comparación de formulaciones de la ecuación de poisson-boltzmann aplicando el método de elementos de frontera. 2021. 3.2
- [18] X. Yang, H. Lei, P. Gao, D. G. Thomas, D. L. Mobley, and N. A. Baker. Atomic radius and charge parameter uncertainty in biomolecular solvation energy calculations. *Journal of chemical theory and computation*, 14(2):759–767, 2018. 1