**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**Maestría y Doctorado en Ciencias Bioquímicas**

# ANÁLISIS DEL IMPACTO QUE TIENE EL USO DE DIVERSOS INDICADORES DE CALIDAD SOBRE LA RESOLUCIÓN DE FASES POR REEMPLAZO MOLECULAR Y EL AFINAMIENTO DE ESTRUCTURAS CRISTALOGRÁFICAS.

## CASO PARTICULAR: M271

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Maestría en Ciencias


PRESENTA:

THANIA QUIROZ HERNÁNDEZ

TUTOR PRINCIPAL

DR. ENRIQUE RUDIÑO PIÑERA INSTITUTO DE BIOTECNOLOGÍA, U.N.A.M.

MIEMBROS DEL COMITÉ TUTOR

DRA. LILIAN GONZÁLEZ SEGURA, FACULTAD DE QUÍMICA, U.N.A.M.

DRA. ADELA RODRÍGUEZ ROMERO, INSTITUTO DE QUÍMICA, U.N.A.M.


CIUDAD DE MÉXICO, FEBRERO, 2022

El presente trabajo se realizó en las oficinas del Departamento de Medicina Molecular y Bioprocesos del Instituto de Biotecnología de la UNAM, bajo la tutoría del Dr. Enrique Rudiño Piñera. Durante la maestría se contó con

## AGRADECIMIENTOS

Por el apoyo institucional:

A la Universidad Nacional Autónoma de México

Al Instituto de Biotecnología

Al CONACYT por la beca otorgada para la realización de esta tesis.

Al Dr. Enrique Rudiño Piñera, por su apoyo, enseñanzas y dirección.

A los miembros del H. Jurado.

A los miembros del Comité Tutor.

Al personal técnico, administrativo y de servicios de la UNAM por la ayuda brindada.

¡GRACIAS!

## DEDICATORIA

A familiares y amigos.

Inserta tu nombre aquí:

_____

Gracias por ser una estrella brillando en el cielo de mi vida.

¡MUCHAS GRACIAS!

**INDICE**

RESUMEN ..... 9
ABSTRACT ..... 10
1.INTRODUCCIÓN ..... 11
**1.2** DETERMINACIÓN DE LA ESTRUCTURA DE PROTEÍNAS POR CRISTALOGRAFÍA DE RAYOS X ..... **13**
1.2.1 PURIFICACIÓN DE UNA PROTEÍNA ..... 13
1.2.3 PROCESAMIENTO DE DATOS ..... 14
1.3 INDICADORES A LO LARGO DEL PROCESO DE LA DETERMINACIÓN DE UNA ESTRUCTURA TRIDIMENSIONAL POR RAYOS X ..... 15
2. ANTECEDENTES ..... 23
2.2 SUSTITUCIÓN Y REEMPLAZO MOLECULAR ..... 25
3.1 HIPÓTESIS. ..... 26
3.2 OBJETIVO GENERAL ..... 26
3.3 OBJETIVOS ESPECÍFICOS ..... 26
4. METODOLOGÍA ..... 27
4.1 CASO DE ESTUDIO ..... 27
4.2 MODELOS CRISTALOGRÁFICOS PARA RESOLVER EL PROBLEMA DE FASES ..... 28
4.3 PROCESAMIENTO DE DATOS DE DIFRACCIÓN ..... 34
5. RESULTADOS ..... 38
5.1 ANALISIS DE INDICADORES DE CALIDAD: ESCALAMIENTO Y REDUCCIÓN DE DATOS ..... 38
5.2 RESOLUCIÓN DE FASES Y AFINAMIENTO ..... 41
5.3 ANALISIS DE LA CALIDAD MODELO TRIDIMENSIONAL Y MAPA DE DENSIDAD ELECTRÓNICA ..... 51
5.3.1 ESCENARIO DE SUSTITUCIÓN MOLECULAR ..... 51
5.4 RMSD ..... 60
6. CONCLUSIONES ..... 62
7. PERSPECTIVAS ..... 63
ANEXO A. SECUENCIAS UTILIZADAS PARA ALINIAMIENTO DE LAS ESTRUCTURAS DE PROTEÍNAS PARA RESOLVER EL PROBLEMA DE FASES ..... 64

Anexo B. Gráficos de indicadores ..... 66
Anexo C. Descripción del caso de estudio:M271. 100% de identidad ..... 75
Anexo D. Estructuras finales sobrelapadas con el modelo final de la proteína M271 ... 77

**INDICE DE TABLAS**

Tabla A. Índices de la etapa de escalamiento y reducción de datos. ..... 15
Tabla B. Índices de la etapa de resolución de fases. ..... 18
Tabla C. Índices de la etapa de afinamiento. ..... 19
Tabla D. Valores de los índices de la etapa de escalamiento y reducción de datos. .... 20
Tabla E. Valores de los índices de la etapa de resolución de fases. ..... 22
Tabla F. Valores de los índices de la etapa de afinamiento. ..... 23
Tabla 4.0. Información cristalográfica de la proteína M271 ..... 27
Tabla 4.1. Tabla cristalográfica de los modelos base utilizados para resolución de fases por reemplazo molecular en esta tesis ..... 29
Tabla 4.2. Valores de integridad y redundancia para cada escenario ..... 34
Tabla. 4.3.Valores de los indicadores de calidad utilizados en este proyecto. ..... 36
Tabla 5.1 Valores de corte. Valores de los indicadores de corte utilizados en este proyecto. ..... 38
Tabla 5.2. Resoluciones de corte utilizando el valor límite inferior para cada indicador de los analizados en esta tesis ..... 40
Tabla 5.3. Escenarios por grupos a partir de los cuales ya no fue posible resolver fases ..... 43

**INDICE DE FIGURAS**

Figura1. Número de estructuras publicadas anualmente en el PDB hasta el año 2021[7]. ..... 12

Figura 4.1. Modelo listón de las estructuras de las proteínas seleccionadas para resolver el problema de fases en esta tesis. ..... 31

Figura 4.2. Alineamiento de las secuencias de aminoácidos de las distintas estructuras de las proteínas seleccionadas utilizadas para la resolución de fases en esta tesis. ... 33

Figura 4.1 Metodología. ..... 37

Figura 5.1. $R_{merge}$ vs resolución. ..... 39

Figura 5.2. Frecuencia de resolución de corte. ..... 41

Figura 5.3. LLG en escala logarítmica contra resolución. ..... 42

Figura 5.4. FTZ en escala logarítmica contra resolución. ..... 45

Figura 5.5. Rwork de afinamiento de cuerpo rígido contra resolución. ..... 46

Figura 5.6. Rfree de afinamiento de cuerpo rígido contra resolución. ..... 47

Figura 5.7. ΔR de afinamiento de cuerpo rígido contra resolución. ..... 48

Figura 5.8. $R_{work}$ de afinamiento restringido contra resolución. ..... 49

Figura 5.9. $R_{free}$ de afinamiento restringido contra resolución. ..... 50

Figura 5.11. Valores locales de CC para dos escenarios extremos. ..... 54

Figura 5.12. Diagrama de Venn de distribución de escenarios que cumplen con criterios de calidad de mapa-modelo y afinamiento ..... 56

Figura 5.13 Porcentaje de escenarios desechados por indicador de corte. ..... 57

Figura 5.14. Diagrama de Veen de distribución de escenarios que cumplen con criterios %aa útiles ≥35% y $R_{free}$≥0.5 ..... 58

Figura 5.15. Porcentaje de aminoácidos útiles versus valores de $R_{free}$. ..... 59

Figura 5.16 Porcentaje de escenarios desechados por indicadores de corte. ..... 60

Figura 5.17. RMSD. ..... 61

Figura 5.18. Carbonos alfa. ..... 62

Figura B.1. $R_{meas}$ vs Resolución. ..... 66

Figura B.2. $R_{pim}$ vs Resolución. ..... 67

Figura B.3. $I/\sigma(I)_{mean}$ vs Resolución. ..... 68

Figura B.4. $I/\sigma(I)$ vs Resolución. ..... 69

Figura B.5. CC½ vs Resolución. ..... 70

Figura B.6. Integridad vs Resolución..... 71
Figura B.7. Redundancia vs Resolución. ..... 72
Figura B-8. LLG en escala logarítmica contra resolución. ..... 73
Figura B-9. FTZ en escala logarítmica contra resolución. ..... 74

**RESUMEN**

En el presente estudio se utilizó información cristalográfica de M271, un inhibidor de proteasas asparricas y serinicas aislado de la papa y cuya estructura tridimensional fue reportada por Campuzano[1], los datos cristalográficos usados en esta tesis, consisten en un *dataset* de 240 *frames*, obtenidos por difracción de rayos X a partir de un cristal de la proteína M271, los cuales se sometieron al proceso de determinación de estructuras tridimensionales por reemplazo molecular. En la etapa de escalamiento y reducción de datos se utilizaron conjuntos consecutivos de *frames* (240, 120, 60 y 30), estableciendo diferentes resoluciones de corte; empezando con la máxima resolución de 1.6 Å e incrementando 0.05 Å hasta reducir la resolución a 2 Å. Y en la etapa de resolución de fases se utilizaron 5 estructuras de proteínas para resolver el problema de fasescon diferente porcentaje de identidad obtenidos del PDB (5DZU, 3TC2,5DVH Y 1AVW), dando origen a 180 escenarios de los cuales 162 dieron lugar a fases que permitieron construir modelos cristalográficos utiles. Durante la etapa de reducción y escalamiento de datos se suele elegir la resolución de corte y para ello, la revisión bibliográfica realizada en esta tesis permitió evidencia el uso de distintas métricas y valores límite: $R_{merge} \leq 0.8$, $R_{meas} \leq 0.8$, $R_{pim} \leq 0.5$, CC ½ ≥ 0.15, Integridad ≥ 70, Redundancia ≥ 2, I/σ(I) ≥ 2, I/σ(I) mean ≥ 2. También se analizó el efecto sobre métricas que permiten validar la calidad de los modelos cristalográficos obtenidos en los escenarios que dieron lugar a fases útiles para construir un modelo ($CC_{Global} > = 0,8$, $R_{work} \leq 0.5$, $R_{free} \leq 0.5$ y $\Delta R \leq 0.05$). El análisis producto de esta tesis, mostró que I/σ es una métrica de corte que desecha la mayor cantidad de datos y escenarios que pueden dar lugar a estructuras útiles (donde un escenario útil es aquel con valores $R_{free} < 0.5$, $R_{work} < 0.5$, $\Delta R > 0.05$, $CC > 0.8$), por lo que su uso en la determinación de la resolución máxima en estructuras producto de difracción de rayos X debe ser descontinuada en favor de métricas como CC1/2.

## ABSTRACT

In the present study, crystallographic information from M271, an inhibitor of aspartic and serine proteases isolated from potato and whose three-dimensional structure was reported by Campuzano1, was used. The crystallographic data from M271 consists of a dataset of 240 frames, obtained by diffraction X-rays from a single crystal. These data were used to determine three-dimensional structures by molecular replacement en¿mploying different scenarios. In the data scaling and data reduction stages, consecutive sets of frames (240, 120, 60, and 30) were used, establishing different highest resolutions from 1.6 Å, and increasing by 0.05 Å, until the resolution reaches 2 Å. In the phase resolution stage, five protein structures were used to solve the phase problem with different sequence identity models obtained from the PDB (5DZU, 3TC2, 5DVH, and 1AVW). Giving rise to 180 scenarios, of which 162 gave rise to phases that allowed us to build correct crystallographic models. The cut-off resolution is usually chosen during the data reduction and scaling stages. A bibliographic review carried out in this thesis allowed evidence of the use of different metrics and limit values: Rmerge ≤ 0.8, Rmeas ≤ 0.8, Rpim ≤ 0.5, CC ½ ≥ 0.15, Integrity ≥ 70, Redundancy ≥ 2, I/σ(I) ≥ 2, I/σ(I) mean ≥ 2. The effect on metrics that allow validating the quality of the crystallographic models obtained in the scenarios that gave rise to valuable phases to build a model ($CC_{Global}I> = 0.8$, $R_{work} \leq 0.5$, $R_{free} \leq 0.5$ and $\Delta R \leq 0.05$). The analysis resulting from this thesis showed that I/σ is a cut-off metric that discards the most significant amount of data and scenarios that can give rise to valuable structures. So its use in determining the maximum resolution in X-ray diffraction product structures should be discontinued in favor of metrics such as CC1/2.

## 1.INTRODUCCIÓN

Los modelos tridimensionales de proteínas permiten realizar análisis a alta resolución con alcances que van desde los procesos evolutivos que dieron lugar a una conformación particular, hasta propuestas sobre las determinantes estructurales que les confieren características bioquímicas especiales. De hecho las estructuras moleculares de proteínas son puntos de partida importantes para elucidar la función de éstas a nivel molecular[2], tales como: defensa ,transporte ,comunicación, almacenaje, enzimas.

Por lo tanto, conocer las estructuras tridimensionales de proteínas permite analizar su arquitectura y comprender su función, y además proponer mecanismos de cómo llevan a cabo dicha función, diseñar experimentos de mutagénesis para comprobar estos mecanismos, diseñar fármacos que inhiban ciertas funciones, o incluso, comprender los mecanismos implicados en nanomáquinas como las ATP sintasas, entre otras [3,4].

Las estructuras tridimensionales de las proteínas reportadas están determinadas principalmente a través del uso de tres técnicas experimentales, que son la cristalografía de rayos X de cristal único, la RMN, resonancia magnética nuclear (NMR, por sus siglas en inglés) y la Crio-ME, crio-microscopía electrónica del alta resolución (Cryo-EM, por sus siglas en inglés)[5]. En la **Tabla1.1** se muestran las características principales de estas técnicas.

**Tabla1.1. Comparación de las técnicas de cristalografía de rayos X, RMN y Crio-ME.** Modificada de Biostructure.[6]

| | Cristalografía de rayos X | RMN | Crio-ME |
|---|---|---|---|
| Ventajas | - Estructuras de alta resolución.<br>-Amplio intervalo de peso molecular.<br>-Facilidad de construcción de modelos. | - Alta resolución.<br>-Estructura 3D en solución.<br>-Ideal para realizar estudios de dinámica. | -Fácil preparación de la muestra.<br>-Estructura en estado nativo.<br>-Tamaño de muestra pequeño. |
| Desventajas | -Dificultad para la cristalización.<br>-Dificultades para la difracción.<br>-Estructura mayoritariamente estática y en estado cristalino. | - Muestra de alta pureza | - Resolución relativamente baja.<br>-Aplicable por el momento sólo a moléculas de alto peso molecular. |

| | | | |
|---|---|---|---|
| Muestras | - Muestras cristalizables.<br>-Proteínas solubles, de membrana, ribosomas, ADN/ARN, complejos de proteínas. | -Biomoléculas con masas moleculares máximos de hasta 40-50 kDa.<br>- Moléculas solubles en agua. | Biomoléculas con masas moleculares superiores a 150 kDa<br>-Viriones, proteínas de membrana, proteínas grandes, ribosomas, complejos de varias proteínas. |
| Resolución | Alta | Relativamente alta | Relativamente baja (> 1.22 Å) |

La mayoría de los modelos tridimensionales de proteínas reportadas en el *Protein Data Bank* (PDB) están determinados por cristalografía de rayos X, como se muestra en la **Figura 1,** representando en conjunto casi 25 veces más estructuras que las determinadas por la técnica de RMN y tres veces más que para Crio-ME.

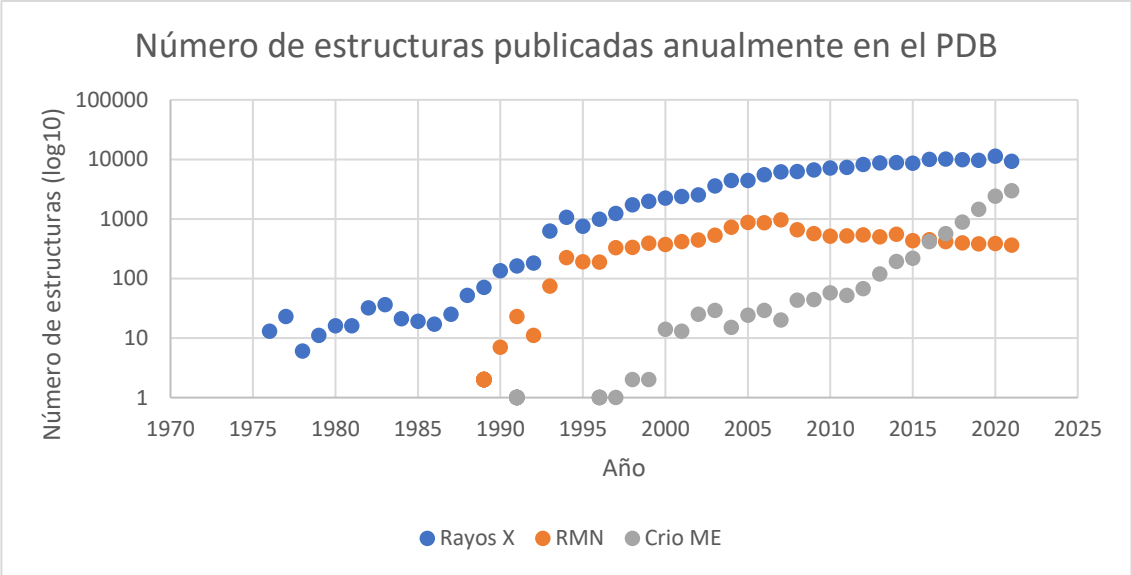



**Figura1. Número de estructuras publicadas anualmente en el PDB hasta el año 2021[7].**

## 1.2 DETERMINACIÓN DE LA ESTRUCTURA DE PROTEÍNAS POR CRISTALOGRAFÍA DE RAYOS X

A partir de los años 50 del siglo pasado, se comenzaron a utilizar los rayos X para la determinación de estructuras tridimensionales de proteínas. En 1958, John Kendrew y colaboradores [8], determinaron experimentalmente, la primera estructura de una proteína, la mioglobina. Desde entonces el número de estructuras de proteínas determinadas por rayos X ha aumento hasta llegar, en enero del 2022 hasta cerca de 162081 estructuras reportadas [7].

A continuación, se describe brevemente el proceso general, para la determinación de estructuras cristalográficas de proteínas por la técnica de cristalografía de rayos X.

### 1.2.1 PURIFICACIÓN DE UNA PROTEÍNA

El primer paso consiste en extraer a la proteína de interés de algún organismo que la contenga o produzca, de forma natural o no, como es el caso del uso de cepas modificadas para expresión de la proteína de interés. Una vez que se tiene el extracto con la proteína, este se somete a procesos de purificación tales como:

- Cromatografías (de tamaño molecular, de interacciones iónicas o especificas o hidrofóbicas, entre otras)
- Diálisis
- Desnaturalización reversible con sales (con temperatura o con agentes caotrópicos).
- Electroforesis.
- Precipitación con sales (temperatura o cambios de pH)

### 1.2.2 CRISTALIZACIÓN

Una vez purificada la proteína y comprobado que se trata de una especie homogénea (mediante SDS-PAGE, geles nativos o estudios biofísicos como la Dispersión Dinámica de la Luz –DLS-) se deben realizar pruebas de cristalización.

La cristalización en sí es otro proceso de purificación, donde las moléculas se ordenan de forma reticular y de manera repetitiva. Sin embargo, existen varios reportes donde aclaran que entre más pura y homogénea sea la muestra a cristalizar, mayores posibilidades se tienen de lograr cristales con el tamaño y características necesarias para difractar los rayos X [9].

Con el fin de encontrar las condiciones que den lugar a un cristal, comúnmente se preparan soluciones, que contienen a la proteína de interés en condiciones solubles, pero próxima a precipitar, y se mezclan con agentes precipitantes (comúnmente llamados *Kits*

de cristalización), diversas soluciones amortiguadoras o incluso solventes no polares, cambios de temperatura, gradientes de pH, entre otros factores[10].

Lo anterior, con el propósito de generar condiciones adecuadas para la cristalización de la proteína deseada. Para ello se deben realizar tantas pruebas de cristalización como sean necesarias y de ahí la necesidad de contar con grandes cantidades de proteína con el fin de explorar la mayor cantidad posible de condiciones, y de esta manera aumentar la probabilidad de encontrar cristales susceptibles de difractar a los rayos X.

Existen distintos métodos para lograr que una proteína cristalice. Algunos se basan en procesos difusivos, como el de la gota colgante o el de la gota sentada. El proceso para identificar una condición de cristalización puede hacerse tanto de forma manual, como automatizada, usando robots, la cual permite probar distintas condiciones en tiempos y volúmenes reducidos.

Una vez que se obtiene un cristal con caras definidas y proveniente de un solo núcleo de cristalización, el siguiente paso para la obtención de su estructura es someterlo a un haz de rayos X, en un sincrotrón o en un equipo de ánodo rotatorio. Si el cristal es de la calidad adecuada, se obtendrán patrones de difracción de tal calidad que permitirán determinar la estructura tridimensional de una proteína [11–13].

Una vez que se realizó la colecta de datos en algún equipo de difracción de rayos X, éstos se someterán a un tratamiento de datos que permitirá la obtención de la estructura tridimensional de la proteína.

### 1.2.3 PROCESAMIENTO DE DATOS

A continuación, se muestran los pasos a seguir para la determinación de una estructura cristalográfica.

I. Indexación del patrón de difracción. Proceso en el cual se determinan los índices[A] de las reflexiones [14].
II. Integración de las intensidades. Proceso en el cual se calcula la intensidad total de cada reflexión. La intensidad integrada es el área de la curva bajo el pico que está por encima del fondo [15].
III. Escalamiento. Se calculan factores de escala para que las intensidades; se representen en una misma escala. Estos factores se ajustan para que la discrepancia residual entre las diversas mediciones de la misma reflexión o las relacionadas con la simetría del cristal, sea la mínima posible.[16]
IV. Problema de fases. Puede definirse como el problema de determinar las fases de los factores de estructura normalizados, cuando solo se dan las magnitudes [17,18].

[A] Son un juego de tres números que permiten identificar unívocamente un sistema de planos cristalográficos.

En la difracción con rayos X sólo se miden las magnitudes de la difracción, y no las fases. Las fases [B]contienen la mayor parte de la información estructural. Las fases se deben derivar de otra manera; usando remplazo isomórfico múltiple, remplazo molecular o dispersión anómala [19].

V. Afinamiento. Es la optimización de una función de un conjunto de observaciones, cambiando los parámetros de un modelo [16].

## 1.3 INDICADORES A LO LARGO DEL PROCESO DE LA DETERMINACIÓN DE UNA ESTRUCTURA TRIDIMENSIONAL POR RAYOS X

Esta información se muestra en las tablas: A, B y C, donde se definen algunos índices necesarios para evaluar el proceso de construcción de un modelo de la estructura cristalográfica de una proteína, que se analizaron en este proyecto de investigación.

**Tabla A. Índices de la etapa de escalamiento y reducción de datos.**

| Escalamiento y reducción de datos | |
|---|---|
| | Descripción |
| $R_{merge}$ | También llamado $R_{sym}$ o $R_{intm}$.<br>Ha sido ampliamente utilizado en el pasado para determinar el corte máximo en la resolución[C]. Cuanto mayor es la multiplicidad, mayor es su valor.<br>$$R_{merge} = R_{sym} = R_{linear} = \frac{\sum_{hkl} \sum_i \lvert I_i(hkl) - \langle I(hkl) \rangle \rvert}{\sum_{hkl} \sum_i I_i(hkl)}$$<br>Donde<br>$\mathbf{I_i(hkl)}$ es la intensidad de una reflexión individual con índices (hkl).<br>$\langle$ **I(hkl)** $\rangle$ Valor promedio de todas las reflexiones dentro de los índices (hkl), incluidas las que son equivalentes por simetría. [14,20,21] |

[B] Fase es una medida de la diferencia de tiempo entre dos ondas senoidales.
[C] Resolución máxima de una estructura tridimensional de proteína.

| | |
|---|---|
| $R_{meas}$ | También llamado $R_{rim}$, evalúa la precisión de la intensidad individual y es independiente de la multiplicidad, por lo que muestra cuán amplia es la distribución de intensidades. Útil para evaluar datos de baja resolución.<br><br>$$R_{meas} = R_{rim} = \sqrt{\sum_{hkl} \frac{N(hkl)}{(N(hkl)-1)}} = \frac{\sum_i \lvert I_i(hkl) - \langle I(hkl)\rangle \rvert}{\sum_{hkl} \sum_i I_i(hkl)}$$<br><br>**$I_i$(hkl)** es la intensidad de una reflexión individual con índices (hkl).<br><br>**⟨ I(hkl) ⟩** Valor promedio de todas las reflexiones dentro de los índices (hkl), incluidas las que son equivalentes por simetría<br><br>**N(hkl)** Número de reflexiones con índices (hkl)<br><br>14,21,22 |
| $R_{pim}$ | No se recomienda para determinar la resolución de corte, ya que su valor se eleva hacia el infinito a medida que disminuyen las intensidades. Debido a que es una suma no ponderada, si se combinan datos de difracción de calidad variable, se subestimará la calidad de los datos globales.<br><br>$$R_{pim} = \sqrt{\sum_{hkl} \frac{1}{(N(hkl)-1)}} \times \frac{\sum_i \lvert I_i(hkl) - \langle I(hkl)\rangle \rvert}{\sum_{hkl} \sum_i I_i(hkl)}$$<br><br>**$I_i$(hkl)** es la intensidad de una reflexión individual con índices (hkl).<br><br>**⟨ I(hkl) ⟩** Valor promedio de todas las reflexiones dentro de los índices (hkl), incluidas las que son equivalentes por simetría<br><br>**N(hkl)** Número de reflexiones con índices (hkl)<br><br>14,21,23 |
| (I)/σ(I) mean | Es el cociente entre la intensidad media y el intervalo de desviación estándar promedio de las intensidades en un experimento de difracción. |
| (I)/σ(I) | Relación señal-ruido promedio de las intensidades combinadas en función de la resolución.[24] |

| | |
|---|---|
| CC ½ | Coeficiente de correlación entre las intensidades de reflexión de dos mitades de un conjunto de datos dividido aleatoriamente.<br><br>$$CC_{1/2} = \frac{\sum_{i=1}^{n}(x - \langle x\rangle)(y - \langle y\rangle)}{\sqrt{\sum_{i-1}^{n}(x - \langle x\rangle^2)}\sqrt{\sum_{i-1}^{n}(y - \langle y\rangle)^2}}$$<br><br>Donde<br><br>**x y** subconjuntos<br><br>**⟨ x ⟩** Valor promedio de x<br><br>**⟨ y ⟩** Valor promedio de y<br><br>21,22,25 |
| Integridad | Es el número de reflexiones cristalográficas medido en un conjunto de datos, expresado como un porcentaje del número total de reflexiones presentes en la resolución especificada, para una celda y grupo espacial específico.<br><br>26,27 |
| Redundancia | Se define como el número de observaciones independientes (después de la combinación de reflexiones parciales[D]) por reflexión única, en el conjunto final de datos combinados y de simetría reducida.[E]<br><br>27 |

[D] Son reflexiones que tienen solo información parcial del plano del que difractan.

**Tabla B. Índices de la etapa de resolución de fases.**

| | Resolución de Fases |
|---|---|
| | Descripción |
| LLG | *Log Likelihood Gain*, en español, ganancia logarítmica de probabilidad, es la diferencia entre la probabilidad del modelo y la probabilidad calculada a partir de una distribución en específico [18].<br><br>Si el LLG es negativo, significa que el modelo es peor que una colección de átomos aleatorios. El LLG siempre debe ser positivo, lo que significa que es probable que el modelo prediga los datos experimentales. [28].<br><br>La función de verosimilitud expresa la probabilidad de que ocurran valores de parámetros correctos, dados los datos observados. [29].<br><br>Log Likelihood Gain:<br><br>$$LLG_1 = log_{p1} - log_{p0}$$<br><br>Donde<br><br>$p_1$ es la probabilidad del modelo<br><br>$p_0$ es la probabilidad calculada a partir de una distribución aleatoria |
| FTZ | La relación señal-ruido se juzga utilizando la puntuación Z, que se calcula comparando los valores de LLG de la búsqueda de rotación o traslación con los valores de LLG para un conjunto de rotaciones o traslaciones aleatoria.<br><br>Es el número de desviaciones estándar sobre la media (puntuación Z) para la ganancia de *loglikelihood* en la intensidad (LLGI) en la función de traslación y rotación (TF) [30]. |

**Tabla C.** Índices de la etapa de afinamiento.

| | Afinamiento |
|---|---|
| $R_{work}$ | Es el valor R para todos los factores de estructura [F]calculados a partir de un modelo cristalográfico. La calidad del ajuste de los factores de estructura del modelo con los datos de difracción es generalmente evaluada con un factor R:<br><br>$$R = \frac{\sum_{hkl} ||F_{obs}| - |F_{calc}||}{\sum_{hkl} |F_{obs}|}.$$<br><br>Donde<br><br>hkl son los puntos de la red reciproca del cristal. Son llamados índices de Miller.<br><br>$F_{obs}$ Factor de estructura observado<br><br>$F_{cal}$ Factor de estructura calculado<br><br>Cuando el 90-95% de las reflexiones se utilizan, se denominan conjunto de trabajo y a partir de ellas se calcula $R_{work}$ [31]. |
| $R_{free}$ | Es el valor R, para un 5-10% de datos seleccionados aleatoriamente que no se utilizan en el afinamiento y por lo tanto constituyen un control en el proceso de afinamiento [32] Valores altos de Rfree pueden revelar errores en la determinación de la estructura [33]. |
| $\Delta(R_{free}- R_{work})$ | La diferencia entre Rfree y Rwork se utiliza como indicador de sobreajuste.<br><br>Esta diferencia siempre debe ser positiva ya que si los datos han sido procesados correctamente, $R_{free}$ siempre debe ser mayor que Rwork. Además, $R_{free}$ debería acercarse gradualmente a $R_{work}$ a medida que el afinamiento progresa [34]. |

[F] Determinan si en una dirección hkl se obtiene o no la correspondiente intensidad difractada. Este factor considera el ordenamientode los átomos dentro de la celda y la distribución de la densidad de carga alrededor de cada átomo presente en la celda unidad. Es variable para cada índice.

**LÍMITES DE LOS ÍNDICES REPORTADOS EN LA LITERATURA**.

A continuación, en tablas: D, E y F, se muestran los límites encontrados en la literatura para los índices mencionados en la sección anterior.

**Tabla D.** Valores de los índices de la etapa de escalamiento y reducción de datos.

| Valores de $R_{merge}$ en la literatura | |
|---|---|
| 0.3 – 0.5 | Valores típicos utilizados en el laboratorio. |
| 0.6 – 0.8 | Los datos se truncan típicamente a una resolución antes de que el valor de Rmerge supere ~ 0.6 a 0.8 [35]. |
| 0.3 – 0.4 (bajas simetrías) | En la capa de resolución más alta, se puede permitir que el $R_{merge}$ alcance el 0.3-0.4 para cristales de baja simetría [36]. |
| 0.6 (altas simetrías) | En la capa de resolución más alta, se puede permitir que el $R_{merge}$ alcance el 0.3-0.4 para cristales de baja simetría [36]. |
| **Valores de $R_{meas}$ en la literatura** | |
| 0.3 – 0.5 | Valores típicos usados en nuestro laboratorio. |
| 0.26 – 0.4 | Powell en 2017, propuso que los valores límite de los valores Rmeas eran aproximadamente 0.8/‹I/σ› [14]. Donde, el valor de I/σ(I) (es decir, el valor promedio) para cada capa de resolución de datos, suele ser de 2 a 3. |
| 0.6 – 0.8 | El límite máximo de resolución se establece antes de que el valor de $R_{meas}$ supere ~ 0.6 a 0.8 [35]. |
| **Valores de $R_{pim}$ en la literatura** | |
| 0.3 – 0.5 | Valores típicos utilizados en nuestro laboratorio. |
| 0.26 – 0.4 | Powell en 2017, propuso que los valores límite de los valores $R_{pim}$ eran aproximadamente 0.8/‹I/σ› [14]. Donde el valor de I/σ(I) (es decir, el valor promedio) para cada capa de resolución de datos, suele usar un valor de 2 a 3. |
| **Valores de (I)/σ(I)mean en la literatura** | |

| | |
|---|---|
| ≥ 2 | Valores típicos utilizados en nuestro laboratorio |
| > 3 | Regla triple para determinar el limite de resolución en la última franja: si la integridad >70% y el valor de $R_{symm}/R_{merge}$ < 30% entonces, el valor de I/σ en la última faja de resolución debe ser >3 [37]. |
| 2 – 3 | El valor de I/σ(I) (es decir, el valor promedio) para la franja de mayor resolución de datos, suele utilizar un límite mínimo de 2 o hasta de 3 [14]. |
| ≥ 2 | Por lo general, los datos se truncan a una resolución antes de que la relación empírica señal / ruido, caiga por debajo de ~ 2.0 [35]. |

| Valores de (I)/σ(I) en la literatura | |
|---|---|
| ≥ 2 | Valores típicos utilizados en nuestro laboratorio. |
| ≥2 ó ≥ 3 | El valor de I/σ(I) (es decir, el valor promedio) para la franja de mayor resolución de datos, suele usar un valor mínimo de 2 a 3 [14]. |

| Valores de CC ½ en la literatura | |
|---|---|
| 0.8 | Valores típicos utilizados en nuestro laboratorio. |
| > 0.15 | El CC½ es un caso especial del coeficiente de correlación de Pearson. La prueba t de Student se puede utilizar para indicar a qué resolución ya no existe una correlación significativa entre las magnitudes pertenecientes a reflexiones idénticas dentro de dos conjuntos de datos seleccionados entre el total de reflexiones colectados en un experimento dado (un valor mínimo de aproximadamente de 0.15, es común como límite de resolución máxima para conjuntos de datos grandes) [14]. |
| 0.5 , 0.75 | Las reglas empíricas actuales sugieren que se deben incluir datos con CC ½ superior a 0.5 en la franja de mayor resolución, aunque la propia experiencia del autor sugiere un límite de alrededor de CC½ ~ 0.75 [38]. |

| Valores de Integridad en la literatura | |
|---|---|
| > 85% | Valores típicos utilizados en nuestro laboratorio para un escenario de reemplazo molecular. |

| | |
|---|---|
| > 70% | Regla triple para determinar el límite de resolución en la última franja: si la integridad >70% y el valor de $R_{symm}/R_{merge}$ < 30% entonces, el valor de I/σ en la última franja de resolución debe ser >3 [37]. |
| > 90% | La integridad debe ser superior al 90% en la faja de mayor resolución externa [39]. |
| **Valores de Redundancia en la literatura** | |
| 2 | Valores típicos utilizados en nuestro laboratorio para un escenario de reemplazo molecular. |
| ≥ 2 | Una multiplicidad ≥ 2 se acepta de forma rutinaria para conjuntos de datos recopilados por medios convencionales [40]. |

**Tabla E.** Valores de los índices de la etapa de resolución de fases.

| | |
|---|---|
| **Valores de LLG en la literatura** | |
| + | Valores típicos utilizados en nuestro laboratorio. |
| > 40 | Un valor superior a 40 suele indicar una solución de reemplazo molecular correcta (al menos en ausencia de factores de complicación como la simetría traslacional no cristalográfica, tNCS). [41]. |
| + | La ganancia de probabilidad logarítmica (LLG) debe ser positiva y lo más alta posible [42]. |
| > 36 | Cuando el valor del LLG es > 36 posiblemente el reemplazo molecular ha funcionado [43]. |
| > 70 | Para considerar que el reemplazo molecular ha sido resuelto, se propone como criterio empírico un valor de LLG superior a 70 [44]. |
| **Valores de TFZ en la literatura** | |
| > 6 | La solución correcta en un reemplazo molecular, generalmente tendrá una puntuación Z (TFZ) superior a 6 [41]. Este estadístico se utiliza cuando se utiliza el software de Phaser. |

Este estadístico se utilizan cuando se utiliza el software de Phaser.

**Tabla F.** Valores de los índices de la etapa de afinamiento.

| Valores de $R_{work}$ en la literatura | |
|---|---|
| 0.2 | Valores típicos utilizados en nuestro laboratorio, si ya se adicionaron al modelo cristalográfico moléculas de agua y ligandos. |
| 0.25 | Valores típicos utilizados en nuestro laboratorio, si aún no se adicionaron al modelo cristalográfico moléculas de agua y ligandos. |
| 0.2 | $R_{work}$, por lo general debe estar alrededor del 20% (0.2 en valores de 0.0 a 1.0) al concluir el afinamiento [39]. |
| 0.2 – 0.5 | Los valores teóricos de $R_{work}$ van desde cero (concordancia perfecta de las intensidades calculadas y observadas), hasta aproximadamente 0.6 para un conjunto de intensidades medidas en comparación con un conjunto de intensidades aleatorias. Muchos modelos con $R_{work} \geq 0.5$ no responderán a los intentos de mejora en el afinamiento (valores de R work menores a 0.5 pueden dar lugar a un modelo correcto al final del proceso de afinamiento). Se considera que un valor de $R_{work}$ deseable para un modelo afinado de una proteína, con datos a 2,5 Å, es ∼0.2 [45]. |
| **Valores de ΔR en la literatura** | |
| ≈ 5 % | Valores típicos utilizados en nuestro laboratorio. |

## 2. ANTECEDENTES

A lo largo del desarrollo de la cristalografía de rayos X, se han propuesto diversos indicadores para delimitar la resolución máxima de los datos de difracción. Sin embargo, hay distintas opiniones en la literatura acerca de su uso, así como de los limites numéricos que deben aplicarse. Por ejemplo, se sabe que la calidad en los datos de difracción determina la precisión del modelo final [46] y para determinar cuántos y cuáles de los datos de difracción tienen la calidad suficiente, se han propuesto indicadores de calidad, tales como $R_{merge}$, $R_{meas}$ y CC½, entre otros, los cuales se utilizan para seleccionar la resolución máxima con el fin de lograr 'un mejor' modelo cristalográfico [22,34].

La elección de la resolución máxima no necesariamente debe realizarse una sola vez durante el estudio cristalográfico, puede realizarse en el procesamiento de datos o en el afinamiento del modelo [47]. Ello con el fin de no desechar datos que pudieran ser útiles. En este sentido, $R_{merge}$ es uno de los parámetros más utilizados, ya que deriva de la fusión de todas las mediciones de intensidad de una reflexión y sus equivalentes de simetría en un único valor [48]. Sin embargo, es intrínsecamente dependiente de la redundancia de los datos, por lo que una baja redundancia conduce a un $R_{merge}$ menor, dando como resultado datos menos precisos [23]. Además, este parámetro no se relaciona con valores de calidad del modelo afinado, como por ejemplo $R_{work}/R_{free}$ ,por lo que algunos autores no lo consideran adecuado para delimitar la resolución máxima en un modelo afinado [49]. Por ello se han propuesto otras métricas independientes de la redundancia como por ejemplo $R_{meas}$.

En el 2012, Karplus y Diederichs [35] demostraron que el CC ½ (coeficiente de correlación de Pearson) es un mejor indicador de la calidad y la resolución de los datos cristalográficos que otras medidas más tradicionales como $R_{merge}$. En contraposición, en 2017, Wang y colaboradores [50] señalaron que el CC ½ acumulativo es un indicador poco sensible a la calidad general y menos informativo que otros indicadores, en especial que $R_{merge}$. Lo cual demuestra las opiniones divergentes en varios artículos ya publicados.

En el 2013 Evans y colaboradores [24], indicaron que aún no se habían definido reglas aceptadas por toda la comunidad cristalográfica, para establecer el corte de máxima resolución en un conjunto de datos de difracción, ya que esto depende del uso que tendrán los datos, por lo que consideran un error aplicar prematuramente un corte severo en la resolución durante la etapa de reducción de datos. De hecho, concluyen que los datos siempre se pueden excluir más adelante, durante el afinamiento de los modelos cristalográficos.

Powell [14] recomienda utilizar todos los datos de difracción disponibles, y posteriormente recortar los límites de resolución al comparar el empalme entre los modelos cristalográficos y los mapas de densidad electrónica. Utilizando para lo anterior, la resolución en la que se alcanza un valor mínimo de $R_{free}$, lo cual se considera un límite razonable para definir la resolución máxima.

Las métricas globales de precisión del modelo de estructura (como $R_{free}$) no identifican errores locales en un modelo. Una mejor métrica de precisión local del modelo es la coherencia con la densidad electrónica en el espacio real. Esto supone que la densidad electrónica en sí misma y, por lo tanto, las fases de las que se deriva son precisas. Esta es una suposición razonable porque la validación basada en la densidad normalmente se realiza cerca de la finalización del proceso de afinamiento, cuando el modelo es mayormente correcto, y solo quedan por resolver una pequeña cantidad de errores en el modelo cristalográfico [51].

.

## 2.2 SUSTITUCIÓN Y REEMPLAZO MOLECULAR

El término "Reemplazo Molecular" (RM) se utiliza generalmente para describir el uso de un modelo molecular conocido para resolver la estructura cristalina desconocida de una molécula relacionada, o más correctamente, homologa estructural a la primera. El RM permite dar solución al problema de las fases cristalográficas proporcionando estimaciones iniciales de las fases de la nueva estructura a partir de una estructura previamente conocida.

El uso de RM se ha vuelto naturalmente más común a medida que se expande la base de datos de estructuras conocidas depositadas en el PDB (https://www.rcsb.org/ ). El RM se utiliza actualmente para resolver hasta el 70% de las estructuras macromoleculares depositadas y, en el mejor de los casos, tiene las ventajas de ser una técnica rápida, barata y altamente automatizada. Sin embargo, tiene la limitante de no ayudar en los casos de proteínas con plegamiento desconocido [52].

Los pasos computacionales en la determinación de la estructura cristalina por reemplazo molecular requieren dos conjuntos de información: (1) las coordenadas de los átomos en la molécula de fases conocidas y (2) un conjunto de datos de difracción de rayos X de los cristales de la proteína desconocida. El protocolo para obtener un conjunto de fases por reemplazo molecular para una proteína desconocida es el siguiente según Banaszak [53]:

- Medir los datos difracción de rayos X de los cristales de la proteína de interés.
- Comparar la función de Patterson[G] de la proteína desconocida con la de una proteína conocida para obtener una transformación rotacional, colocando la molécula conocida en la orientación correcta en la celda unitaria desconocida.
- Encontrar la posición de traslación correcta de la molécula conocida correctamente orientada en la celda unitaria desconocida.
- Calcular un conjunto de fases de prueba, utilizando las coordenadas orientadas y trasladadas de la molécula conocida en la celda unitaria desconocida. Las amplitudes del cristal desconocido se combinan con las fases antes mencionadas para producir un mapa de densidad de electrones de prueba.

El caso más simple de remplazo molecular es conocido como sustitución molecular. En este caso el grupo espacial y la celda unitaria son iguales a los del modelo base. Este método se utiliza comúnmente cuando un cristal ya formado se “remoja” con compuestos para ocupar sitios activos, sitios de unión a metales, etc. O bien, en los casos en que el grupo espacial y la celda se mantienen, para evidenciar cambios en mutantes puntuales.

[G] La función de Patterson, se define como la función de autocorrelación de la densidad electrónica.

## 3.1 HIPÓTESIS.

El uso de parámetros tradicionales para la determinación de la resolución máxima en estructuras cristalográficas desechará información útil que permitiría aumentar el detalle de la descripción molecular.

## 3.2 OBJETIVO GENERAL

Determinar las diferencias entre los mapas de densidad electrónica y modelos, obtenidos bajo diferentes condiciones y contrastarlo con el uso de métricas, con el fin de proponer las mejores métricas para evitar al máximo desperdiciar datos ya colectados y aumentar la calidad y detalle de la estructura cristalográfica resultante.

## 3.3 OBJETIVOS ESPECÍFICOS

1) Indexar e integrar datos experimentales del *dataset* de la proteína M271.
2) Hacer la reducción y escalamiento de datos, utilizando conjuntos consecutivos[H] de *frames* (240, 120, 60 y 30), estableciendo diferentes resoluciones de corte; empezando con la máxima resolución medida de 1.6 Å y se aumentará de 0.05 Å en 0.05 Å hasta 2 Å.
3) Resolver las fases por sustitución y por reemplazo molecular con distintos modelos con identidad decreciente con la proteína M271.
4) Realizar un afinamiento de cuerpo rígido[I] y afinamiento restringido[J], para cada condición descrita anteriormente de manera automática; es decir, sin realizar un afinamiento manual.
5) Analizar el comportamiento de la calidad mapa-modelo para cada condición.
6) Correlacionar la calidad del mapa-modelo con diversas métricas para proponer la mejor combinación.

[H] Ejemplo: 1,2,3…120
[I] Afinamiento de coordenadas utilizando bloques rígidos de átomos, normalmente un dominio o cadena completos (o posiblemente todo el contenido de la unidad asimétrica).
[J] Afinamiento de coordenadas utilizando átomos.

## 4. METODOLOGÍA

Para el cumplimiento de los objetivos del presente trabajo, se eligió un caso de estudio aleatorio, el cual consiste en un *dataset* proveniente de un experimento de difracción de rayos X que se diseñó para resolver fases por reemplazo molecular. Es decir, se tomó la información experimental de una estructura ya determinada, obtenida por la técnica de difracción por rayos X [1].

### 4.1 CASO DE ESTUDIO

Se trabajó con el *dataset* del cristal de la proteína M271, colectado con el objetivo de determinar la estructura de M271 por reemplazo molecular y cuya estructura fue reportada por Campuzano [1].

Dicho *dataset* está compuesto por 240 patrones de difracción (*frames),* tomados a 1° de rotación por imagen en la línea 19-I-D del APS (Advanced Photon Source), con un detector ADSC Q315r CCD. M271 es un inhibidor de proteasa de la familia *Kunitz-STI*, aislado de *Solanum tuberosum*[K]. Un resumen acerca de lo que Campuzano [1] reportó en su tesis con respecto a M271 se encuentra en el **Anexo C**. Así mismo, la información cristalográfica de M271 se puede consultar en la **Tabla 4.0**.

**Tabla 4.0.** Información cristalográfica de la proteína M271 [1].

| Resolución | 1.7 |
|---|---|
| Longitud de onda (Å) | 0.9791 |
| Grupo espacial | $P2_12_12$ |
| Celda unitaria: | |
| a, b, c (Å) | 62.39, 76.75, 82.37 |
| α, β, γ (°) | 90, 90, 90 |
| *Mosaicity* (°) | 0.14 |
| Resolución (Å) | 48.41 – 1.69 |
| $R_{merge}$ (%) | 10.5 (75.7) |
| CC ½ (%) | 99.9 (85) |
| *Completeness* (%) | 99.9 (99.8) |
| I/σ (I) | 21.58 (3.36) |
| Redundancia | 12.58 |
| R-work (%) | 17 |
| R-free (%) | 20 |

[K] Esta información corresponde a un cristal obtenido en la condición 6 del Kit Crystal Screen Cryo: Cloruro de magnesio hexahidratado 0.22 M, TRIS 0.08 M, PEG 4000 26% p/v y glicerol 20% v/v. La estructura final reportada por Campuzano (Campuzano, 2019) tiene una resolución de 1. 7 Å, utilizando como índices de corte de resolución máxima a CC ½ mayor a 80% y I/σ en la franja de mayor resolución mayor a dos.

## 4.2 MODELOS CRISTALOGRÁFICOS PARA RESOLVER EL PROBLEMA DE FASES

Se realizó un análisis tipo BLAST en la página de *National Center for Biotechnology Information* (https://blast.ncbi.nlm.nih.gov/),empleando la secuencia de aminoácidos de la cadena B, sin etiqueta de histidinas, del modelo cristalográfico final de M271[1], y una matriz BLOSUM45.

La secuencia de aminoácidos, en código de una letra, usada para el análisis fue la siguiente:

"SPLPKPVLDTNGKKLNPNSSYRIISTFWGALGGDVYLGKSPNSDAPCPDGVFRYNSD
VGPSGTPVRFIPLSGANIFEDQLLNIQFNIPTVKLCVSYTIWKVGNINAHLRTMLLETGGT
IGQADSSYFKIVKSSKFGYNLLYCPLTRHFLCPFCRDDNFCAKVGVVIQNGKRRLALVN
ENPLDVLFQEV".

Se buscaron modelos cristalográficos con identidad de secuencia de entre 20% y 100%, los cuales se muestran en la Tabla 4.1. Cabe mencionar que no se encontraron modelos cristalográficos con identidad de secuencia de entre 35% y el 76%. Para emplear estos modelos como base para la resolución de fases, únicamente se utilizó una cadena de aminoácidos, sin moléculas de aguas ni ligandos. En el caso de los modelos cristalográficos que tenían más de una cadena en la unidad asimétrica[LL], se empleó la secuencia de aminoácidos de una sola de ellas. La cadena utilizada se muestra en el último renglón de la Tabla 4.1. Es decir, que al archivo pdb original descargado del *Protein Data Bank* para cada modelo, se le dejó únicamente una cadena de la proteína.

### 4.2.1 Descripción de los modelos que se emplean como modelo base para el cálculo de fases.

Los códigos PDB y un resumen con algunas características de las coordenadas utilizadas para calcular fases, se muestran a continuación.

**1AVW (25%de identidad)** es un inhibidor de tripsina Kunitz STI de la planta de soya, tiene 181 aminoácidos, consta de dos enlaces disulfuro. La estructura reportada, corresponde a un complejo, entre la tripsina pancreática porcina y el inhibidor utilizado en ese proyecto. Tiene una resolución de 1.75 Å, con un $R_{work}$ de 18.9% [54].

**5DVH (35% de identidad)**. Corresponde a un inhibidor de proteasas de cisteína extraído de papa. Tiene actividad inhibitoria en endopeptidasas. Pertenece a la familia Kunitz-STI. La cadena A tiene 193 aminoácidos, con una masa teórica de 21.31 kDa [55].

[LL] La parte más pequeña de la celda unitaria a partir de la cual se puede reproducir toda la estructura cristalina mediante la acción de todas las operaciones de simetría

**3TC2 (76% de identidad)** Corresponde a un inhibidor de proteasas serínicas de la familia Kunitz STI, aislado de papa. Cuenta con 3 cadenas en la unidad asimétrica (cadena A,B,C), cuya longitud es de 187 aminoácidos. Tiene masa teórica de 20.32 kDa [56]

**5DZU (89% de identidad).** Corresponde a un inhibidor de catepsina de papa D de la familia Kunitz STI. Se clasifica con actividad inhibitoria tipo serínica. Tiene una masa de 20.61 kDa [57].

En la **Tabla 4.1** se muestran las características cristalográficas de los modelos utilizados en esta tesis. De cada modelo estructural se tomó únicamente la información de una sola cadena de proteína (ya que en algunos casos tenían más de una). También se muestra la resolución máxima de cada una de estas estructuras, así como su grupo espacial.

En la **Figura 4.1** se muestras los modelos listón, de las proteínas elegidas para realizar el remplazo molecular, y en la **Figura 4.2** se muestra el alineamiento de las secuencias de aminoácidos de las proteínas elegidas.

**Tabla 4.1. Tabla cristalográfica de los modelos base utilizados para resolución de fases por reemplazo molecular en esta tesis**. En la tabla se muestran diversos datos provenientes de las estructuras cristalográficas que se eligieron después de realizar un BLAST con la secuencia de la cadena B, sin histidinas y usando los depósitos del PDB, contra el modelo final de M271. Las distintos valores mostrados se extrajeron del *Protein Data Bank* (https://www.rcsb.org/) y de Campuzano, 2019.

| **% DE IDENTIDAD** | **25** | **35** | **76** | **89** | **100** |
|---|---|---|---|---|---|
| **Código PDB** | 1AVW | 5DVH | 3TC2 | 5DZU | M271*** |
| **Descripción** | Complejo tripsina pancreática porcina/ inhibidor de tripsina de soya, cristal ortorrómbico. | Estructura de PCP1-3, inhibidor de cisteína del tipo Kunitz. | Estructura de un cristal de un inhibidor de proteasa serina de papa. | Estructura de un inhibidor de catepsina D de papa. | Inhibidor de proteasa de la familia Kunitz-STI, aislado de *Solanum tuberosum*. |
| **Resolución (Å):** | 1.75 | 1.80 | 1.60 | 2.12 | 1.70 |
| **$R_{free}$:** | 0.214 | 0.223 | 0.221 | 0.270 | 0.200 |
| **$R_{work}$:** | 0.189 | 0.197 | 0.168 | 0.191 | 0.170 |
| **Grupo especial** | $P2_12_12_1$ | $P\ 4_1\ 2_1\ 2$ | $P\ 1\ 2_1\ 1$ | C 1 2 1 | $P2_12_12$ |
| **a (Å)** | 58.91 | 76.64 | 54.82 | 75.46 | 62.39 |
| **b (Å)** | 62.33 | 76.64 | 93.92 | 124.77 | 76.75 |
| **c (Å)** | 151.46 | 107.31 | 55.44 | 37.88 | 82.37 |

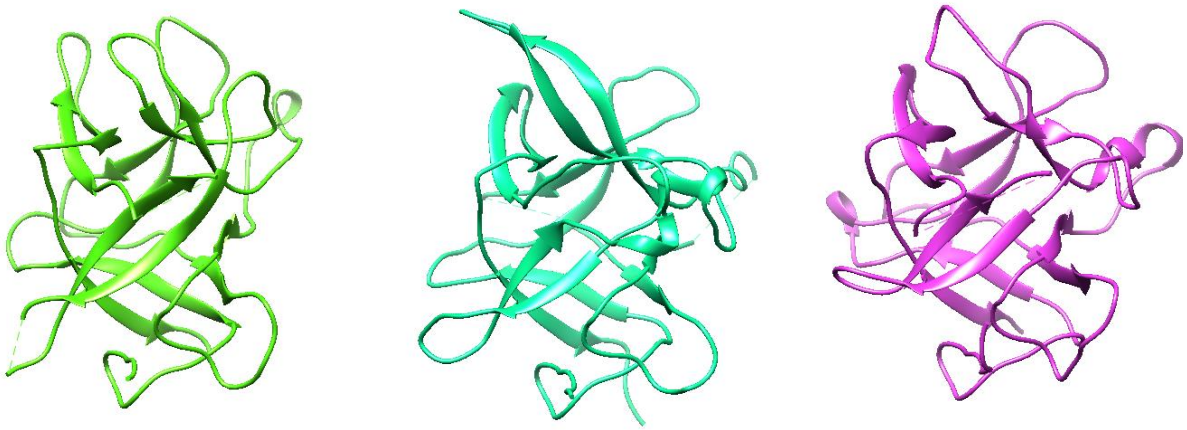
| | | | | | |
|---|---|---|---|---|---|
| **α (°)** | 90.00 | 90.00 | 90.00 | 90.00 | 90.00 |
| **β (°)** | 90.00 | 90.00 | 100.69 | 95.14 | 90.00 |
| **γ (°)** | 90.00 | 90.00 | 90.00 | 90.00 | 90.00 |
| **% Integridad (Limites de resolución en Å)** | 92.0 (8Å - 1.75Å) | 99.7 (48.37Å - 1.80 Å) | 69.9 (54.23-1.60Å) | 96.4 (64.38Å-2.12Å) | 99.9 (40.0-1.70 Å) |
| **$R_{merge}$** | 0.08 | 0.11 | 0.04 | 0.07 | 0.75 |
| **I/σ(I) *** | 3.43 (en la faja de mayor resolución) | 1.15 (en la faja de mayor resolución) | 1.59 (en la faja de mayor resolución) | 1.05 (en la faja de mayor resolución) | 3.36 (en la faja de mayor resolución) |
| **Cadenas reportadas en el PDB** | A \ 223 aa Tripsina<br>B \ 171 aa Inhibidor de tripsina | A \ 185 aa \ Inhibidor de proteasas | A \ 181 aa<br>B \ 177 aa<br>C \ 178 aa<br>Inhibidor de proteasas | A\187 aa<br>B\176 aa<br>Inhibidor de proteasas | A \ 190 aa Inhibidor de proteasas<br>B \ 194 aa |
| **Cadena utilizada en esta tesis (en paréntesis se declara el número de aa)**** | B \ 171 aa | A \ 185 aa | A \ 181 aa | A \ 187 aa | B \ 188 aa |

*Intensidades calculadas a partir de las amplitudes depositadas por los autores. Extraídos del *PDB X-ray Structure Validation Report* de cada depósito (https://www.rcsb.org/).

** Cadena tomada como modelo para resolver las fases en esta tesis.

***Esta estructura aún no se ha depositado en el PDB y fue obtenida de Campuzano [1].

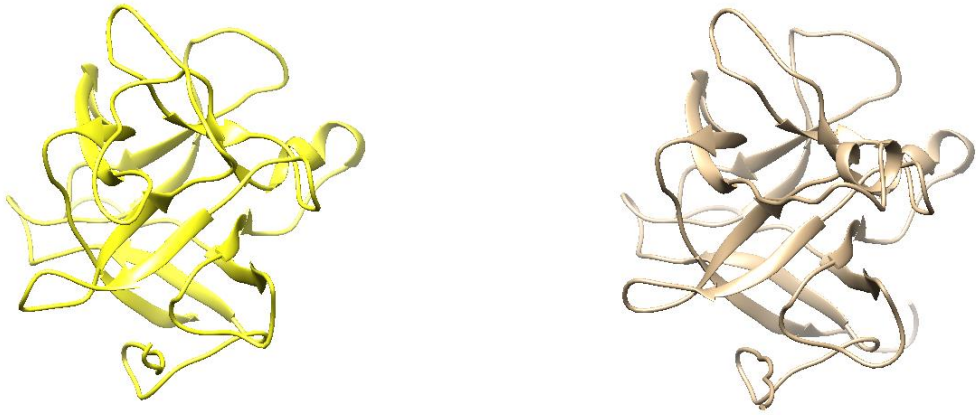
En la **Figura 4.1** se muestran las coordenadas que se utilizaron como modelo base para la resolución de fases utilizando una representación tipo listón.



1AVW (25% identidad)

5DVH (35% identidad)

3TC2 (76% identidad)



5DZU (89% identidad)

M271 (100% identidad)

**Figura 4.1. Modelo listón de las estructuras de las proteínas seleccionadas para resolver el problema de fases en esta tesis.** Dichas estructuras tienen entre 25 y 100%de identidad de secuencia de aminoácidos con relación a la secuencia de M271. Imágenes generadas en Chimera 1.15

En la **Tabla 4.0** se muestran los valores de RMSD de cada una de las coordenadas empleadas en los reemplazos moleculares, con respecto a las coordenadas tridimensionales de M272, donde se observa, tal como se esperaba, un incremento de este valor conforme disminuye el valor de identidad de secuencia. Cabe mencionar que entre los modelos 5DVH y 1AVW, pareciera que, a menor identidad, hay una mejora en el valor de RMSD, sin embargo, el número de carbonos alfa usados en el cálculo de RMSD, pasó de 157 a 147, algo similar ocurre entre 5DZU y 3TC2. Estas observaciones resaltan el hecho de que realmente lo que hace que un modelo para reemplazo molecular funcione mejor que otro, son la similitud de coordenadas y la cobertura de las mismas. Entre más similares en secuencia y más similares en el número de aminoácidos, mejor será este modelo para obtener fases por reemplazo molecular, por lo que el uso de la identidad de secuencia solo tiene sentido bajo el principio de que modelos con alta identidad en secuencia, deberían significar valores bajos de RMSD entre si.

**Tabla 4.0 RMSD de los modelos base para resolver el problema de fases en esta tesis.** Valores de RMSD y valores de carbonos alfa superpuestos.

| | Referencia | | | | |
|---|---|---|---|---|---|
| Proteína: | M271 | 5DZU | 3TC2 | 5DVH | 1AVW |
| ID: | 100% | 89% | 76% | 35% | 25% |
| RMSD | 0 | 0.7904 | 0.7539 | 1.6818 | 1.5425 |
| Carbonos alfa | 194 | 176 | 173 | 157 | 147 |


**Figura 4.2. Alineamiento de las secuencias de aminoácidos de las distintas estructuras de las proteínas seleccionadas utilizadas para la resolución de fases en esta tesis.** Se realizó un alineamiento múltiple de secuencia [58], utilizando la plataforma de T-Coffe (http://tcoffee.crg.cat/apps/tcoffee/do:expresso), e introduciendo las secuencias en formato fasta.

Residuos en colores morado y verde representan una mala calidad de alineamiento y es poco confiable. En amarillo se colorean residuos que tienen una alineación más o menos confiable. Yen color rosa son residuos que se alinearon correctamente y de manera confiable.

## 4.3 PROCESAMIENTO DE DATOS DE DIFRACCIÓN

La información utilizada para este proyecto se obtuvo de las diferentes etapas y condiciones del proceso para la determinación de estructuras cristalográficas por rayos X. A continuación, se describen brevemente los pasos que se siguieron. El proceso de afinamiento se realizó de forma automática.

Las imágenes de difracción del caso de estudio fueron indexadas e integradas con el programa MOSFLM [59] con el que se auto indexó e integró todo el conjunto de imágenes (*frames*) del *Dataset* del caso de estudio, utilizando como coordenadas para el centro de las imágenes determinadas en la colecta de datos de M271 (x=100.57mm e y=107.22 mm).y obteniéndose un archivo de extensión MTZ con las reflexiones indexadas.

El escalamiento y reducción de datos se realizó con SCALA dentro de la suite CCP4 [60]. En este paso se utilizaron 240, 120, 60 y 30 *frames*, estableciendo diferentes resoluciones de corte. Se empezó con la máxima resolución medida de 1.6 Å con incrementos de 0.05 Å hasta alcanzar 2 Å, seleccionando el 5% de las reflexiones para evaluar el $R_{free.}$ Para ello se ingresó el archivo MTZ proveniente de MOSFLM, dando 36 combinaciones únicas de resolución máxima y de número de *datasets.* Por lo tanto, se generaron 36 archivos MTZ con las reflexiones, amplitudes y fases correspondientes. De esta sección se extrajeron los siguientes indicadores: $R_{merge}$, $R_{meas}$, $R_{pim}$, redundancia, integridad, $I/\sigma\ (I)_{mean}$, $I/\sigma\ (I)$ los cuales pueden revisarse en el **ANEXO B.**

En la **Tabla 4.2.** se muestran los valores de redundancia e integridad para cada una de las condiciones o escenarios utilizados en esta tesis.

**Tabla 4.2. Valores de integridad y redundancia para cada escenario**. Se muestran los valores de integridad y redundancia correspondientes a las diferentes resoluciones de corte utilizando diferente número de *frames*.

| *Frames* | 240 | | 120 | | 60 | | 30 | |
|---|---|---|---|---|---|---|---|---|
| Resolución | Integridad (%) | Redundancia | Integridad (%) | Redundancia | Integridad (%) | Redundancia | Integridad (%) | Redundancia |
| 2.00 Å | 100 | 9.8 | 100 | 4.9 | 65.4 | 3.7 | 59.7 | 2 |
| 1.95 Å | 100 | 9.8 | 100 | 4.9 | 66.2 | 3.7 | 60.6 | 2 |
| 1.90 Å | 100 | 9.8 | 100 | 4.9 | 66.2 | 3.7 | 60.8 | 2 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1.85 Å | 100 | 9.8 | 100 | 4.9 | 66.9 | 3.7 | 61.4 | 2 |
| 1.80 Å | 100 | 9.8 | 100 | 4.9 | 67.4 | 3.6 | 61.9 | 2 |
| 1.75 Å | 100 | 9.7 | 100 | 4.8 | 67.9 | 3.6 | 62.1 | 1.9 |
| 1.70 Å | 100 | 9 | 100 | 4.5 | 68.5 | 3.3 | 61 | 1.8 |
| 1.65 Å | 100 | 8 | 100 | 4 | 68.9 | 2.9 | 58.4 | 1.7 |
| 1.60 Å | 97.5 | 6.6 | 97.5 | 3.3 | 67.1 | 2.4 | 53.8 | 1.5 |

Las fases se estimaron por remplazo molecular utilizando las estructuras de las proteínas reportadas en la Tabla 4.1.y el programa PHASER [18]. Para ello se utilizaron los archivos MTZ provenientes de SCALA y los archivos PDB con las coordenadas de las diferentes estructuras de las proteínas obtenidos directamente del PDB, excepto M271, que fue generado por nuestro grupo y aún no se deposita en el PDB (de esta sección se extrajeron los valores de LLG y FTZ, ver las Figuras 5.3 y 5.4).

Posteriormente se realizó un afinamiento de cuerpo rígido en REFMAC [61] utilizando los archivos MTZ de Phaser (F y sigmaF) y los archivos PDB con las coordenadas de las estructuras de las proteínas seleccionadas para resolver el problema de fases, generados por PHASER (de esta sección se extrajeron los valores de $R_{work}$ y $R_{free}$ ,ver las Figura 5.5 y 5.6)

Posteriormente se hizo un afinamiento restringido con PHENIX [62] de tres macrociclos (cada macrociclo consiste en un ajuste de coordenadas, afinamiento de parámetros de desplazamiento atómico y cálculo de ocupaciones), utilizando el PDB generado en el afinamiento de cuerpo rígido y los MTZ generados en SCALA. Se obtuvieron un archivo de extensión MTZ con las reflexiones, amplitudes, fases y un archivo PDB, con las coordenadas y parámetros de desplazamiento atómico (se extrajeron los valores de $R_{work}$ y $R_{free}$, ver las Figuras 5.8 y 5.9)

Finalmente se utilizó el programa COOT [63] y la herramienta de PHENIX, *Real Space Correlation* [64], para visualizar y analizar la calidad de los mapas de densidad electrónica

(2FO-FC). Es importante mencionar que posteriormente, ningún modelo se sometió a afinamiento manual (se extrajeron los siguientes indicadores: CC, Figura 5.11)

Los límites de los indicadores de calidad utilizados en el presente trabajo, así como su origen, se muestran a continuación:

**Tabla. 4.3.Valores de los indicadores de calidad utilizados en este proyecto.**

$R_{merge} \leq 0.8$ [35].

$R_{meas} \leq 0.8$ [35].

$R_{pim} \leq 0.5$ (Valores típicos utilizados en nuestro laboratorio).

CC ½ ≥ 0.15 [14].

Integridad ≥ 70 [37].

Redundancia ≥ 2 (Valores típicos utilizados en nuestro laboratorio).

$I/\sigma(I) \geq 2$ [14].

$I/\sigma(I)$ mean ≥ 2 [35].

LLG ≥ 36 [43].

FTZ ≥ 6 [41]

$R_{work} \leq 0.5$ [45].

$R_{free} \leq 0.5$ [45].

$\Delta R \leq 0.05$ (Valores típicos utilizados en nuestro laboratorio).

$CC_{global} \geq 0.8$

$CC_{local} \geq 0.85$

En la **Figura 4.3** se esquematiza la metodología que se utilizó en el presente trabajo.

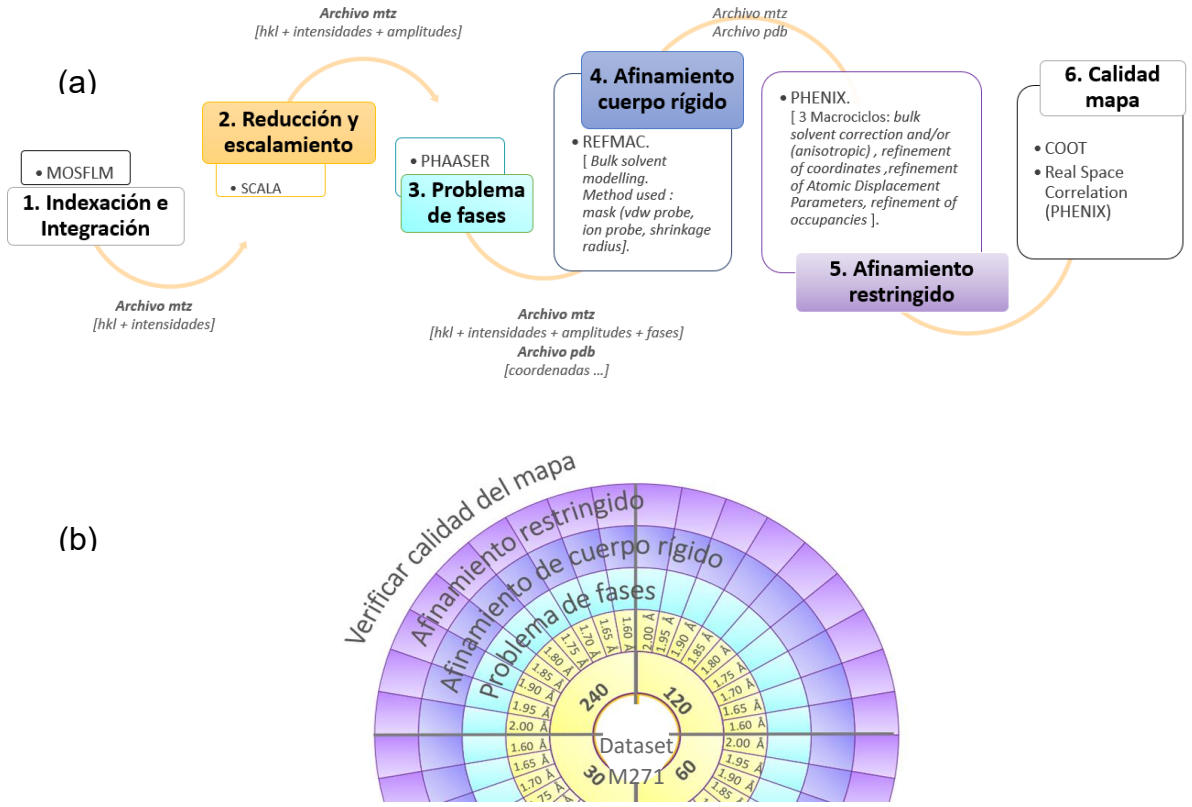



**Figura 4.1 Metodología.** (a) Procesamiento de datos: 1. Indexación e integración de los *spots* de las imágenes obtenidas por rayos X, utilizando el programa MOSFLM, 2. Reducción y escalamiento, utilizando el programa SCALA en CCP4i, 3. Resolución del del problema de fases y utilizando el programa PHASER, 4. Afinamiento de cuerpo rígido y utilizando el programa REFMAC5 en CCP4i, 5. Afinamiento restringido, utilizando el programa PHENIX. 6.Evaluación de la calidad del mapa, utilizando el programa COOT y la herramienta *Real Space Correlation* dentro de PHENIX.

(b) Se muestran los 36 escenarios y el proceso a que se sometió cada estructura de las proteínas seleccionadas para resolver el problema de fases. En la etapa de escalamiento y reducción de datos se utilizaron 30, 60, 120 y 240 imágenes (*frames*) y 9 diferentes resoluciones de corte.

# 5. RESULTADOS

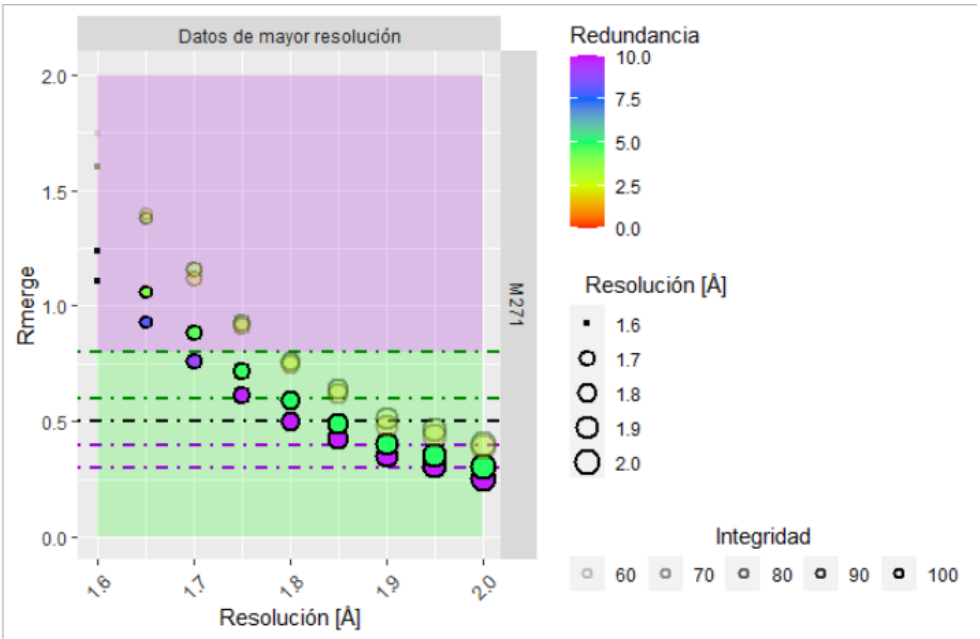
## 5.1 ANALISIS DE INDICADORES DE CALIDAD: ESCALAMIENTO Y REDUCCIÓN DE DATOS.

Después de indexar e integrar al *dataset* de M271, se realizó la reducción y escalamiento de datos en SCALA [60]. Se extrajeron los siguientes indicadores de calidad para cada una de estas condiciones en la última franja de resolución (para los que aplica): $R_{merge}$, $R_{meas}$, $R_{pim}$, CC½, Integridad, redundancia, I/σ(I), I/σ(I) mean**.** Esta información se graficó para determinar la resolución de corte para cada grupo utilizado en esta tesis (240, 120, 60 y 30 *frames*). Un ejemplo de dicha gráfica se muestra en la **Figura 5.1** y el resto de graficas se presentan en el **Anexo B**. De los límites mencionados para los diversos indicadores de resolución máxima de corte, indicados en el **Capítulo 1.4.,** se seleccionaron aquellos más permisibles; es decir, los que abarcan la mayoría de los límites encontrados en la literatura. Estos límites se muestran en la **Tabla 5.1.** y con ellos se eligieron las resoluciones máximas para cada grupo.

**Tabla 5.1 Valores de corte. Valores de los indicadores de corte utilizados en este proyecto.** En la tabla se muestra el valor mínimo o máximo, indicando la fuente de dichos límites.

| |
|---|
| $R_{merge} \leq 0.8$ [35]. |
| $R_{meas} \leq 0.8$ [35]. |
| $R_{pim} \leq 0.5$ (Valores típicos utilizados en nuestro laboratorio). |
| CC ½ ≥ 0.15 [14]. |
| Integridad ≥ 70 [37]. |
| Redundancia ≥ 2 (Valores típicos utilizados en nuestro laboratorio). |
| I/σ(I) ≥ 2 [14]. |
| I/σ(I) mean ≥ 2 [35]. |

En la **Figura 5.1** se muestra $R_{merge}$ versus resolución, donde el valor de corte de $R_{merge}$ es 0.8 [35], el cual queda señalado como la interfaz entre el área verde y el área morada (siendo el área verde la que cumple los requisitos de Karplus & Diederichs). Los puntos cercanos a esta línea, son considerados como resoluciones de corte para cada uno de los cuatro grupos seleccionados en esta tesis. La resolución de corte que aplicaría para cada grupo de acuerdo con la **Tabla 5.1**, sería la siguiente: 240 Frames:1.7Å, 120 Frames:1.75 Å, 60 Frames: 1.80 Å, 30 Frames: 1.80 Å.




---- $R_{merge}$:0.3-0.5. Valores típicos utilizados en nuestro laboratorio.

---- $R_{merge}$: 0.6 – 0.8. [35].

---- $R_{merge}$:0.3-0.4. [36].

**Figura 5.1. $R_{merge}$ vs resolución.** El valor de redundancia se representa por la variación de color, la resolución por el tamaño del punto y la integridad por la nitidez del punto. Se observa que el valor de $R_{merge}$ va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más elevados de $R_{merge}$, que los grupos de alta redundancia e integridad, haciéndose la brecha menor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura y mencionados en la parte baja de la figura. El área verde representa a los escenarios que cumplen con un valor de $R_{merge}$ de al menos 0.8. El área morada representa los escenarios que no cumplen con el límite anterior. Este gráfico fue realizado en RStudio [65].

De las gráficas del **Anexo B**, se extrajo el valor de resolución de corte para cada indicador, como se mostró en el ejemplo de la **Figura 5.1.** y utilizando como límites los indicados en la **Tabla 5.1.** En la **Tabla 5.2** se muestran estos valores, observándose que utilizando el mismo grupo de datos (240, 120, 60 o 30 *frames*), se presentan diferentes resoluciones de corte de resolución para cada uno de los indicadores. En particular, en el grupo de 60 y 30 *frames*, no se cumple con el valor mínimo requerido de integridad ≥ 70%.

**Tabla 5.2. Resoluciones de corte utilizando el valor límite inferior para cada indicador de los analizados en esta tesis**. Se muestran cuatro valores de resolución máxima, por parámetro de corte, es decir, un valor de resolución por cada grupo (240, 120, 60, 30 *frames*). Se encontró una amplia gama de resoluciones de corte de acuerdo con el parámetro que se decida utilizar.

| Número de *frames* | | 240 | 120 | 60 | 30 |
|---|---|---|---|---|---|
| Redundancia | | 6.6 -9.8 | 3.3 – 4.9 | 2.4 - 3.7 | 1.5 - 2 |
| Integridad | | 97.5 - 100 | 97.5 - 100 | 65.4 – 68.9 | 53.8 – 62.1 |
| Indicador | Rango de indicador | Resolución de corte por grupo e indicador de corte | | | |
| $R_{merge}$ | 0 - 0.8 | 1.70 Å | 1.75 Å | 1.80 Å | 1.8 Å |
| $R_{meas}$ | 0 – 0.8 | 1.70 Å | 1.75 Å | 1.85 Å | 1.85 Å |
| $R_{pim}$ | 0 – 0.5 | 1.60 Å | 1.70 Å | 1.75 Å | 1.85 Å |
| CC ½ | 0.15 – 1.00 | 1.60 Å | 1.60 Å | 1.60 Å | 1.60 Å |
| Integridad | 70 – 100 | 1.60 Å | 1.60 Å | No cumple | No cumple |
| Redundancia | 2 -10 | 1.60 Å | 1.60 Å | 1.60 Å | 1.70 Å |
| I/σ(I) | > 2 | 1.85 Å | 1.90 Å | 1.95 Å | 1.95 Å |
| I/σ(I) mean | >2 | 1.80 Å | 1.70 Å | 1.85 Å | 1.95 Å |

Lo descrito anteriormente, se muestra en un gráfico de frecuencias en la **Figura 5.2**, donde se observa que existe una variabilidad de estas resoluciones de corte, en el rintervalo de 1.6 Å a 2.0 Å. Se obtuvo un pico máximo en la resolución de 1.6 Å. La mayoría de estos datos corresponden a grupos de alta redundancia e integridad (los grupos de 240 y 120 *frames*),los que comparados con los grupos de menor redundancia e integridad (los grupos de 60 y 30 *frames*) presentan una mayor distribución de valores.

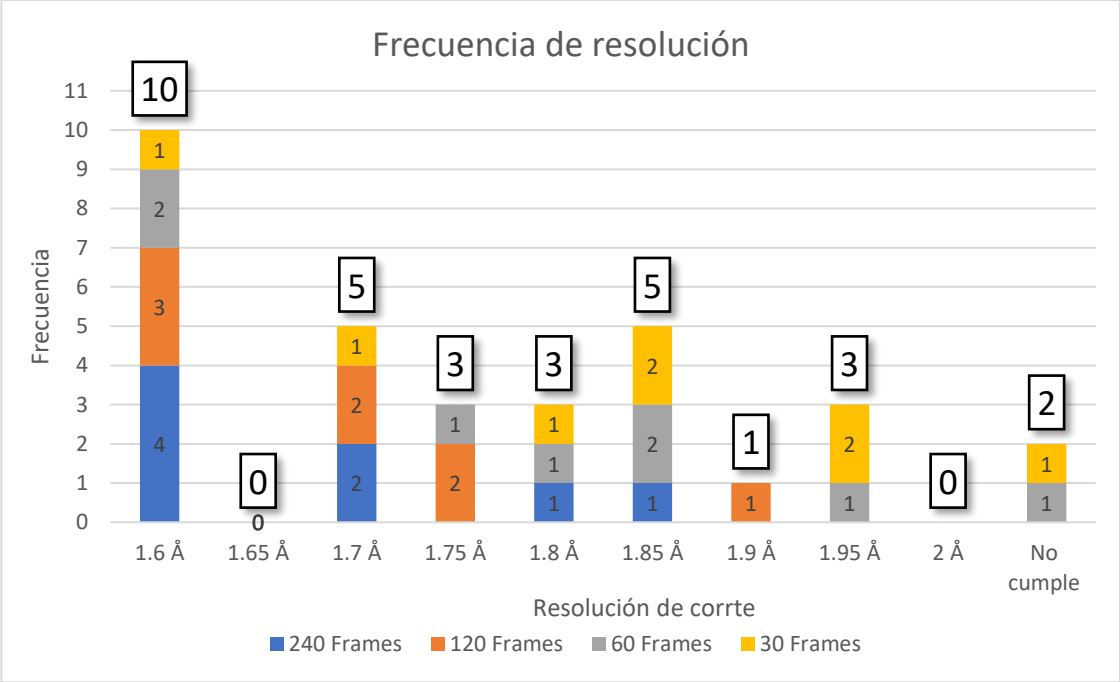



**Figura 5.2. Frecuencia de resolución de corte.**
Muestra las frecuencias de las resoluciones de corte para los grupos: 240, 120, 60, 30 *frames,* obtenidas utilizando todos los indicadores: $R_{merge}$, $R_{meas}$, $R_{pim}$, CC½, integridad, redundancia, I/ σ (I), I/ σ (I) mean. En el eje horizontal se muestran los diferentes valores de resoluciones de corte, donde la última casilla corresponde a los casos que no cumplen con el valor mínimo del indicador. El eje vertical, corresponde a la frecuencia de casos. Este gráfico se construyó con los datos de la **Tabla 5.2.**

## 5.2 RESOLUCIÓN DE FASES Y AFINAMIENTO

Después de la etapa de escalamiento y reducción de datos se prosiguió con la resolución del problema de fases utilizando los modelos mostrados en la **Tabla 1.4** mediante la técnica de reemplazo molecular.

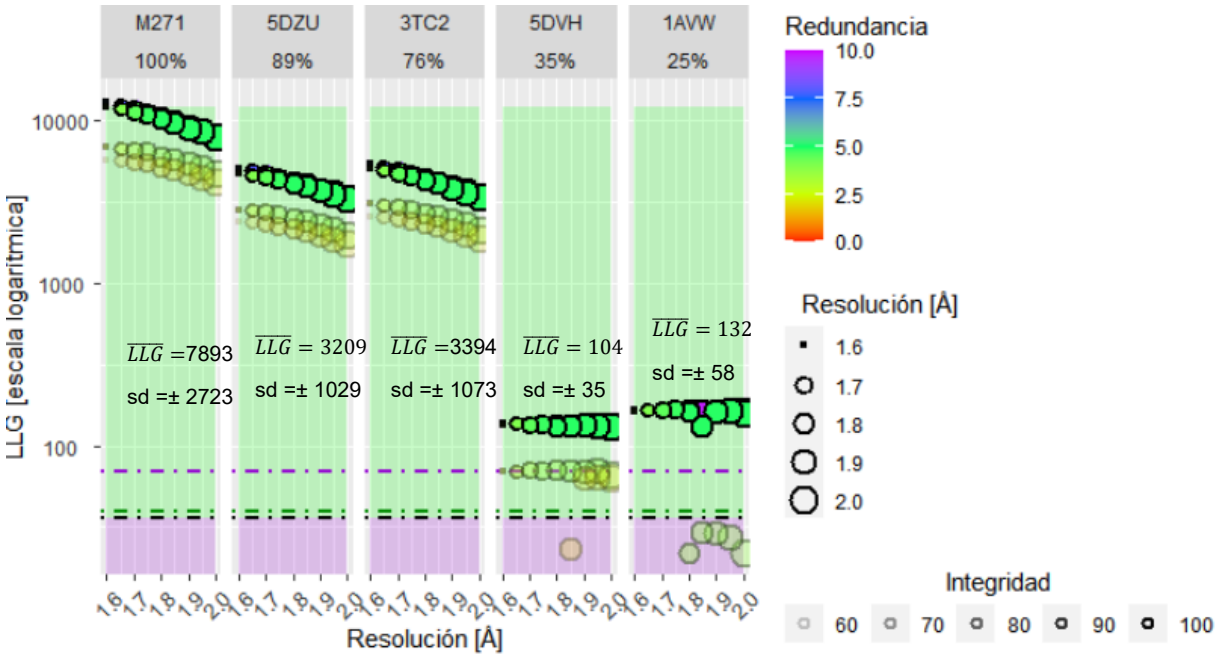
De esta etapa se obtuvieron los valores del parámetro LLG y FTZ para cada uno de los escenarios, mediante el programa PHASER [18]

En la **Figura 5.3** se observa el valor de LLG en escala logarítmica contra la resolución máxima de corte, es mayor conforme el porcentaje de identidad es mayor. El valor de LLG también fue disminuyendo conforme la resolución fue bajando, pero para las estructuras de las proteínas seleccionadas para resolver el problema de fases con un porcentaje de identidad de secuencia de 35% y 25% , esta tendencia ya no es tan evidente. Lo anterior debido a que los valores se comportan de manera muy similar dentro de un mismo grupo, sin importar la resolución de corte.

A menores valores de redundancia e integridad, también disminuyeron los valores de LLG, a tal punto, que para el conjunto del modelo con identidad de secuencia de 35 % (1.85Å utilizando 30 *frames*) y 25% (1.80Å utilizando 60 *frames*), Phaser no es capaz de encontrar una solución para el reemplazo molecular (estos escenarios se describen con mayor detalle en la **Tabla 5.3**).

.

La redundancia y la integridad juegan un papel importante, cuando se tienen estructuras de proteínas para resolver el problema de fases con un porcentaje de identidad de aminoácidos cada vez menor. Utilizando la información de la **Tabla 5.2** los indicadores que predijeron con mayor exactitud las resoluciones de corte para el modelo de 35% de identidad a 1.85 Å, fueron $R_{meas}$ y $R_{pim}$, mientras que para el modelo de 25% de identidad a una resolución de 1.80 Å, $R_{merge}$ fue el mejor parámetro de corte.




**----** LLG ≥ 36 [43]

**----** LLG >40 [41].

**----** LLG >70. [44].

**Figura 5.3. LLG en escala logarítmica contra resolución.** El valor de redundancia se representa por la variación de color, la resolución por el tamaño del punto y la integridad por la nitidez de este. Se observa que el valor de LLG va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de LLG, en comparación con los grupos de alta redundancia e

integridad. El área verde representa a los escenarios que cumplen con un valor de LLG mayor a 36 y el área morada representa los escenarios que no cumplen con esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada estructura de proteína seleccionada para resolver el problema de fases. Las líneas punteadas simbolizan los límites encontrados en la literatura y marcados en la parte baja de la figura.

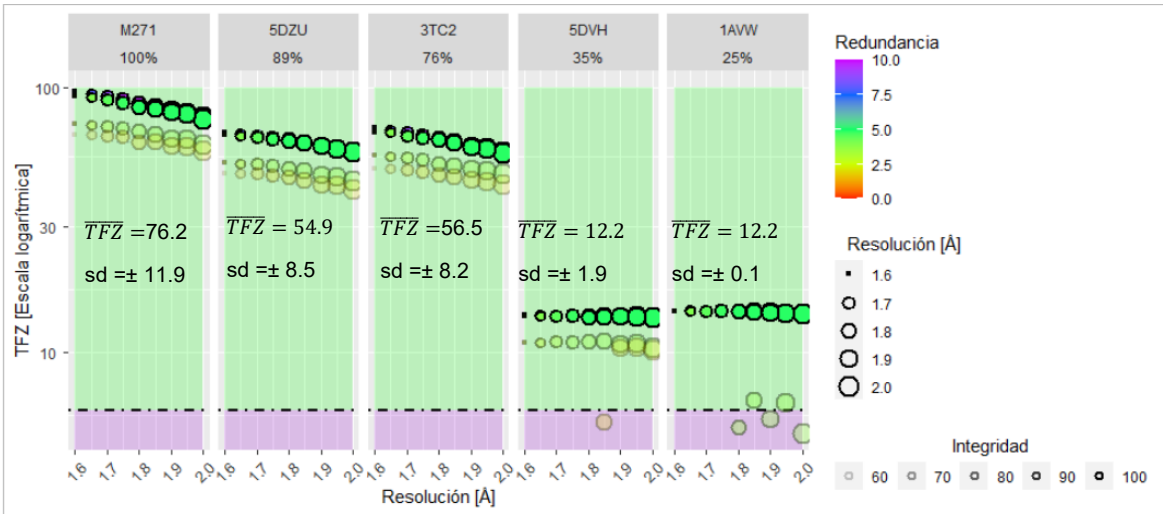
**Tabla 5.3. Escenarios por grupos a partir de los cuales ya no fue posible resolver fases.** En esta tabla se describen las características de los escenarios donde no se obtuvieron fases, ordenados por grupos de *frames:* 30, 60, 120 y 240. Se incluyen, además, los valores de los indicadores en cada una de las etapas, y estos se contrastan con datos provenientes de la Tabla 5.2 (resoluciones de corte utilizando el valor límite inferior para cada indicador).

| | Modelo (código PDB) | | 5DVH | | 1AVW | |
|---|---|---|---|---|---|---|
| | %id de aa | | 35% | | 25% | |
| | *# frames* | | 30 | | 60 | |
| | Resolución máxima | | 1.85Å | | 1.80Å | |
| | Grupo | | 1 | | 2 | |
| | Indicador | Intervalo de indicador | Valor del indicador | Resolución de corte por grupo * | Valor del indicador | Resolución de corte por grupo* |
| ESCALAMIENTO | $R_{merge}$ | 0 - 0.8 | 0.621 | 1.8 Å | 0.763 | 1.8 Å |
| | $R_{meas}$ | 0 – 0.8 | 0.792 | 1.85 Å | 0.891 | 1.85 Å |
| | $R_{pim}$ | 0 – 0.5 | 0.484 | 1.85 Å | 0.448 | 1.85 Å |
| | CC ½ | 0.15–1.00 | 0.554 | 1.60 Å | 0.645 | 1.60 Å |
| | Integridad | 70% – 100% | 61.4% | No cumple | 67.4% | No cumple |
| | Redundancia | 2 -10 | 2 | 1.70 Å | 3.6 | 1.70 Å |

| | | | | | | |
|---|---|---|---|---|---|---|
| | I/σ(I) | > 2 | 1.45 | 1.95 Å | 1.25 | 1.95 Å |
| | I/σ(I) mean | >2 | 1.5 | 1.95 Å | 1.7 | 1.95 Å |
| FASES | LLG | > 36 | 23 | - | 22 | - |
| | FTZ | > 6 | 5.4 | - | 5.2 | - |
| AFINAMIENTO DE CUERPO RÍGIDO | $R_{free}$ | >0.5 | 0.5031 | - | 0.5426 | - |
| | $R_{work}$ | >0.5 | 0.5293 | - | 0.5491 | - |
| | ΔR | <0.05 | 0.0259 | - | 0.0065 | - |
| AFINAMIENTO RESTRINGIDO | $R_{free}$ | >0.5 | 0.4939 | - | 0.543 | - |
| | $R_{work}$ | >0.5 | 0.4153 | - | 0.4916 | - |
| | ΔR | <0.05 | 0.0786 | - | 0.0514 | - |
| CALIDAD MAPA | $CC_{global}$ | >0.7 | 0.672 | - | 0.581 | - |
| | %aa con un $CC_{local}$>0.85 | | 42.1% | - | 9.9% | - |

* Infomación extraída de la **Tabla 5.2.** Resoluciones de corte utilizando el valor límite inferior para cada indicador (tabla 5.1).

También de la etapa de resolución de fases se obtuvo el valor de FTZ (sección 1.4), que se muestra en la **Figura 5.4.** donde se observa el valor de FTZ en escala logarítmica contra la resolución máxima de corte. Se aprecia que de manera general el valor de FTZ es mayor conforme el porcentaje de identidad es mayor, y va disminuyendo conforme la identidad lo hace. El valor de FTZ también va disminuyendo conforme la resolución va bajando. Finalmente se observa que para la estructura de proteína con 35% de identidad a una resolución de 1.85 Å, con 30 *frames* ya no se cumple con el valor mínimo de 6 (Tabla 4.3). Esto también sucede para todos los escenarios con 30 *frames* de la estructura de proteína con 25% de identidad.

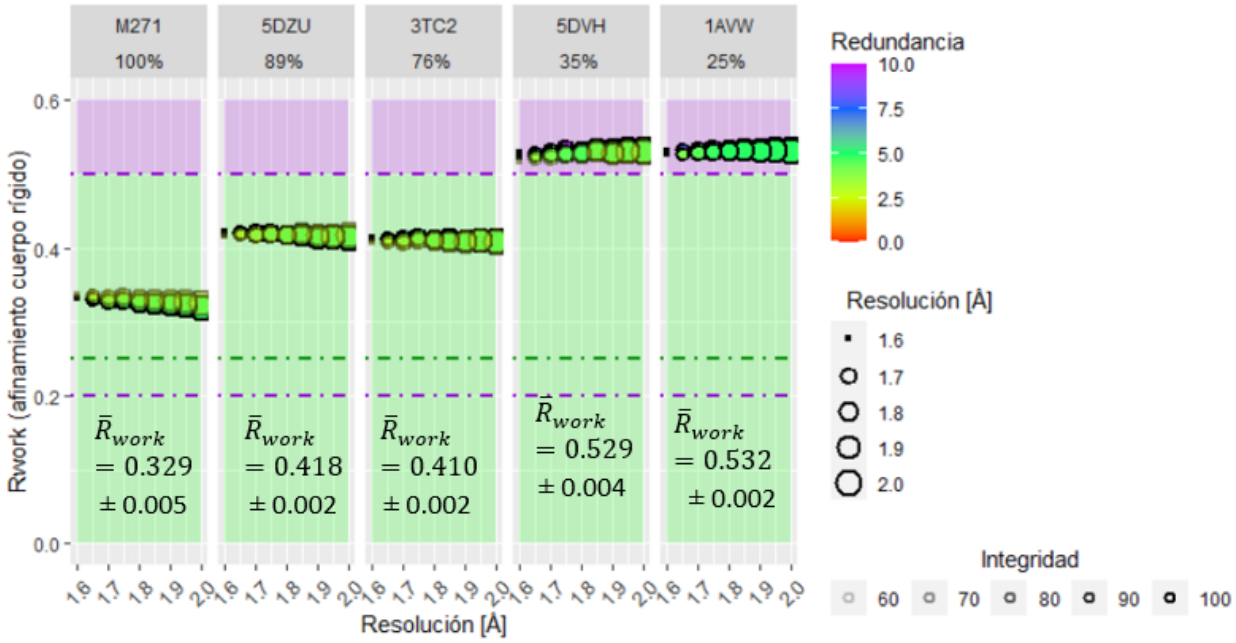



**----** FTZ >6 [41].

**Figura 5.4. FTZ en escala logarítmica contra resolución.** El valor de redundancia se representa por la variación de color, la resolución por el tamaño del punto y la integridad por la nitidez de éste. Se observa que el valor de FTZ va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de FTZ en comparación de los grupos de alta redundancia e integridad. Las líneas punteadas simbolizan los límites (Tabla 4.3) encontrados en la literatura. El área verde representa a los escenarios que cumplen con un valor de FTZ mayor a 6 y el área morada representa los escenarios que no cumplen con esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada estructura de proteína seleccionada para resolver el problema de fases.

Después de realizar la resolución de fases por sustitución molecular se prosiguió con un afinamiento de cuerpo rígido en el programa REFMAC [61]. En las **Figuras 5.5, 5.6 y 5.7** se representan los valores de $R_{work}$, $R_{free}$ y ΔR. Cabe recordar que ningún escenario fue sometido a modificaciones de coordenadas producto de afinamiento restringido, ni debido a ajustes utilizando un programa gráfico como COOT.

En la **Figura 5.5** se observa que $R_{work}$ tuvo valores más bajos conforme el porcentaje de identidad de aminoácidos va siendo mayor, lo que es lógico ya que esto permite explicar mejor la información experimental. Se tomó como límite un valor de $R_{work}$ menor a 0.5, que es el área verde en la **Figura 5.5.** Cuando los valores de $R_{work}$ fueron mayores a 0.5 se muestra en un área de color morado. Llama la atención que los escenarios que parten de estructuras de proteína con 35% y 25% de identidad de secuencia dieron lugar a los valores peores de $R_{work}$.

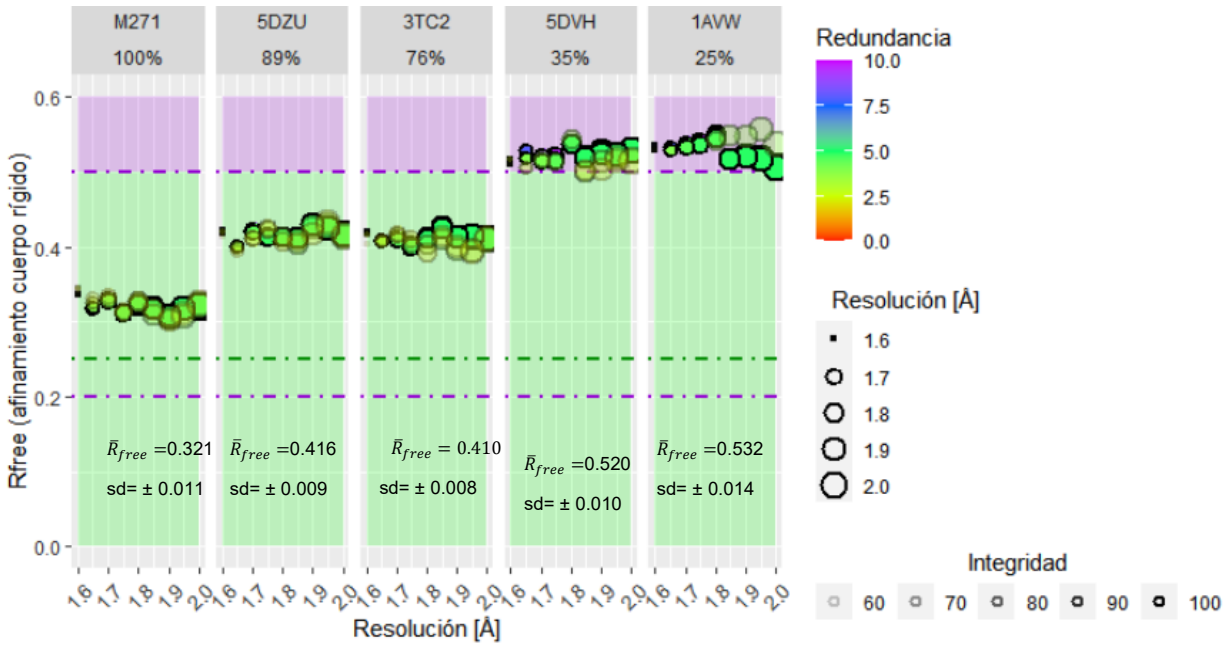



---- $R_{work}$: 0.25 Valores típicos utilizados en nuestro laboratorio.

---- $R_{work}$:0.3-0.5. [45].

**Figura 5.5. Rwork de afinamiento de cuerpo rígido contra resolución.** Se observan los valores de $R_{work}$ obtenidos de un afinamiento de cuerpo rígido, representados como círculos y su variación respecto a la resolución máxima de corte. Los valores de $R_{work}$ son menores conforme el porcentaje de identidad de aminoácidos es mayor. El área verde representa a los escenarios que cumplen con un valor de $R_{work}$ menor a 0.5 y el área morada representa los escenarios que no cumplen con esta restricción (modelos con el 35% y 25% de identidad). También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada una de las estructuras de las proteínas seleccionadas para resolver el problema de fases.

En la **Figura 5.6,** se muestran los valores de $R_{free}$, los que fueron aumentando conforme el porcentaje de identidad va disminuyendo, comportándose de manera similar a los valores de $R_{work}$. Sin embargo, los valores de $R_{work}$ se comportaron de una manera más uniforme que los valores de $R_{free}$. Los escenarios de los modelos de bajo porcentaje de identidad, 35% y 25%, los valores de $R_{free}$ fueron los más altos, obtenidos en esta tesis.

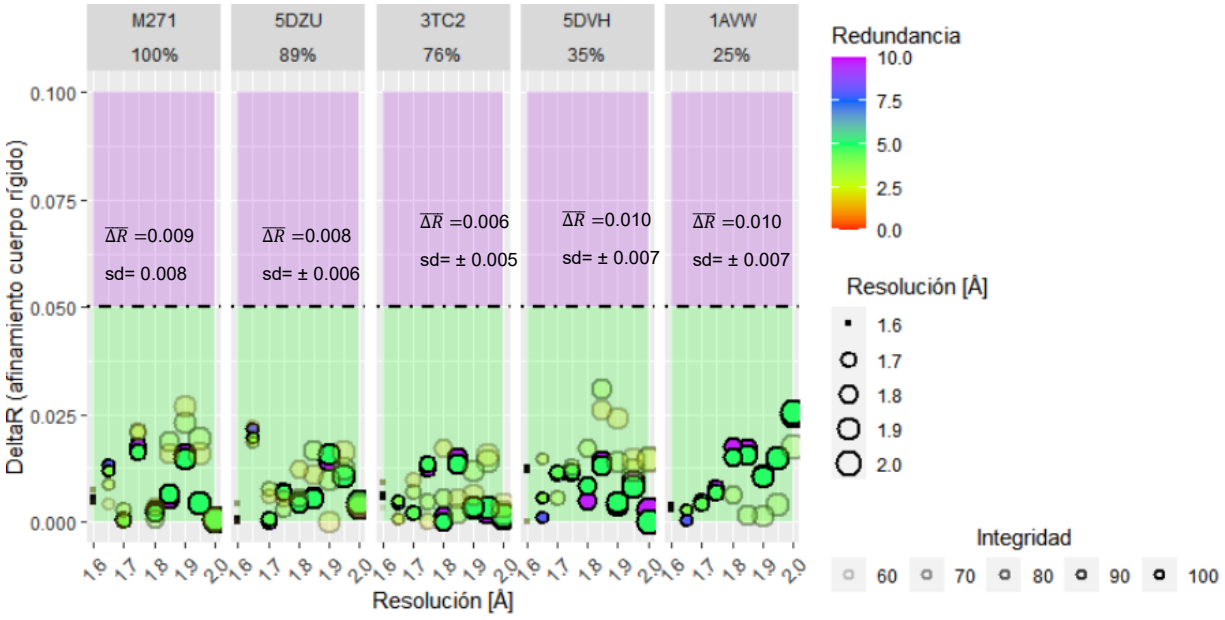



---- $R_{free}$: 0.25 Valores típicos utilizados en el laboratorio.

---- $R_{free}$:0.3-0.5. [45].

**Figura 5.6. Rfree de afinamiento de cuerpo rígido contra resolución.** Se muestran los valores de $R_{free}$ obtenidos de un afinamiento de cuerpo rígido, representados como círculos y su variación respecto a la resolución máxima de corte. Los valores de $R_{free}$ son menores conforme el porcentaje de identidad de aminoácidos es mayor. El área verde representa a los escenarios que cumplen con un valor de $R_{free}$ menor a 0.5 y el área morada representa los escenarios que no cumplen con esta restricción (modelos con el 35% y 25% de identidad de secuencia). También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada modelo base.

En la **Figura 5.7,** están representados los valores de ΔR. A diferencia de otros valores analizados en esta tesis, no se observa una tendencia clara, los valores oscilan entre 0.030 y 0.

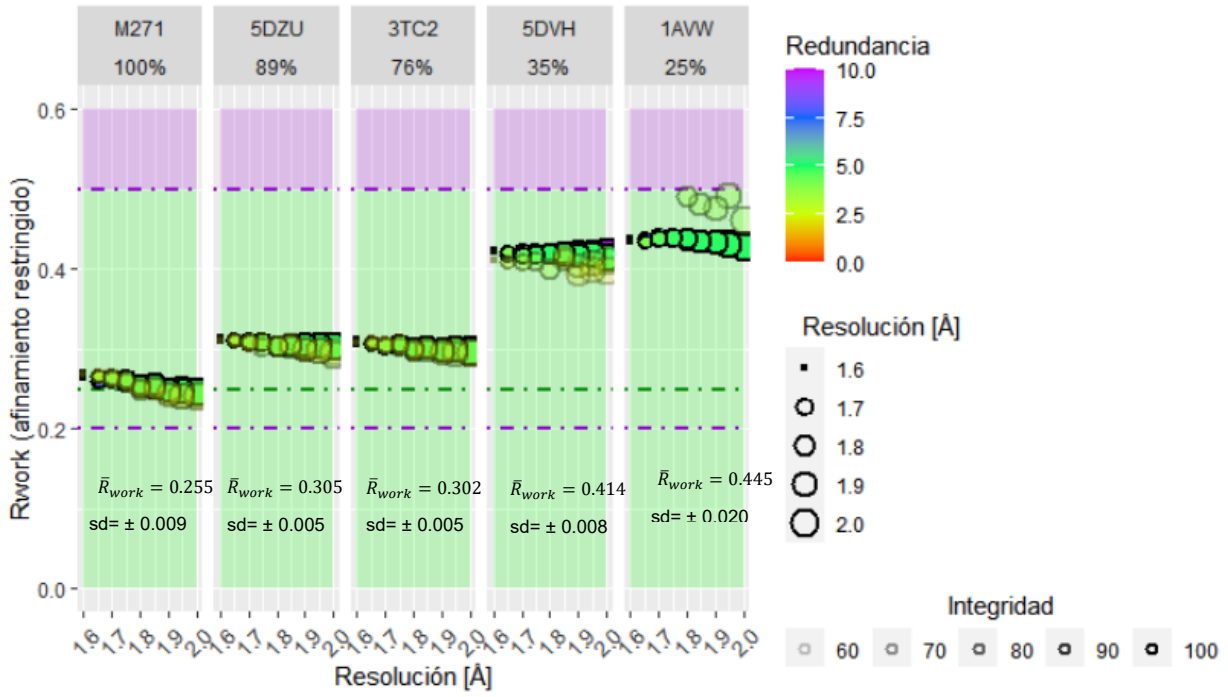



---- **ΔR**: 0.05 [66].

**Figura 5.7. ΔR de afinamiento de cuerpo rígido contra resolución.** Se muestran los valores de ΔR obtenidos de un afinamiento de cuerpo rígido, representados como círculos y su variación respecto a la resolución máxima de corte. Los valores de ΔR no presentan una tendencia; sin embargo, son valores bajos que oscilan entre 0.03 y 0. El área verde representa a los escenarios que cumplen con un valor de ΔR menor a 0.05 (todos los escenarios cumplen) y el área morada representaría los escenarios que no cumplen con esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada modelo base.

Después de realizar el afinamiento de cuerpo rígido se prosiguió a realizar un afinamiento restringido que se muestra en la **Figuras 5.8, 5.9 y 5.10**, donde se muestra el comportamiento de los valores de $R_{work}$, $R_{free}$ y ΔR, respectivamente. Cabe recordar que ningún escenario fue sometido a un afinamiento manual utilizando algún programa gráfico.

En la **Figura 5.8** se muestra que conforme el porcentaje de identidad de aminoácidos va siendo mayor $R_{work}$ tuvo valores más bajos. Lo anterior es lógico ya que esto permite explicar mejor la información experimental. Los valores de $R_{work}$ pertenecientes a los grupos de baja redundancia e integridad de las estructuras de las proteínas con 35% y 25% de identidad de secuencia, son los que presentaron un peor desempeño.

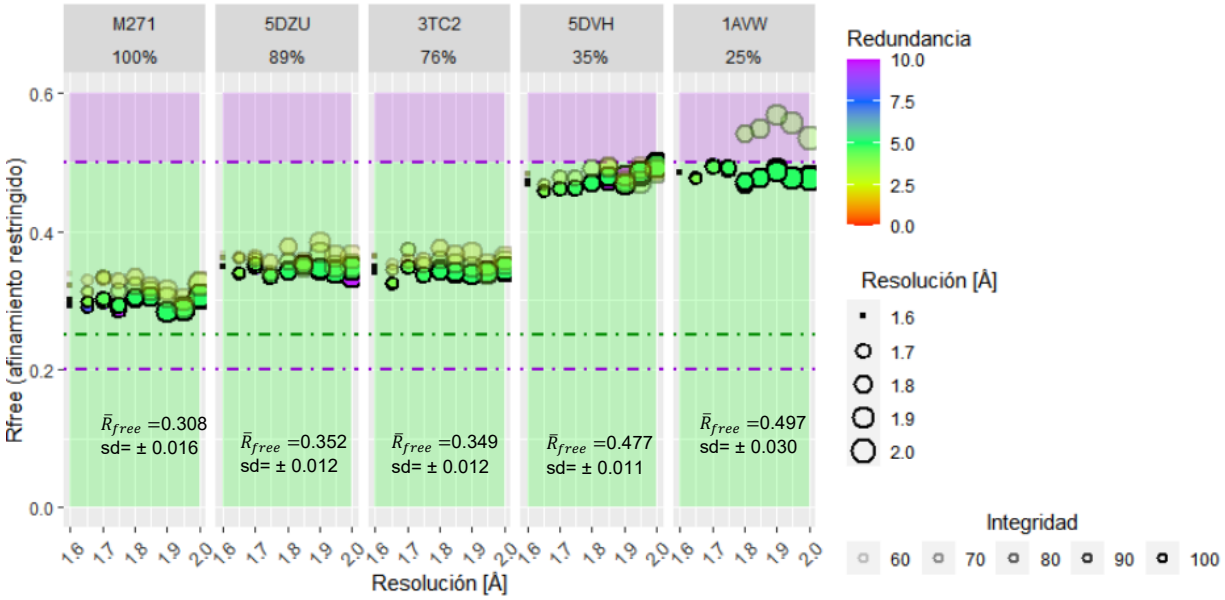



---- $R_{work}$: 0.25 Valores típicos utilizados en nuestro laboratorio.

---- $R_{work}$:0.3-0.5. [45].

**Figura 5.8. $R_{work}$ de afinamiento restringido contra resolución.** Se muestran los valores de $R_{work}$ obtenidos de un macrociclo de afinamiento restringido realizado en Phenix. El diámetro de los círculos indica la variación con respecto a la resolución máxima de corte. Los valores de $R_{work}$ son menores conforme el porcentaje de identidad de aminoácidos es mayor. El área verde representa a los escenarios que cumplen con un valor de $R_{work}$ menor a 0.5 y el área morada muestra a los escenarios que no cumplen esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada moldeo base.

En la **Figura 5.9,** se muestran los valores de $R_{free}$, que fueron aumentando conforme el porcentaje de identidad disminuye. Por tanto se comportan de manera similar a los valores de $R_{work}$. Sin embargo, estos últimos se comportan de una manera más uniforme, además de que son menores a los valores de $R_{free}$ Los valores de $R_{free}$ fueron más elevados en los grupos de menor redundancia e integridad; estos detalles no se percibían en los valores de $R_{free}$ por afinamiento de cuerpo rígido (comportamiento normal para este tipo de afinamientos). Se observa que los valores de $R_{free}$ para el grupo de 60 *frames*, usando la estructura de la proteína de 35% de identidad de secuencia, no cumple con un valor menor a 0.5.




---- $R_{free}$: 0.25 Valores típicos utilizados en nuestro laboratorio.

---- $R_{free}$:0.3-0.5. [45].

**Figura 5.9. $R_{free}$ de afinamiento restringido contra resolución.** Se observan los valores de $R_{free}$ obtenidos de un afinamiento restringido. El diámetro de los círculos cambia con respecto a la resolución máxima de corte. Los valores de $R_{free}$ son menores conforme el porcentaje de identidad de aminoácidos entre el modelo de búsqueda y M271 es mayor. El área verde representa a los escenarios que cumplen con un valor de $R_{free}$ menor a 0.5 y el área morada representa los escenarios que no cumplen con esta restricción (grupo de 60 *frames* para el modelo base con 35%). También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada moldeo base.

En la **Figura 5.10,** se muestran los valores de ΔR, No se observó tendencia alguna, y los valores oscilaron en intervalos amplios, que van de 0.25 hasta casi 0.1. Se puede observar que los grupos de baja redundancia e integridad tuvieron valores superiores a los de alta redundancia e integridad, lo que se debe a que estos últimos grupos tienen más información experimental.

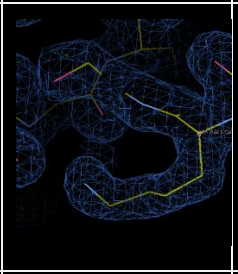
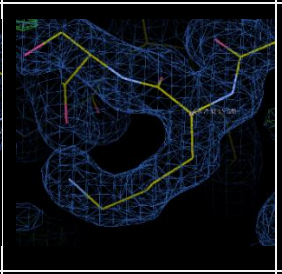
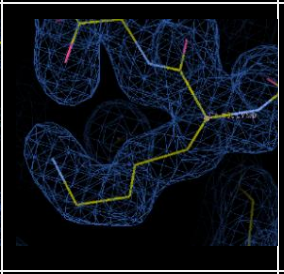
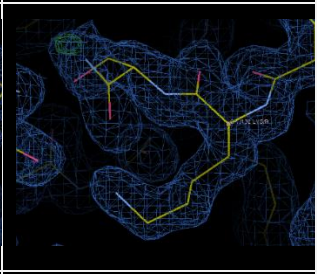
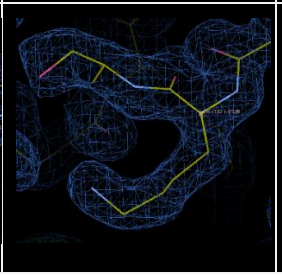
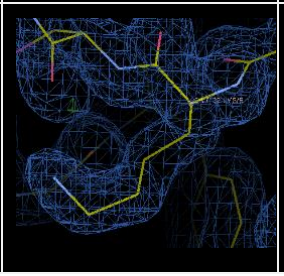
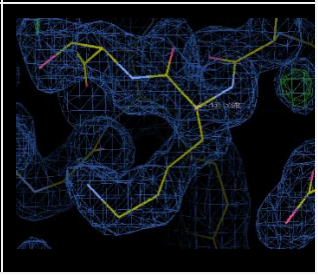
## 5.3 ANALISIS DE LA CALIDAD MODELO TRIDIMENSIONAL Y MAPA DE DENSIDAD ELECTRÓNICA

### 5.3.1 ESCENARIO DE SUSTITUCIÓN MOLECULAR

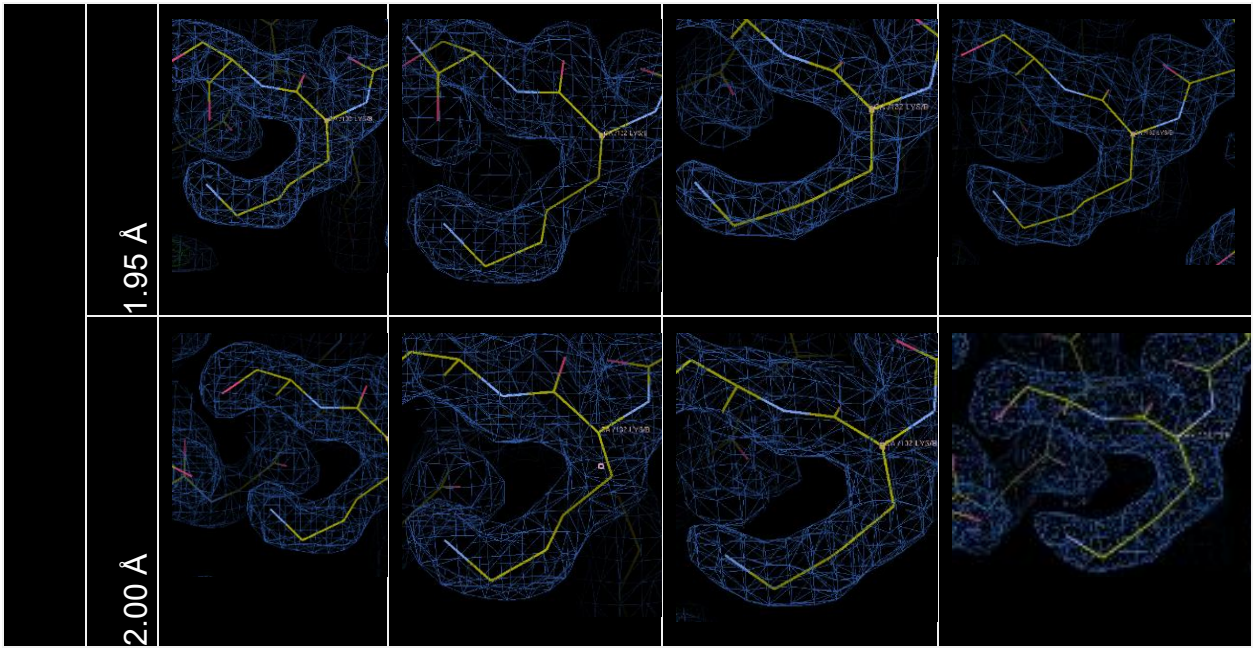
Se analizó la densidad electrónica de los escenarios para la proteína M271, que es el escenario más simple, donde las fases fueron resueltas por sustitución molecular.

En la **Tabla 5.3** se muestran los mapas de densidad electrónica para la Lys 132B, utilizando COOT, y los mapas 2FO-FC (a un valor de sigma de 1) y FO-FC (aun valor de sigma ± 3.5). Se observa que incluso en los escenarios más críticos con resolución de 1.60 Å y con 30 *frames,* los mapas muestran una integridad que permite ajustar al aminoácido. Resultado difícil de distinguir de cualquier otro escenario analizado en esta tesis. Esto significa que todos los escenarios de sustitución molecular tienen información útil y de calidad similar, incluso el caso más simple de sustitución molecular. Dicho de otra forma, este experimento muestra un hecho conocido en cristalografía: las fases correctas son lo más importante para la calidad del mapa.

**Tabla 5.3. Mapas de densidad electrónica 2FO-FC a 1 sigma y FO-FC a 3.5.sigma.** Se muestran los mapas de densidad electrónica del aminoácido Lys 132/B para todos los escenarios del modelo base de sustitución molecular usando las coordenadas finales de M271.

| *LYS 132 B* | | | | | |
|---|---|---|---|---|---|
| | | *Frames* | | | |
| | | 30 | 60 | 120 | 240 |
| RESOLUCIÓN MÁXIMA DE CORTE | 1.60 Å |  |  |  |  |
| | 1.65 Å |  |  |  |  |

1.70 Å
1.75 Å
1.80 Å
1.85 Å
1.90 Å




### 5.7.2 Análisis de la calidad mapa-modelo.

Los modelos y mapas de densidad electrónica al final del afinamiento restringido fueron analizados utilizando la herramienta *Real Space Correlation* en Phenix. Que arrojó valores de coeficiente de correlación (CC) globales y locales. Este coeficiente toma valores de entre 0 y 1; 1 corresponde a un ajuste perfecto del mapa, y cero a que no hay correlación entre el mapa y el modelo. En la **Figura 5.11** se contrasta un valor alto, $CC_{local}$ igual a 1 (LEU 550/B) y otro bajo, $CC_{local}$ igual a 0.5 (PRO 590/B). Además se muestra una imagen del mapa de densidad 2F2-FC a una sigma, con el modelo tridimensional correspondiente. La información mostrada corresponde a la cadena B para el escenario de la estructura 1AVW con 25% de identidad utilizando 60 *frames* a 1.80 Å. Este escenario fue el más problemático de los analizados en esta tesis.

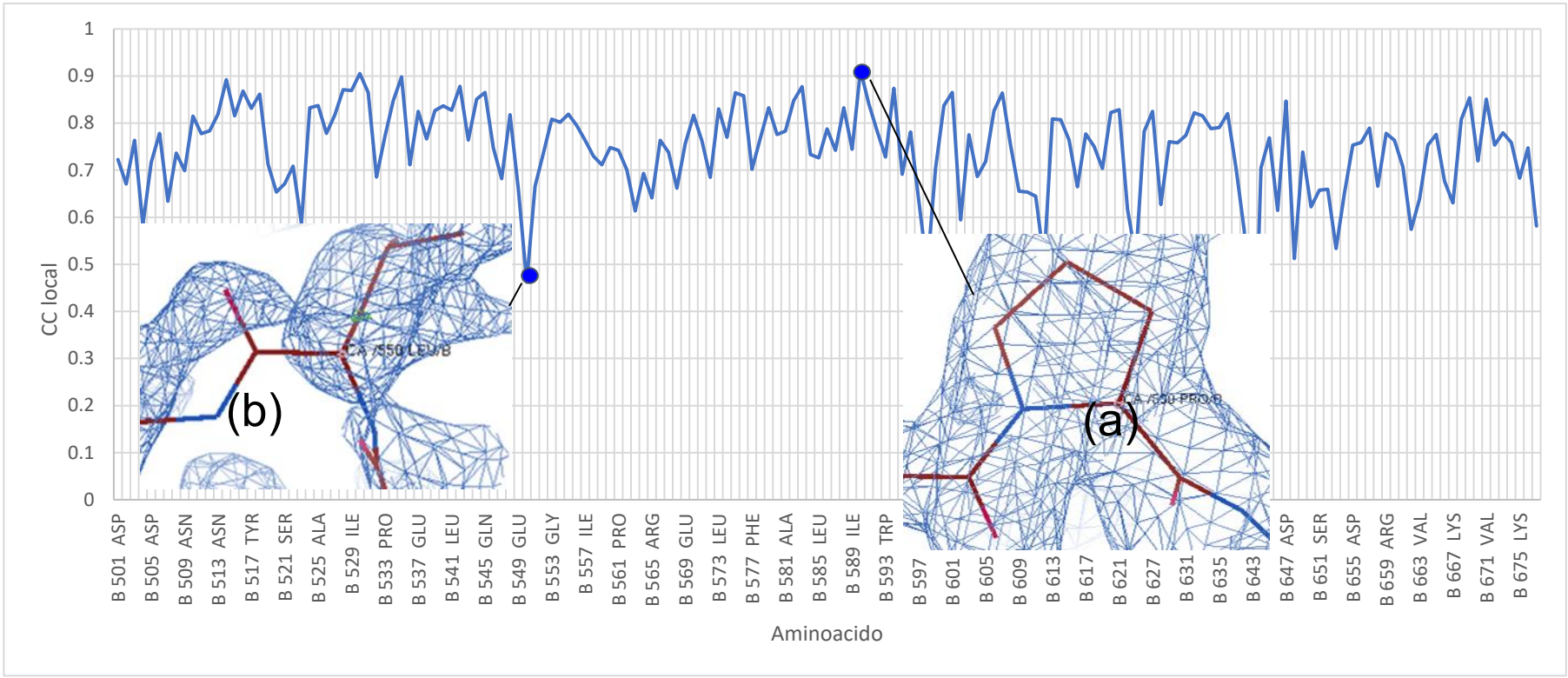
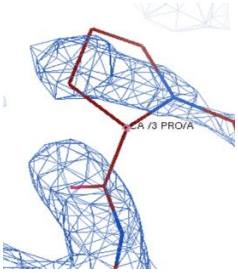
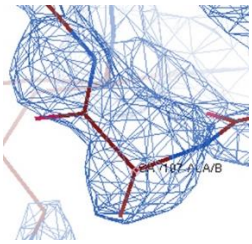
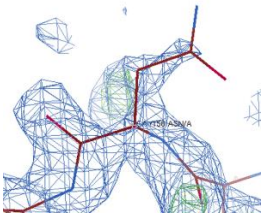
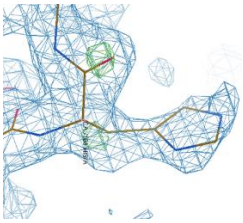



**Figura 5.11. Valores locales de CC para dos escenarios extremos.** (a) B 590 PRO, CC= 0.913, (b) B 550 LEU , CC= 0.462. En esta figura se muestra el gráfico generado con la herramienta *Real Space Correlation* de Phenix. En el eje horizontal se muestra el número de aminoácido correspondiente a la cadena B de cada escenario y en uno de los ejes verticales, en color azul a CC, de 1AVW (25 % de identidad) como modelo base para reemplazo molecular, a una resolución de 1.80 Å, utilizando 60 *frames*, con una redundancia de 3.9, integridad de 62.5%. Además, se muestra el mapa de densidad electrónica visualizado en COOT, utilizando 2FO-FC a 1 sigma y FO-FC a ±3.5 σ.

En la **Tabla 5.4** se muestran los mapas de densidad electrónica y sus respectivos valores de $CC_{local}$. Se puede observar que, de manera local, a partir de un valor cercano a 0.85, se observan mapas de densidad electrónica que ajustan perfectamente con sus respectivos modelos.

**Tabla 5.4. Mapas de densidad electrónica de aminoácidos y su respectivo valor de $CC_{local}$.** Se muestran diferentes valores de $CC_{local}$ de forma creciente y sus respectivos mapas de densidad electrónica. Los mapas de densidad electrónica fueron visualizados en COOT, utilizando 2FO-FC a 1 sigma y FO-FC a 3.5.sigma, y $CC_{local}$ utilizando Phenix de *Real Space Correlation.*

| $CC_{local}$ | Mapa de densidad electrónica | $CC_{local}$ | Mapa de densidad electrónica |
|---|---|---|---|
| 0.700 | M271 (100%)<br>240 *Frames (2.00 Å)*<br>**PRO3/A**<br>  | 0.756 | M271 (100%)<br>240 *Frames (2.00 Å)*<br>**ALA107/B**<br>  |
| 0.802 | M271 (100%)<br>240 *Frames (2.00 Å)*<br>**ASN158/A**<br>  | 0.852 | M271 (100%)<br>240 *Frames (2.00 Å)*<br>**HIS108/A**<br>  |

### 5.7.3 Análisis global de la calidad mapa-modelo respecto a los diferentes indicadores de corte convencionales

La finalidad de este análisis es contrastar el valor de $CC_{global}$ y los indicadores de afinamiento, $R_{work}$, $R_{free}$, $\Delta R$ y contrastar con el caso hipotético del uso de los indicadores de resolución máxima de corte empleados en esta tesis. Se aplicaron los siguientes criterios a todos los escenarios que presentaron solución de fases: $CC_{global} \geq 0.8$, $R_{work} \leq 0.5$, $R_{free} \leq 0.5$, $\Delta R \leq 0.05$. En la **Figura 5.12** se muestra un diagrama de Venn de la distribución de los escenarios que cumplieron con los criterios de calidad del mapa-modelo, y de afinamiento. Sólo 16 de los 162 escenarios en los que se pudieron resolver las fases, cumplieron con los 4 criterios, lo que representan el 9.9% del total de los escenarios. Es decir, cerca del 90% de los escenarios donde hubo resolución de fases, fueron desechados al no cumplir con las restricciones de $CC_{global} \geq 0.8$, $R_{free} \leq 0.5$, $R_{work} \leq 0.5$ y $\Delta R \leq 0.05$ lo que denota que el aplicar estos límites desecha información estructural valiosa.

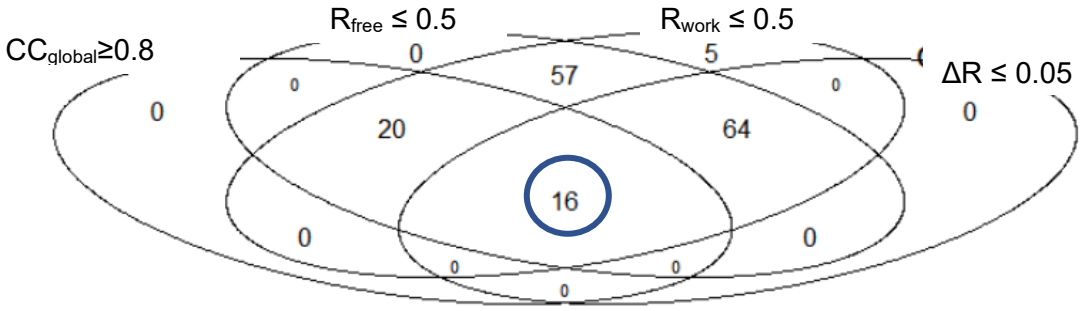



**Figura 5.12. Diagrama de Venn de distribución de escenarios que cumplen con criterios de calidad de mapa-modelo y afinamiento.** Se muestran los escenarios que cumplen con: $R_{free} \leq 0.5$, $R_{work} \leq 0.5$, $\Delta R \leq 0.05$, y CC global >0.7. De los 162 escenarios con fases resueltas, solo 16 cumplieron con estos cuatro criterios. 36 cumplen con $CC_{global} \geq 0.8$, 157 cumplen con $R_{free} \leq 0.5$, 162 cumplen con $R_{work} \leq 0.5$ y 80 cumplen con $\Delta R$.

De estos, 16 escenarios cumplen satisfactoriamente con los siguientes criterios:

$CC_{global} \geq 0.8$, $R_{free} \leq 0.5$, $R_{work} \leq 0.5$, $\Delta R \leq 0.05$, es decir, aquellos escenarios que cumplen con los criterios de calidad del mapa y el modelo son desechados cuando se utiliza un corte de resolución máxima con los distintos indicadores de corte utilizados en esta tesis: $I/\sigma(I)$, $R_{meas}$, $R_{merge}$, $I/\sigma(I)_{mean}$, $R_{pim}$, integridad, redundancia, CC ½. El porcentaje de escenarios útiles desechados por indicador de corte de resolución se muestran en la Figura 5.13, en la que el 100% corresponden a los 16 escenarios útiles.Se observa que el indicador de corte con el porcentaje de escenarios útiles más elevado fue $I/\sigma$ (I), desechando el 75% de escenarios útiles. CC ½ y redundancia fueron los indicadores con mayor asertividad al no desechar ningún escenario útil. Por lo anterior este análisis

parece indicar que utilizar a I /σ como parámetro para definir la resolución máxima puede representar un error para el usuario que implique la pérdida de información estructural útil.

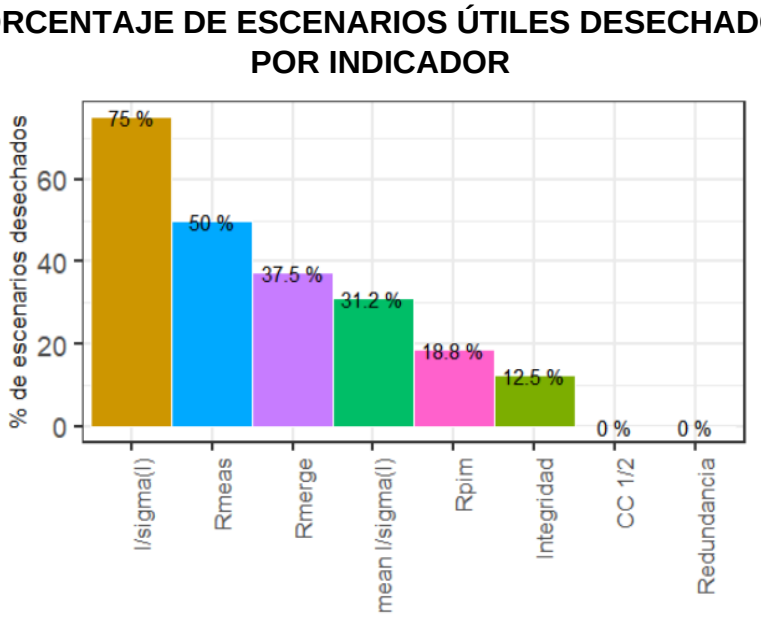



**Figura 5.13 Porcentaje de escenarios desechados por indicador de corte.** Gráfico del porcentaje de escenarios útiles, desechados que cumplen con: $R_{free} \leq 0.5$, $R_{work} \leq 0.5$, $\Delta R \leq 0.05$, y $CC_{global} \geq 0.8$. Estos escenarios desechados tienen información útil para la construcción de un modelo. Los16 escenarios que cumplieron con estos cuatro criterios representan al 100% de todos los escenarios con información útil. En el gráfico se muestra el porcentaje, en orden descendente, de los escenarios exitosos que se hubieran desechado utilizando diferentes indicadores de corte.

### 5.7.4 Análisis local-global de la calidad mapa-modelo respecto a los diferentes indicadores de corte convencionales

Para este análisis se calculó el porcentaje de aminoácidos útiles en cada uno de los escenarios, definiendo a un aminoácido útil como aquel que tiene un valor de $CC_{local}$ ≥0.85, como se muestra en la **Tabla 5.4.** Conociendo el número de aminoácidos que cumplen con esta condición, y conociendo el número total de aminoácidos de la proteína, se calcula el porcentaje de aminoácidos útiles para cada escenario, en esta tesis utilizamos 35% como un valor límite.

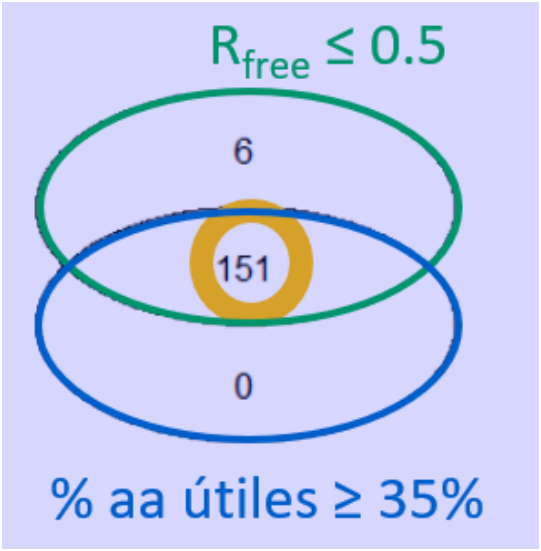
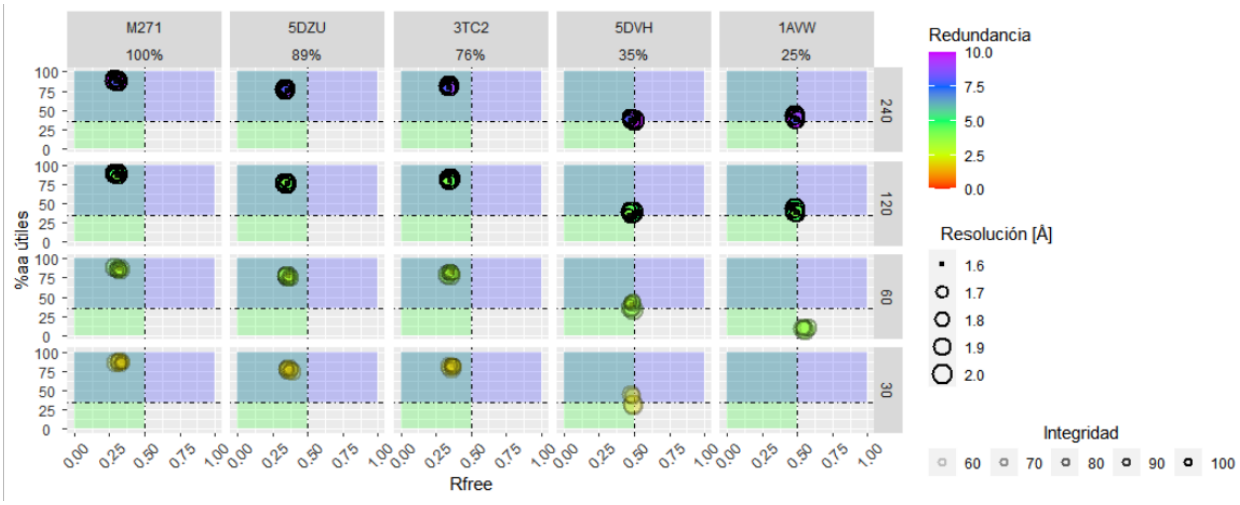



**Figura 5.14. Diagrama de Veen de distribución de escenarios que cumplen con criterios %aa útiles ≥35% y $R_{free}$≥0.5**.

De los 162 escenarios con fases resueltas151 cumplieron con estos dos criterios

En la **Figura 5.14** se muestra un diagrama de Venn de la distribución de los escenarios que cumplieron con los criterios %aa útiles≥35% y $R_{free}$ ≤ 0.5. Sólo151 escenarios de los 162 en los que se pudieron resolver las fases, cumplieron con estos dos criterios. Lo anterior representa el 93% del total de los escenarios, es decir, poco menos del 7% fueron desechados al no cumplir con las restricciones de $CC_{local}$≥0.85, $R_{free}$ ≤ 0.5, $R_{work}$ ≤ 0.5, ΔR ≤ 0.05. Los cuales se consideran escenarios con información cristalográfica útil. En la **Figura 5.15** se muestra la relación del porcentaje de aminoácidos útiles contra $R_{free}$. En este gráfico, se puede observar que cuando se resuelven fases con modelos de bajo porcentaje de identidad y además con baja redundancia e integridad de datos de difracción, ya no se tiene información útil que permita obtener un modelo tridimensional adecuado. Es decir, conforme el valor de $R_{free}$ va aumentando debido a que las diferencias entre los factores observados y calculados se va haciendo mayor, la calidad del mapa empeora al punto de volverse inútil para construir un modelo.




%aa útiles ≥ 35%

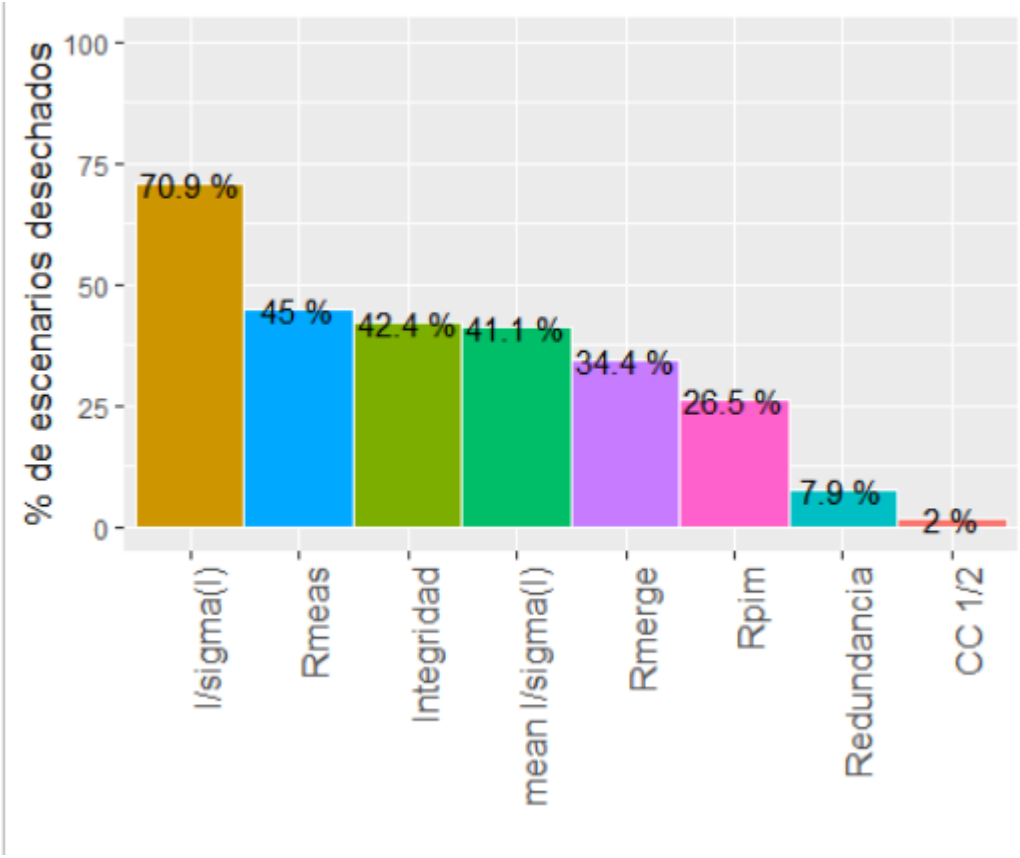
$R_{free} \leq 0.05$[66].

**Figura 5.15. Porcentaje de aminoácidos útiles versus valores de $R_{free}$.** Los valores de $R_{free}$ finales, se obtuvieron después de un afinamiento restringido y se representan como círculos. Se puede observar que conforme los valores de $R_{free}$ aumentan también lo hacen los valores del %aa útiles. El área verde representa a los escenarios que cumplen con un valor de $R_{free}$ menor a 0.5 y el área azul son los escenarios que cumplen con un %aa útiles de 35%. Los escenarios que están en las intersecciones son los que cumplen con ambas restricciones.

El porcentaje de escenarios útiles desechados por indicador de corte de resolución se muestran en la Figura 5.16. El 100% corresponden a los 151 escenarios útiles que cumplieron con los criterios: $R_{free} \leq 0.5$ y %aa ≥ 35%.

Se observa que el indicador de corte con el porcentaje de escenarios útiles desechados más elevado fue I/σ(I), el cual desecha el 70% de escenarios útiles, mientras que CC ½ fue el indicador con mayor asertividad, al desechar solo el 2% de los escenarios que dieron lugar a un mapa de densidad útil para construir un modelo cristalográfico.

**PORCENTAJE DE ESCENARIOS ÚTILES DESECHADOS POR INDICADORES DE CORTE**




$I/\sigma(I) > R_{meas} >$ Integridad$> I/\sigma(I)_{mean} > R_{merge} > R_{pim} >$Redundancia$>$ CC ½

[70.9%] [45%] [42.4%] [41.1%] [34.4%][26.5%] [7.9%] [2%]

**Figura 5.16 Porcentaje de escenarios desechados por indicadores de corte**. Gráfico del porcentaje de escenarios con información cristalográfica útil para construir modelos y desechados por los indicadores marcados al pie de cada columna, a pesar de cumplir con: Rfree≤0.5 y %aa útiles≥35%. El porcentaje mostrado en el gráfico representa el porcentaje, en orden descendente, de escenarios desechados por cada indicador mostrado al pie de cada columna y que hubieran dado lugar a mapas de densidad aptos para construir un modelo.

## 5.4 RMSD

Este análisis se realizó comparando las posiciones de los carbonos alfa de cada modelo tridimensional analizado, utilizando la función LSQ Superpose dentro de Coot [63]. Se tomó la cadena B del modelo final de M271 como coordenada base, y se comparó contra la cadena B de cada modelo obtenido en cada condición analizada en esta tesis. Los valores de RMSD se muestran en la **Figura 5.17**, donde se observa que el valor de

RMSD va aumentando conforme el porcentaje de identidad de aminoácidos va disminuyendo. Se pueden observar valores de RMSD bajos, menores a 0.5Å, para el caso de los escenarios de sustitución molecular. El número de carbonos alfa alineados se muestran en la Figura 5.18, donde se observa que estos disminuyen conforme disminuye el porcentaje de identidad de secuencia entre M271 y los modelos cristalográficos usados para el reemplazo molecular, de la tal forma, que como era de esperarse, las coordenadas con menor identidad y/o menor cobertura con respecto a M271, son también las que muestran un RMSD más alto.

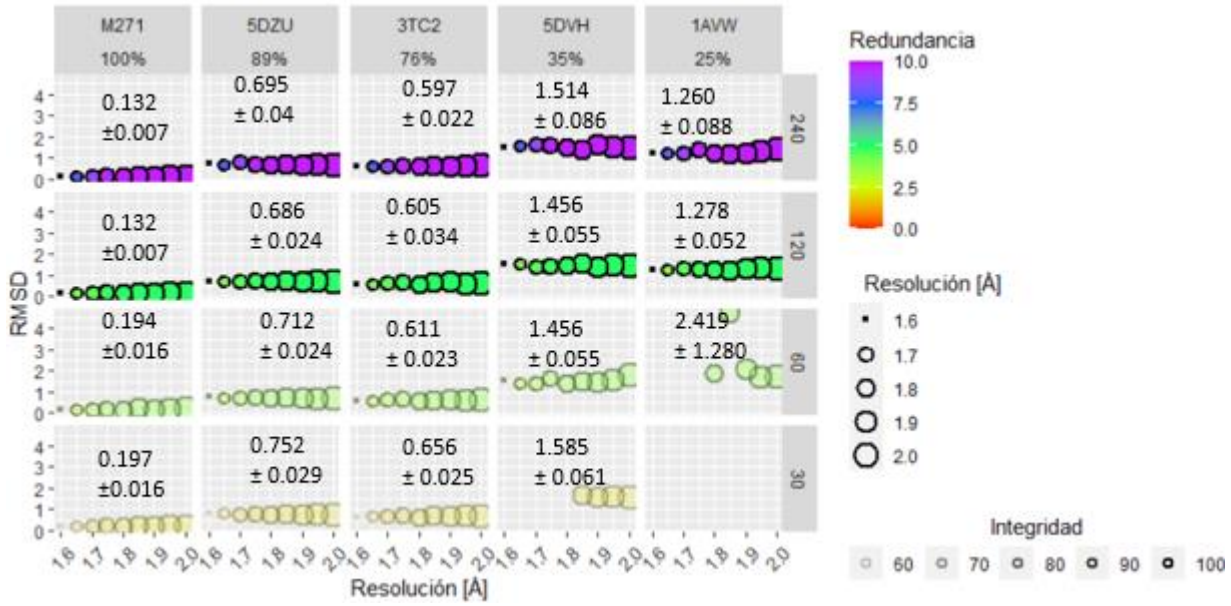



**Figura 5.17. RMSD**. Se muestran los valores de RMSD obtenidos utilizando la herramienta LSQ Superpose del software COOT [63] y representados como círculos, así como su variación respecto a la resolución máxima de corte, la cual se indica con el cambio en el diámetro del círculo. Los valores de RMSD se obtuvieron comparando las coordenadas de la cadena B del modelo original de M271 con cada uno de los modelos de los diferentes escenarios.

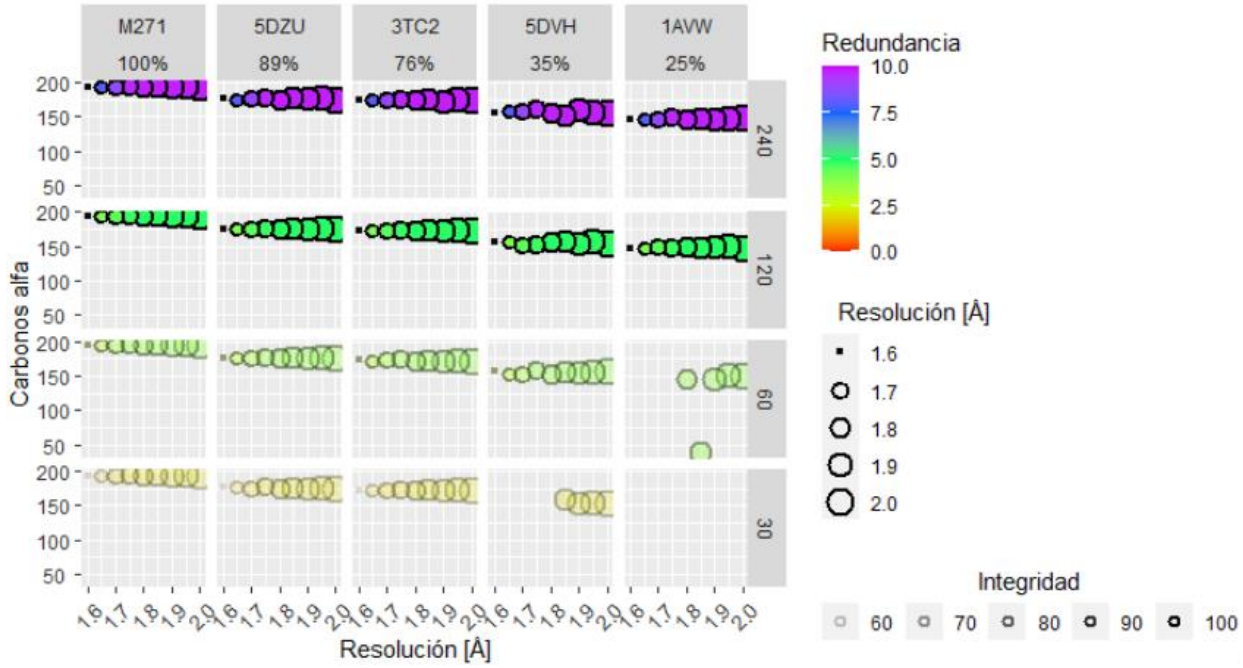



**Figura 5.18. Carbonos alfa.** Se muestran el número de carbonos alfa alineados, representados como círculos y su variación respecto a la resolución máxima de corte, la cual cambia con respecto al diámetro del círculo empleado para representarla.

Los modelos de todos los escenarios se muestran en el anexo D.

## 6. CONCLUSIONES

Después de analizar los resultados de esta tesis, podemos confirmar que la hipótesis planteada en este proyecto es cierta, ya que el uso de los parámetros tradicionales para la determinación de la resolución máxima en estructuras cristalográficas, desecha información útil que permitirían al cristalógrafo aumentar el detalle de la descripción molecular, e indirectamente, mejorarían la calidad del modelo generado del proceso de determinación estructural útil para los investigadores que utilicen a estos modelos cristalográficos.

En un escenario de sustitución molecular, utilizar cualquier indicador de corte para la elección de una resolución máxima, estaría desechando datos útiles, que servirían para la construcción de una estructura cristalográfica con mayor detalle. Si las fases son correctas, cualquier otra métrica usada como criterio para desechar datos cristalográficos es un error que repercutirá en el nivel de detalle de la estructura cristalográfica determinada. Respecto a los escenarios de reemplazo molecular, al utilizar el I/σ(I) como parámetro de corte, una de las métricas más comunes en el campo de la cristalografía de proteínas, se desecha información útil. Por lo que su uso, debería de evitarse en favor de otros valores como lo es CC1/2. De acuerdo con los datos generados en esta tesis,

un escenario es considerado una estructura final “útil” si cumple con los siguientes indicadores, $CC_{Global}$> = 0,8, $R_{work}$ ≤0.5, $R_{free}$≤0.5 y ΔR≤0.05).

Como resultado de los estudios realizados en esta tesis, se muestra claramente que CC ½ es el mejor parámetro de corte de máxima resolución de máxima en la mayoría de los casos, pero en casos donde se tienen porcentajes de identidad de aminoácidos bajos (30% -25%), $R_{merge}$, $R_{meas}$ y $R_{pim}$ también pueden funcionar para delimitar la resolución máxima. De nuestros estudios concluimos que por simplicidad CC1/2 debe ser utilizado de manera preferente para determinar el límite de resolución máxima, y dejar de usar otras métricas que resulta en perdida, sin sustento matemático, de información importante para mejorar la calidad y robustes del modelo cristalográfico obtenido.

## 7. PERSPECTIVAS

Probar que las conclusiones de esta tesis son las mismas si se amplía el universo de datos y estructuras cristalográficas analizadas.

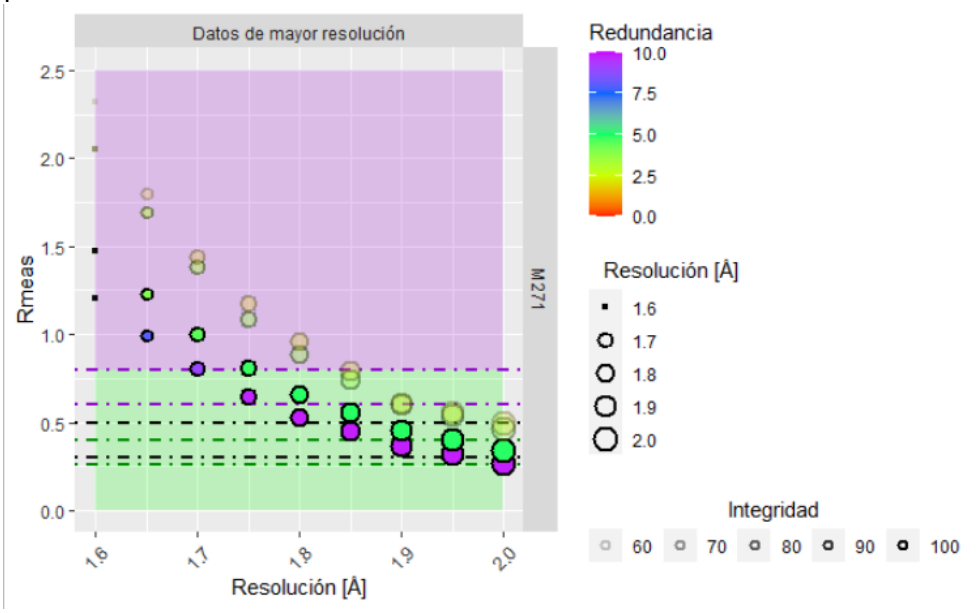
## ANEXO A. SECUENCIAS UTILIZADAS PARA ALINIAMIENTO DE LAS ESTRUCTURAS DE PROTEÍNAS PARA RESOLVER EL PROBLEMA DE FASES

| Código PDB (% de identidad) | Secuencia de aminoácidos para alineamiento |
|---|---|
| 1AVW (25% identidad) | DFVLDNEGNPLENGGTYYILSDITAF<br>GGIRAAPTGNERCPLTVVQSRNEL<br>DKGIGTIISSPYRIRFIAEGHPLSLKF<br>DSFAVIMLCVGIPTEWSVVEDLPEG<br>PAVKIGENKDAMDGWFRLERVSEF<br>NNYKLVFCPQDKCGDIGISIDHDDG<br>TRRLVVSKNKPLVVQFQKLD |
| 5DVH (35% identidad) | VSPPVLDMDGEPLKIDEEYSIISIPF<br>GGGSVYLANLGNTKCPNGVVQDSS<br>NKTPVLFYTMKLGSHFVSENQDVSI<br>KFSTKSCINETVWKVAYSIVGPTHS<br>PLRFVITGGTFGFPGPNNIENWFKIE<br>KYETGRPHSYKLRYCPSQYICPTC<br>QFDCADVGLYENKGYARLALNNKP<br>YPFGFSKVNKN |
| 3TC2 (76% identidad) | PSDATPVLDVTGKELDPRLSYRIIST<br>FWGALGGDVYLGKSPNSDAPCAN<br>GVFRYNSDVGPSGTPVRFIGSSSH<br>FGQGIFEDELLNIQFAISTSKMCVSY<br>TIWKVGDYDASLGTMLLETGGTIGQ<br>ADSSWFKIVKSSQFGYNLLYCPVD<br>QFCLKVGVVHQNGKRRLALVKDNP<br>LDVSFKQVQ |
| 5DZU (89% identidad) | SPLPKPVLDTNGKELNPNSSYRIISI<br>GRGALGGDVYLGKSPNSDAPCPD<br>GVFRYNSDVGPSGTPVRFIPLSGGI<br>FEDQLLNIQFNIATVKLCVSYTIWKV<br>GNLNAYFRTMLLETGGTIGQADSSY<br>FKIVKLSNFGYNLLYCPITPPFLCPF<br>CRDDNFCAKVGVVIQNGKRRLALV<br>NENPLDVLFQEV |

| | |
|---|---|
| M271 (100% identidad) * | SPLPKPVLDTNGKKLNPNSSYRIIST FWGALGGDVYLGKSPNSDAPCPD GVFRYNSDVGPSGTPVRFIPLSGA NIFEDQLLNIQFNIPTVKLCVSYTIWK VGNINAHLRTMLLETGGTIGQADSS YFKIVKSSKFGYNLLYCPLTRHFLCP FCRDDNFCAKVGVVIQNGKRRLAL VNENPLDVLFQEV |

* M271 es el nombre de la proteína y no es un código del PDB.

## Anexo B. Gráficos de indicadores.

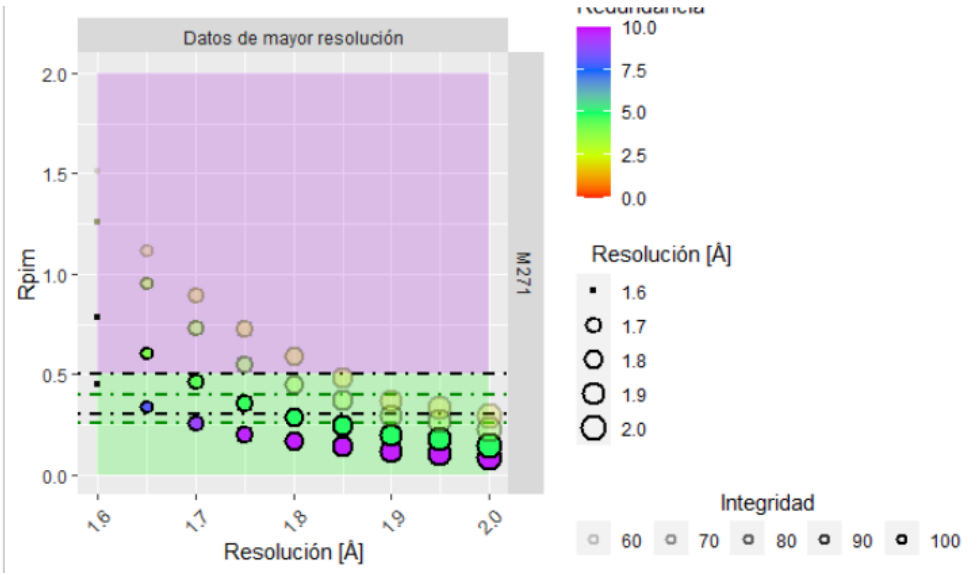



---- $R_{meas}$:0.3-0.5. Valores típicos usados en nuestro laboratorio.

---- $R_{meas}$: 0.26-0.4. [14].

---- $R_{meas}$:0.6-0.8. [35].

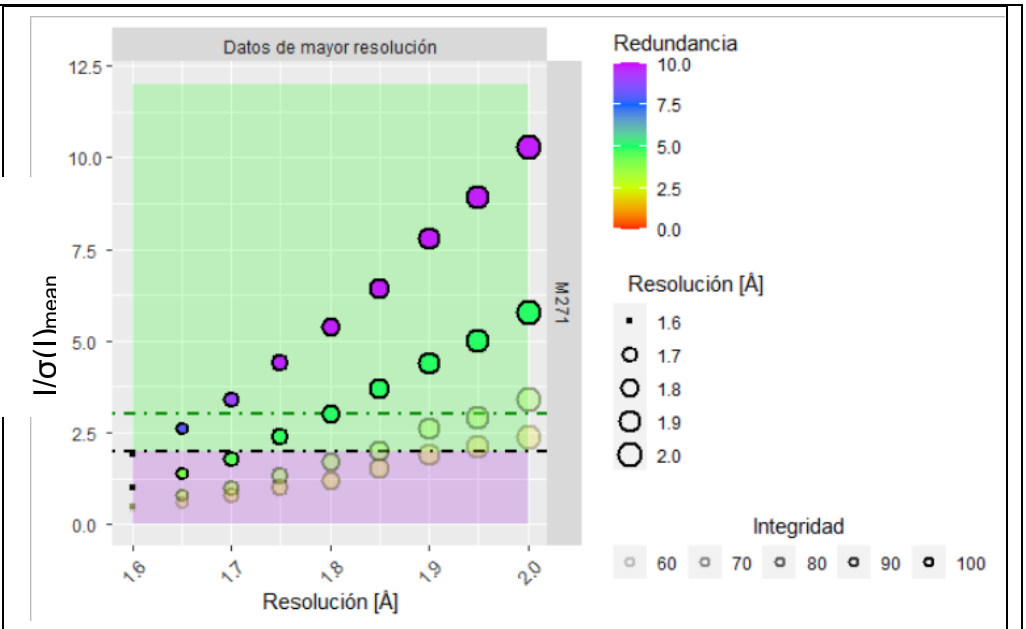
**Figura B.1. $R_{meas}$ vs Resolución.** El valor de redundancia representa por la variación de color, la resolución, por tamaño del punto y la integridad por la nitidez del punto. Se observa que el valor de $R_{meas}$ va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más elevados de $R_{meas}$, que los grupos de alta redundancia e integridad, haciéndose la brecha menor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de $R_{meas}$ de 0.8. Y el área morada representa los escenarios que serían desechados con un valor de $R_{meas}$ de 0.8.




---- $R_{pim}$:0.3-0.5. Valores típicos utilizados en nuestro laboratorio.

---- $R_{pim}$: 0.26 – 0.4 [14].

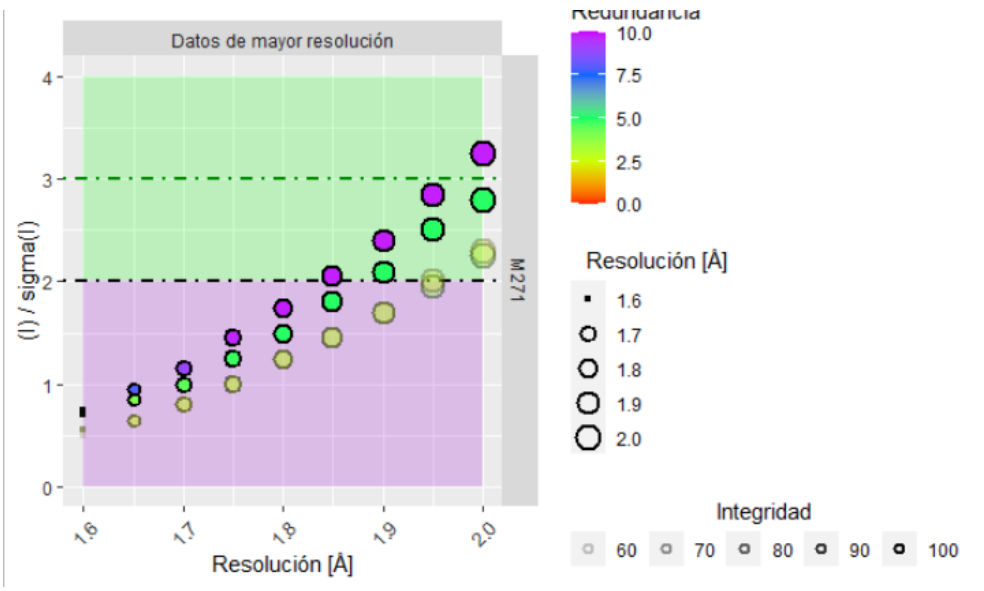
**Figura B.2. $R_{pim}$ vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de $R_{pim}$ va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más elevados de $R_{pim}$, que los grupos de alta redundancia e integridad, haciéndose la brecha menor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de $R_{pim}$ de 0.5. Y el área morada representa los escenarios que serían desechados con un valor de $R_{pim}$ de 0.8.




---- $(I)/\sigma(I)_{mean} \geq 2$ [35].

---- $(I)/\sigma(I)_{mean} \geq 3$ [37].

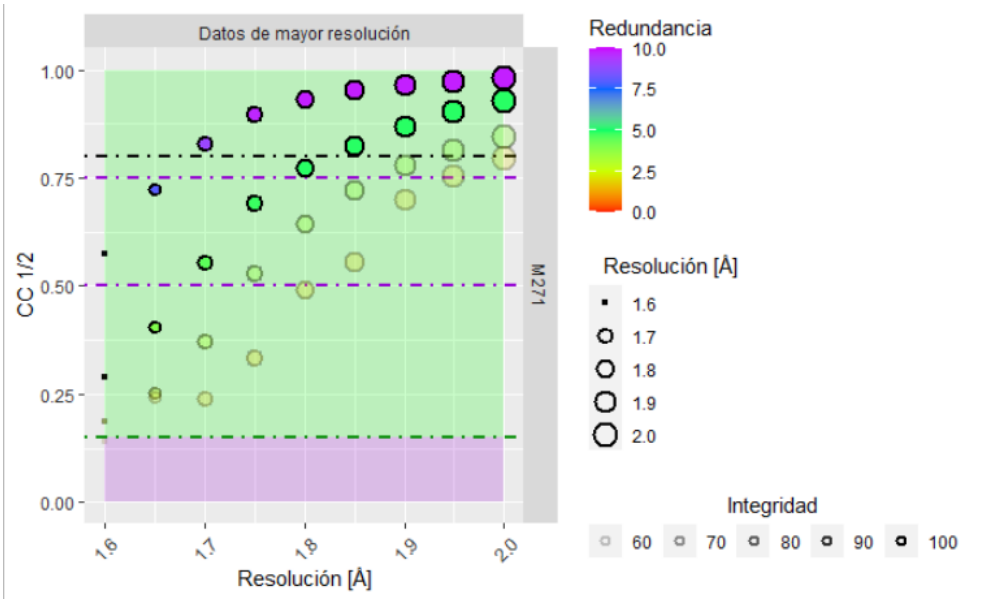
**Figura B.3. $I/\sigma(I)_{mean}$ vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de $I/\sigma(I)_{mean}$ va aumentando conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de $I/\sigma(I)_{mean}$, que los grupos de alta redundancia e integridad, haciéndose la brecha mayor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de $I/\sigma(I)_{mean}$ de 2.0. Y el área morada representa los escenarios que serían desechados con un valor de $I/\sigma(I)_{mean}$ de 2.




**----** (I)/σ(I) ≥ 2 Valores típicos utilizados en nuestro laboratorio.

**----** (I)/σ(I) ≥ 3 [14].

**Figura B.4. I/σ(I) vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de I/σ(I) va aumentando conforme la resolución va bajando. Los grupos de baja redundancia tienen valores menores de I/σ(I), que los grupos de alta redundancia e integridad, haciéndose la brecha mayor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de I/σ(I) de 2.0. Y el área morada representa los escenarios que serían desechados con un valor de I/σ(I) de 2.

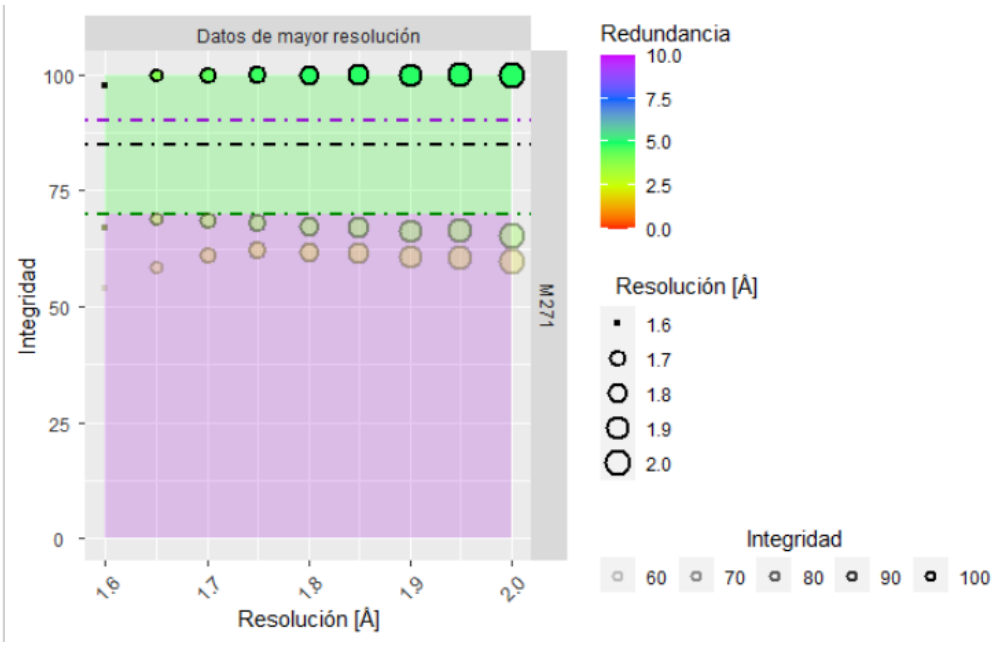



**----** CC ½ ≥:0.8 Valores típicos utilizados en nuestro laboratorio.

**----** CC ½ ≥:0.15[14].

**----** CC ½ ≥:0.5-0.75. [38].

**Figura B.5. CC½ vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de CC½ va aumentando conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de CC½, que los grupos de alta redundancia e integridad, haciéndose la brecha menor conforme la resolución va disminuyendo. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de CC½ de 0.15. Y el área morada representa los escenarios que serían desechados con un valor de CC½ de 0.15.

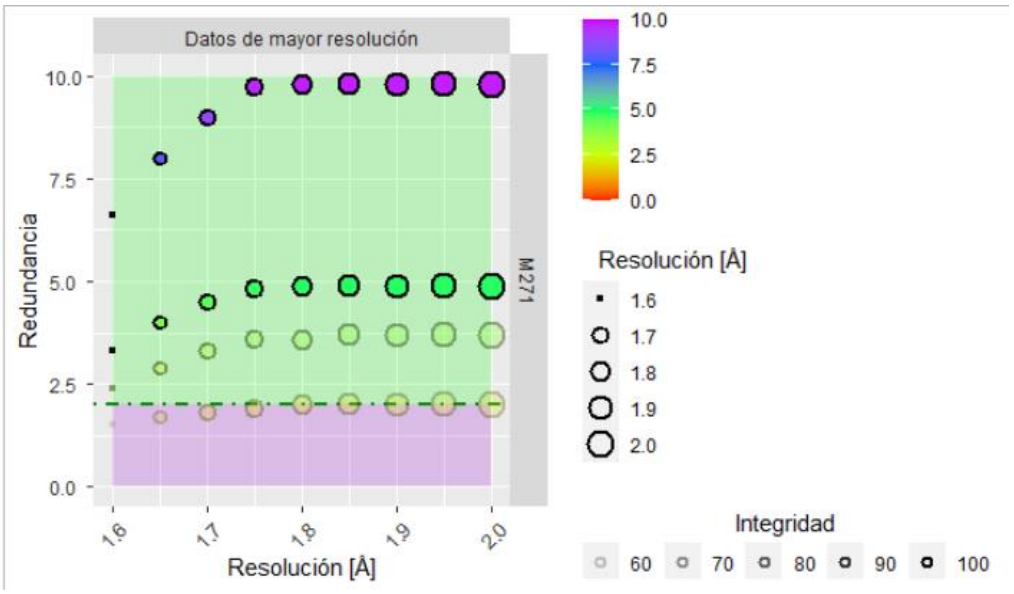



**----** Integridad ≥ 85%. Valores típicos utilizados en nuestro laboratorio.

**----** Integridad ≥ 70% [35].

**----** Integridad ≥ 90% [35].

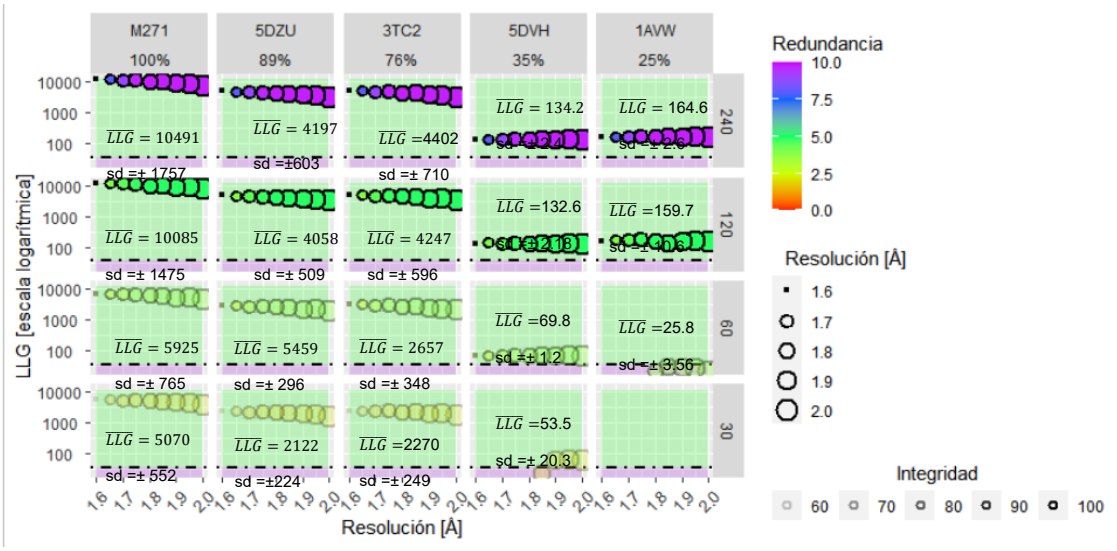
**Figura B.6. Integridad vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa que el valor de la integridad va aumentando conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de integridad, que los grupos de alta redundancia e integridad. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de integridad de 70%. Y el área morada representa los escenarios que serían desechados con un valor de integridad de 70% que corresponden a los escenarios de baja redundancia e integridad (utilizando 30 y 60 *Frames*).




---- Redundancia ≥ 2. [40].

**Figura B.7. Redundancia vs Resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de redundancia va aumentando conforme la resolución va bajando. Los grupos de baja redundancia tienen valores menores de redundancia, que los grupos de alta redundancia e integridad. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que serían útiles utilizando un valor de redundancia de 2.0. Y el área morada representa los escenarios que serían desechados con un valor de redundancia de 2.0.

## Resolución de Fases



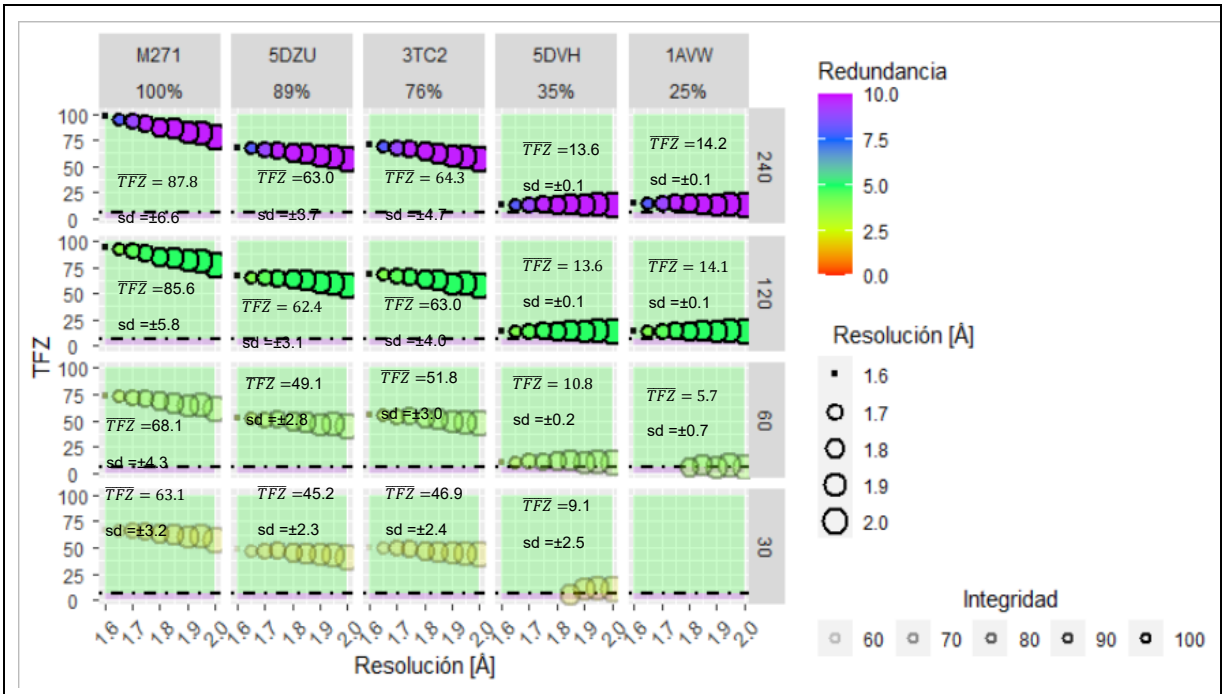

**----** LLG >36. (A. McCoy, 2014).

**----** LLG >40 [41].

**----** LLG >70. [44].

**Figura B-8. LLG en escala logarítmica contra resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de LLG va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de LLG, que los grupos de alta redundancia e integridad. El área verde representa a los escenarios que

cumplen con un valor de LLG mayor a 36 y el área morada representa los escenarios que no cumplen con esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada moldeo base. Las líneas punteadas simbolizan los límites encontrados en la literatura.




---- FTZ >6 [41]

**Figura B-9. FTZ en escala logarítmica contra resolución.** Donde el valor de redundancia queda representado por la variación de color, la resolución, por tamaño del punto e integridad por la nitidez del punto. Se observa valor de FTZ va disminuyendo conforme la resolución va bajando. Los grupos de baja redundancia tienen valores más bajos de FTZ, que los grupos de alta redundancia e integridad. Las líneas punteadas simbolizan los límites encontrados en la literatura. El área verde representa a los escenarios que cumplen con un valor de FTZ mayor a 6 y el área morada representa los escenarios que no cumplen con esta restricción. También se muestran los valores promedio y la desviación estándar para todos los escenarios de cada moldeo base. Las líneas punteadas simbolizan los límites encontrados en la literatura.

## Anexo C. Descripción del caso de estudio:M271. 100% de identidad

La siguiente sección fue escrita tomando como guía a la información reportada por Campuzano [1].

En el estudio titulado: “Estudio estructural del asa L9 de la proteína M271 de *Solanum tuberosum,* un inhibidor de proteasas de la familia Kunitz-STI”, se tiene como objeto de estudio al inhibidor M271 de la familia Kunitz-STI[M], extraído de *Solonum tuberosum,* el cual es una variante de dos inhibidores: E3ad (inhibidor bifuncional bifuncional de proteasas serínicas y aspárticas)[N] y PDI (inhibidor de catepsina D de papa)[O].

Este inhibidor se estudió por medio de cristalografía de Rayos X, los patrones de difracción se obtuvieron a una longitud de 0.9791 Å, cuyo grupo espacial es el P21212 con las dimensiones de la celda unitaria: A≈62, B≈ 76, C≈ 820, α=90°, β=90° , γ=90°. Las principales consideraciones post afinamiento fueron: CC ½ ≥80% , I/σ ≥2 y una integridad mayor a 99%.

Superpusieron las tres estructuras tridimensionales (M271, E3Ad y PDI). Se observó que el asa L9 posee conformaciones diferentes, la cual corresponde a un asa, ubicada en diferentes regiones, para la proteína M271, L9 se encuentra precedida por un puente disulfuro entre las cisteínas 151 y 154.

En este estudio, también se realizaron comparaciones de RMSD entre estas tres proteínas, sin incluir al L9, el cuál oscila entre 0.69 para la proteína E3Ad y 0.56 para la proteína PDI.

También se realizó un análisis anisotrópico utilizando los datos cristalográficos de la proteína M271 a la resolución máxima de 1.7 Å.

Se eligió esta estructura por ser el cristal con mayor resolución, por ello presenta una mayor relación de datos-parámetros. Análisis por anisotrópico se realizó un seccionamiento de la cadena de M271:

Tabla C-1 Seccionamiento de la cadena A y B para análisis anisotrópico.

| Cadena | Sección | Intervalo aminoácidos | No. Aminoácidos por sección |
|---|---|---|---|

[M] La familia Kunitz, se distingue por inhibir mayormente a las proteasas serínicas,pero también actúa sobre proteasas aspárticas y cisteínicas [1].
[N] Es un inhibidor de soya, bifuncional del tipo Kunitz, presenta glicosilaciones capaces de inhibir proteasas de seriana y aspártico Se estudiaron 6 estructuras que presentan un arreglo tipo trébol β [67].
[O] Inhibidor de catepsina de papa. Es una glicoproteína de 188 aminoácidos que inhibe a la proteasa aspártica catepsina D. Su resolución máxima por la técnica de Rayos X, es de 2.1 Å, Presenta un arreglo tipo trébol β [68]

| | | | |
|---|---|---|---|
| Cadena A | A | 4-40 | 37 |
| | B | 41-45 | 5 |
| | C | 46-146* | 101 |
| | D | 147-159 ** | 13 |
| | E | 160-178*** | 19 |
| | F | 179-194 | 16 |
| Cadena B | A | 2-27 | 26 |
| | B | 28-46 | 19 |
| | C | 47-103 | 57 |
| | D | 104-110 | 7 |
| | E | 111-151* | 41 |
| | F | 152-158** | 7 |
| | G | 159-194*** | 36 |

*Zona inicial del asa L9

** Asa L9

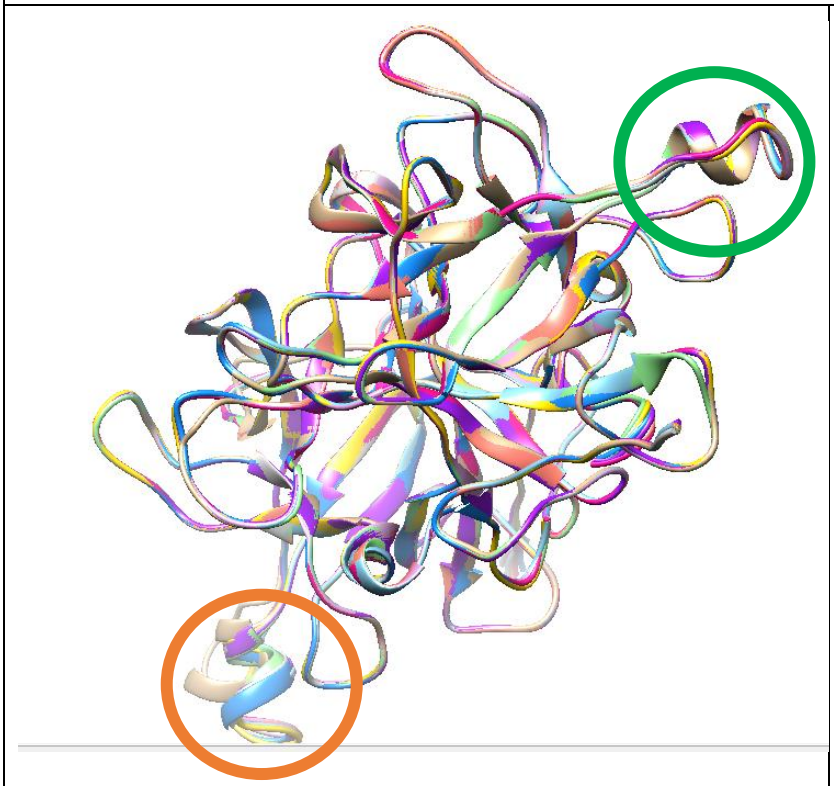
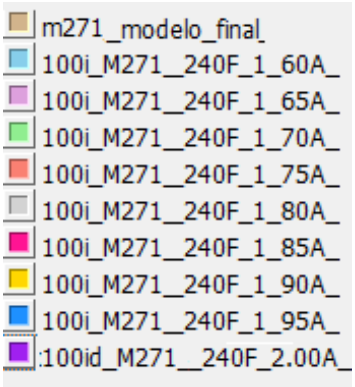
*** Zona final del asa L9

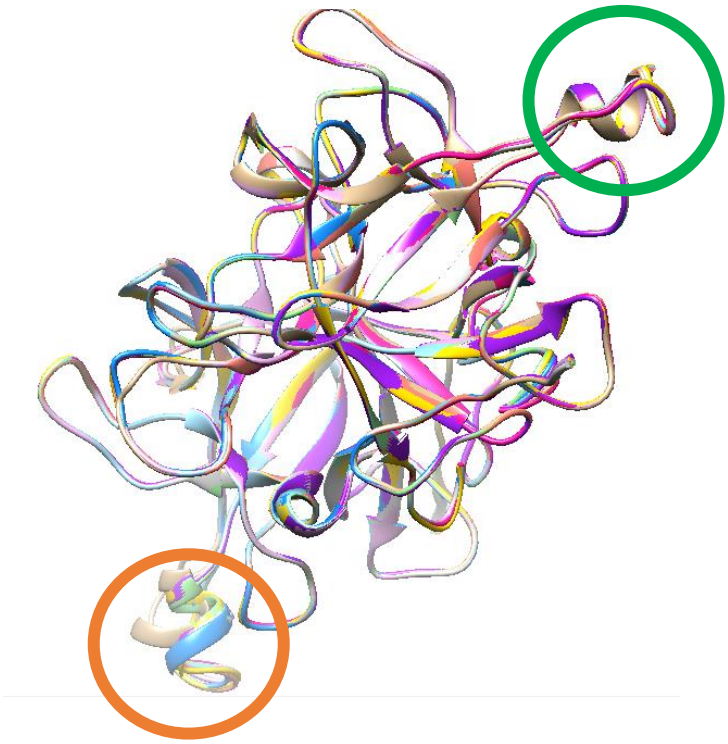
De este estudio se observó que la estructura cristalográfica de M271, tiene poca flexibilidad en el núcleo, y las asas presentan mayor flexibilidad, en especial en la cadena B, por el ambiente 'químico" que tiene a su alrededor, presentando impedimentos estéricos, por mencionar.

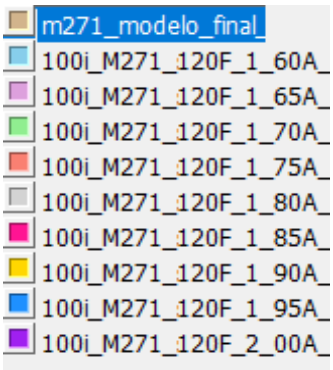
También se realizaron simulaciones de dinámica molecular (para mayor información consultar Campuzano, 2019. Donde se realizaron dinámicas de minimización de energía. Donde a nivel global para el caso de M271, no se ve mayormente afectado comparado con el inhibidor PDI.

**Anexo D. Estructuras finales sobrelapadas con el modelo final de la proteína M271**

Se muestran en círculos las diferencias más notables.

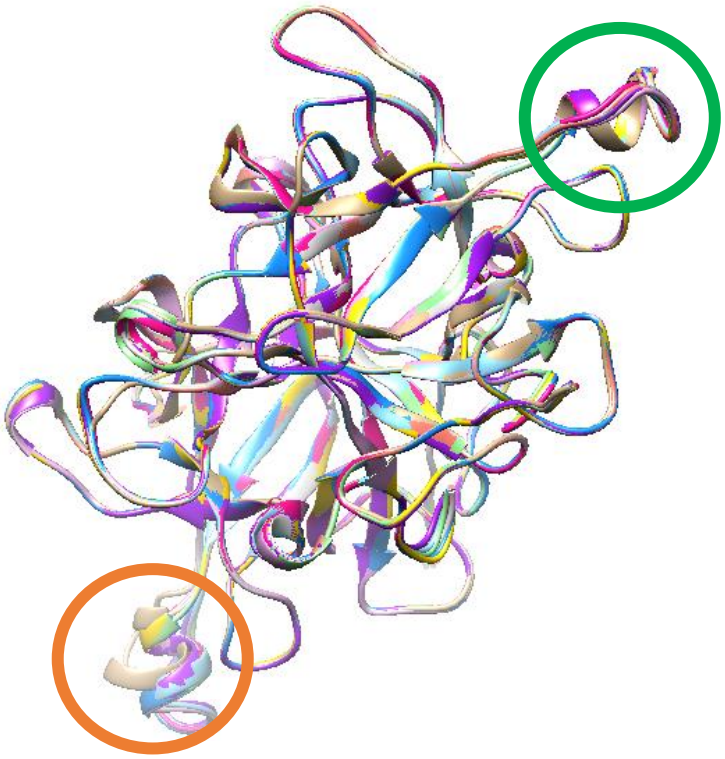
| **Figura 5.19. Modelos finales** | |
|---|---|
| M271 - 240 Frames a diferentes resoluciones | |
|  |  <br>Se observa que en la región correspondiente al *Loop* 9 (cadena A en un círculo naranja y cadena B en un círculo verde), es la que más diverge, si bien, el cambio es menor y no altera el sobrelape de todos los modelos finales presentados en esta figura. En términos de los valores apareados de RMSD, hay un promedio de 0.132, con valores máximos y mínimos entre 0.146 y 0.124. |
| M271 - 120 Frames a diferentes resoluciones | |






Se observa que en la región correspondiente al *Loop* 9 (cadena A en un círculo naranja y cadena B en un círculo verde), es la que más diverge, si bien, el cambio es menor y no altera el sobrelape de todos los modelos finales presentados en esta figura. En términos de los valores apareados de RMSD, hay un promedio de 0.132, con valores máximos y mínimos entre 0.144 y 00117.

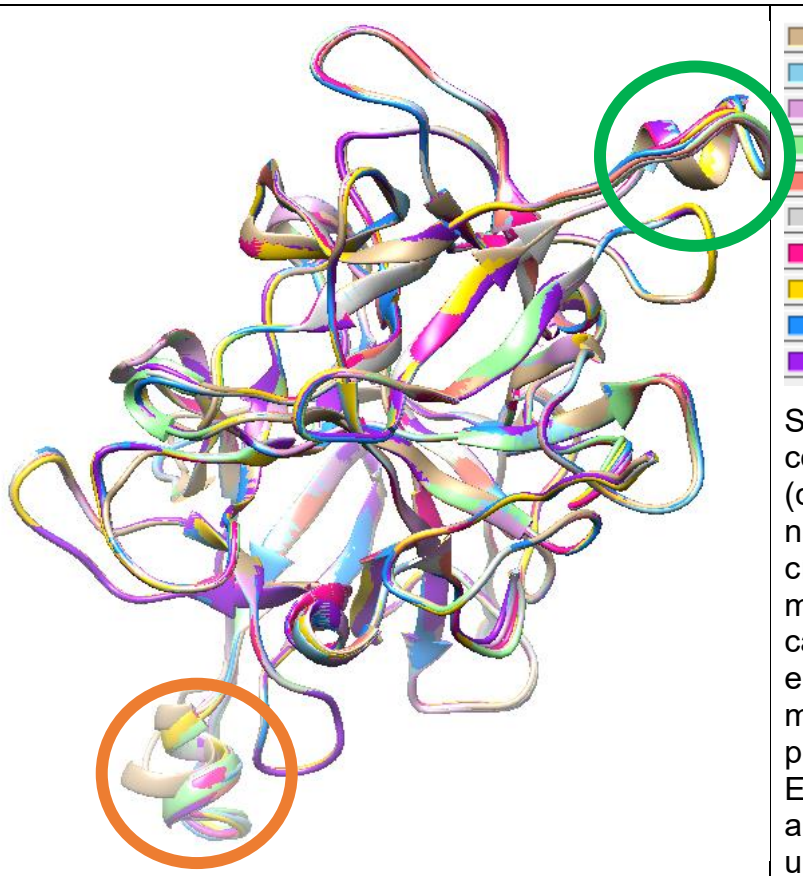
M271 - 60 Frames a diferentes resoluciones



m271_modelo_final_
100i_M271_60F_1_60A_
100i_M271_60F_1_65A_
100i_M271_60F_1_70A_
100i_M271_60F_1_75A_
100i_M271_60F_1_80A_
100i_M271_60F_1_85A_
100i_M271_60F_1_90A_
100i_M271_60F_1_95A_
100i_M271_60F_2_00A_

Se observa que en la región correspondiente al *Loop* 9 (cadena A en un círculo naranja y cadena B en un círculo verde), es la que más diverge, si bien, el cambio es menor y no altera el sobrelape de todos los modelos finales presentados en esta figura. En términos de los valores apareados de RMSD, hay un promedio de 0.194, con valores máximos y mínimos entre 0.222 y 0.176.

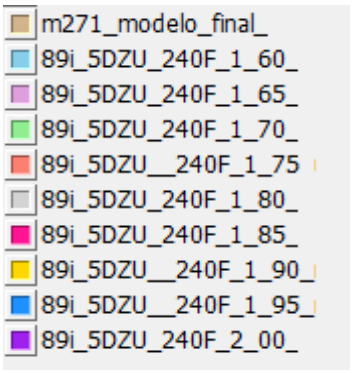
M271 – 30 Frames a diferentes resoluciones



m271_modelo_final_
100i_M271_30F_1_60A_
100i_M271_30F_1_65A_
100i_M271_30F_1_70A_
100i_M271_30F_1_75A_
100i_M271_30F_1_80A_
100i_M271_30F_1_85A_
100i_M271_30F_1_90A_
100i_M271_30F_1_95A_
100i_M271_30F_2_00A_

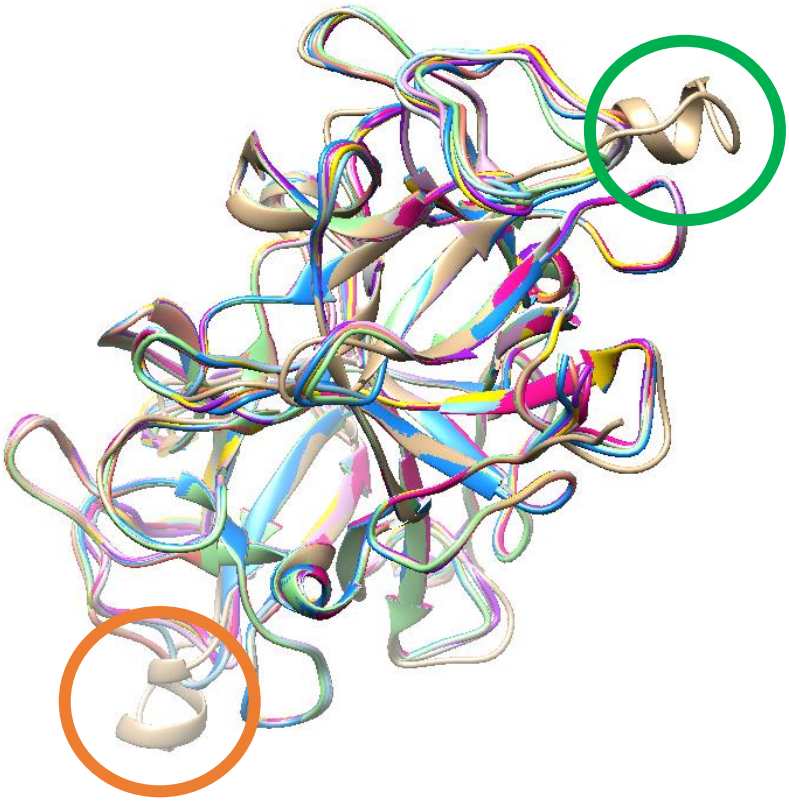
Se observa que en la región correspondiente al *Loop* 9 (cadena A en un círculo naranja y cadena B en un círculo verde), es la que más diverge, si bien, el cambio es menor y no altera el sobrelape de todos los modelos finales presentados en esta figura. En términos de los valores apareados de RMSD, hay un promedio de 0.197, con valores máximos y mínimos entre 0.223 y 0.172.

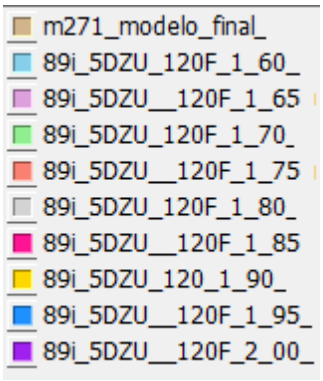
5DZU - 240 Frames a diferentes resoluciones




Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 5DZU fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.695, con valores máximos y mínimos entre 0.774 y 0.652 y considerando 176 y 175 número de carbono alfa respectivamente.

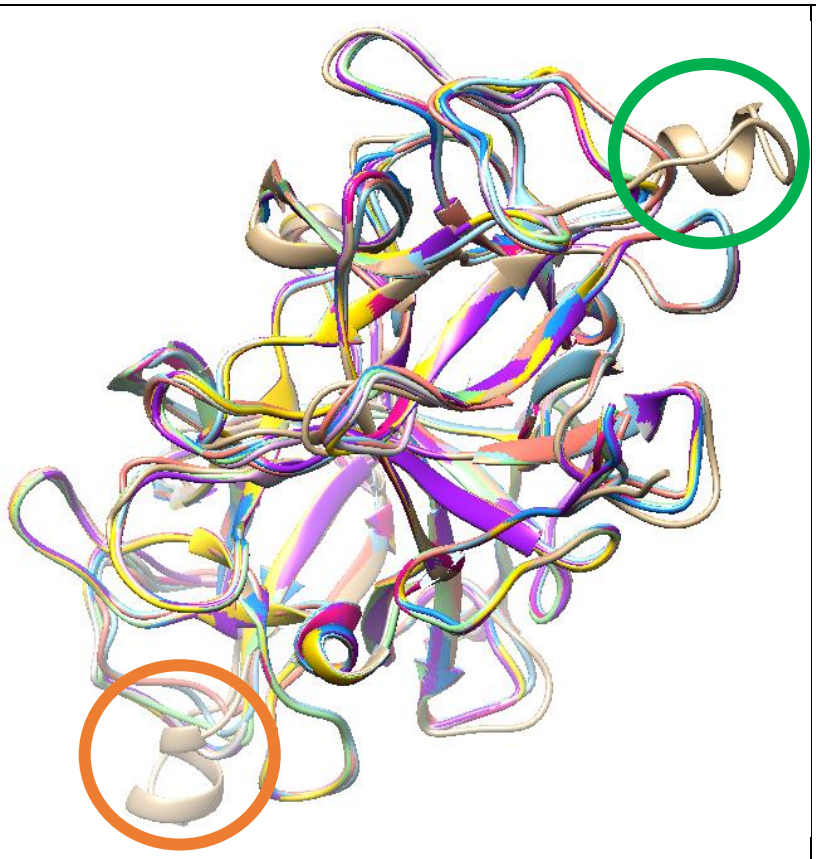
5DZU - 120 Frames a diferentes resoluciones






Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 5DZU fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.686, con valores máximos y mínimos entre 0.723 y 0.6589 y considerando 175 y 175 número de carbono alfa respectivamente.

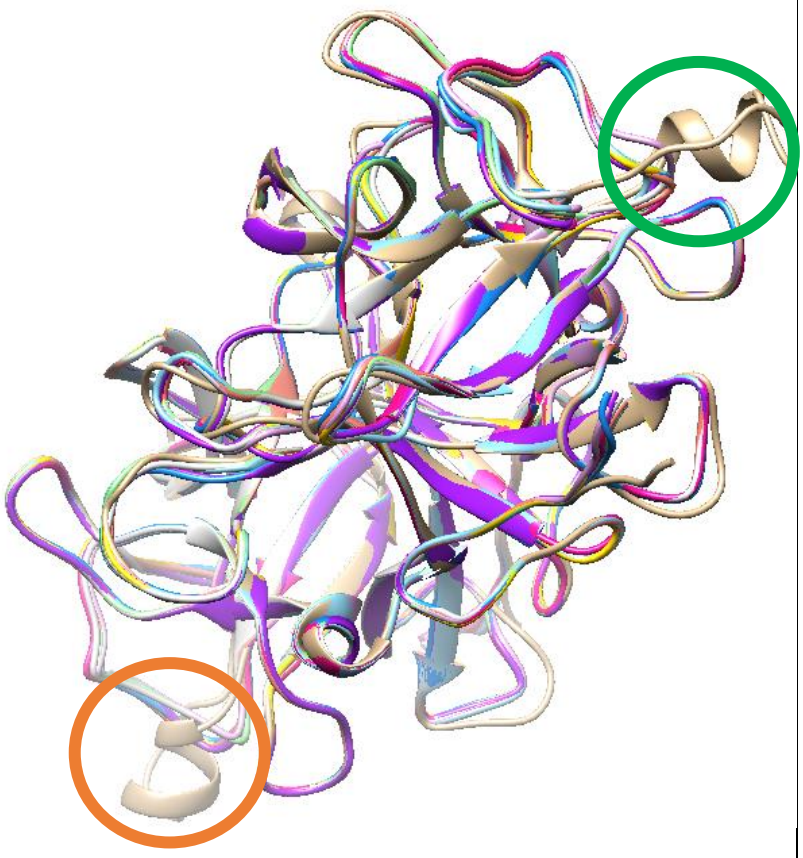
5DZU - 60 Frames a diferentes resoluciones

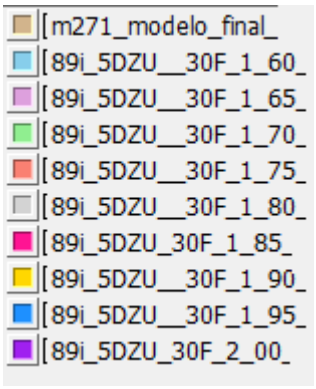


[m271_modelo_final_dos_cadenas.
[89i_5DZU__60F_1_60
[89i_5DZU_60F_1_65_
[89i_5DZU__60F_1_70_
[89i_5DZU__60F_1_75_
[89i_5DZU__60F_1_80_
[89i_5DZU_60F_1_85_
[89i_5DZU_60F_1_90_
[89i_5DZU__60F_1_95_
[89i_5DZU_60F_2_00_

Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 5DZU fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.712, con valores máximos y mínimos entre 0.758 y 0.672 y considerando 176 y 175 número de carbono alfa respectivamente.

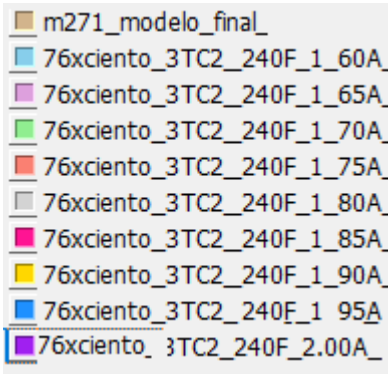
5DZU - 30 Frames a diferentes resoluciones





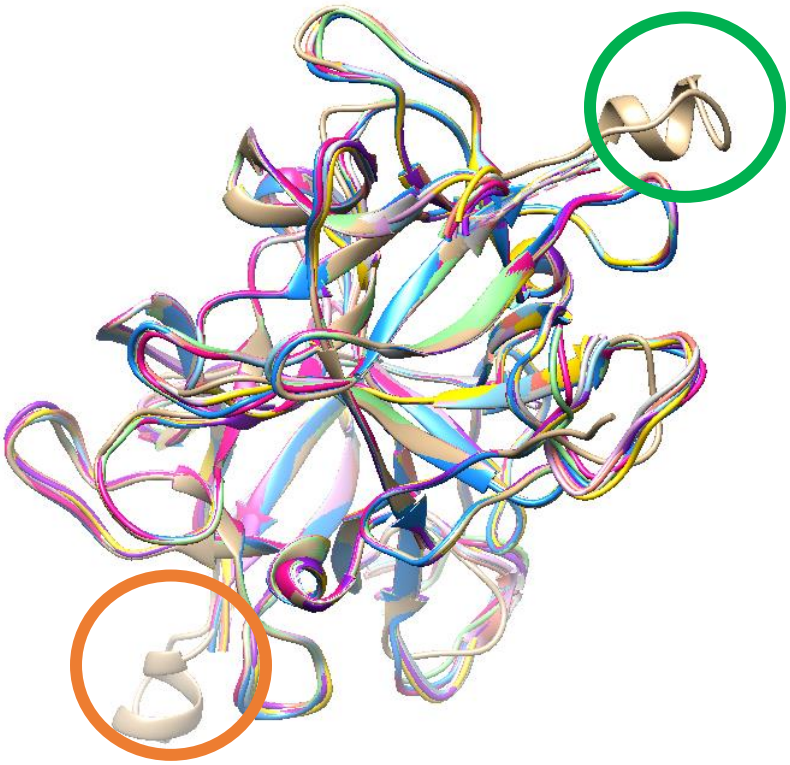

Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 5DZU fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.752, con valores máximos y mínimos entre 0.802 y 0.714 y considerando 176 y 175 número de carbono alfa respectivamente.

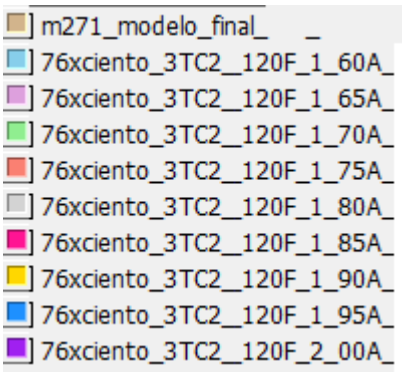
3TC2 - 240 Frames a diferentes resoluciones




Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 3TC2 fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.597, con valores máximos y mínimos entre 0.640 y 0.562 y considerando 173 y 173 número de carbono alfa respectivamente.

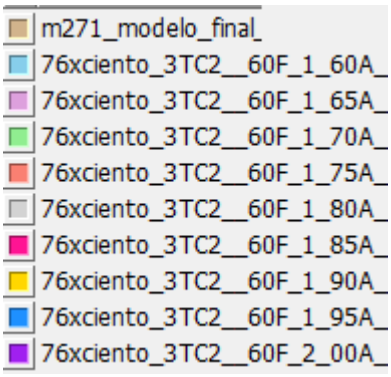
3TC2 - 120 Frames a diferentes resoluciones





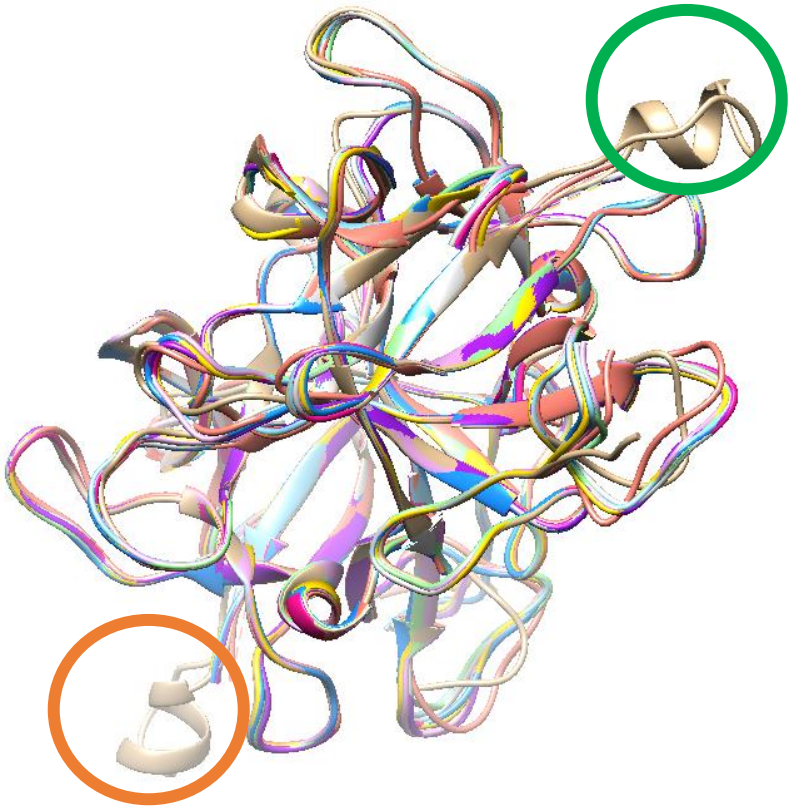

Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 3TC2 fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.605, con valores máximos y mínimos entre 0.671 y 0.564 y considerando 173 y 173 número de carbono alfa respectivamente.

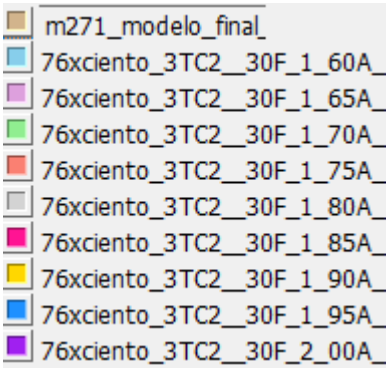
3TC2 - 60 *Frames* a diferentes resoluciones




Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 3TC2 fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.611, con valores máximos y mínimos entre 0.652 y 0.578 y considerando 173 y 172 número de carbono alfa respectivamente.

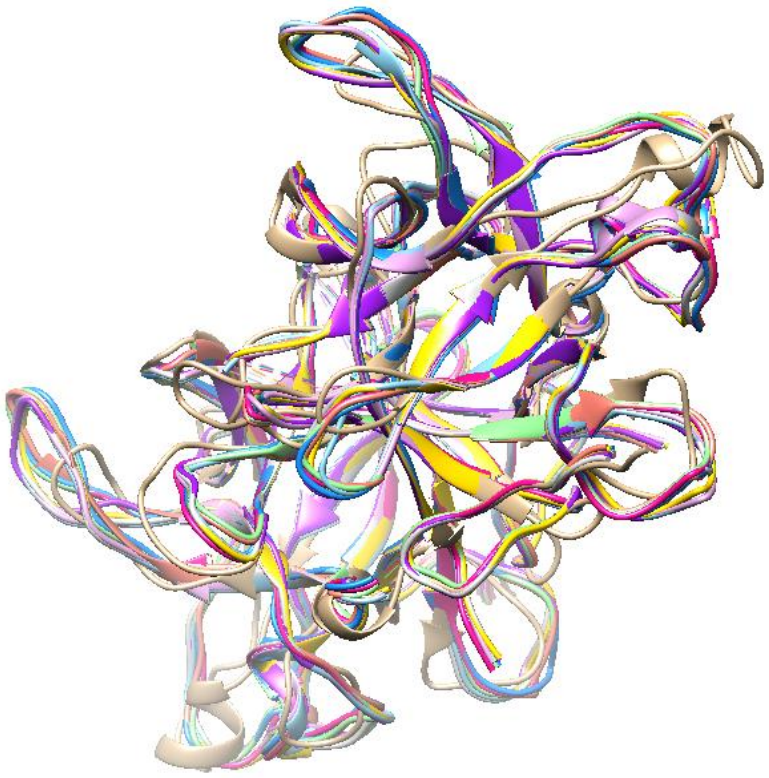
3TC2 - 30 Frames a diferentes resoluciones






Se observa que en la región correspondiente al *Loop* 9 de M271 (cadena A en un círculo naranja y cadena B en un círculo verde), ninguno de los modelos basados en 3TC2 fueron capaces de construirlo. En términos de los valores apareados de RMSD, hay un promedio de 0.656, con valores máximos y mínimos entre 0.696 y 0.606 y considerando 173 y 172 número de carbono alfa respectivamente.

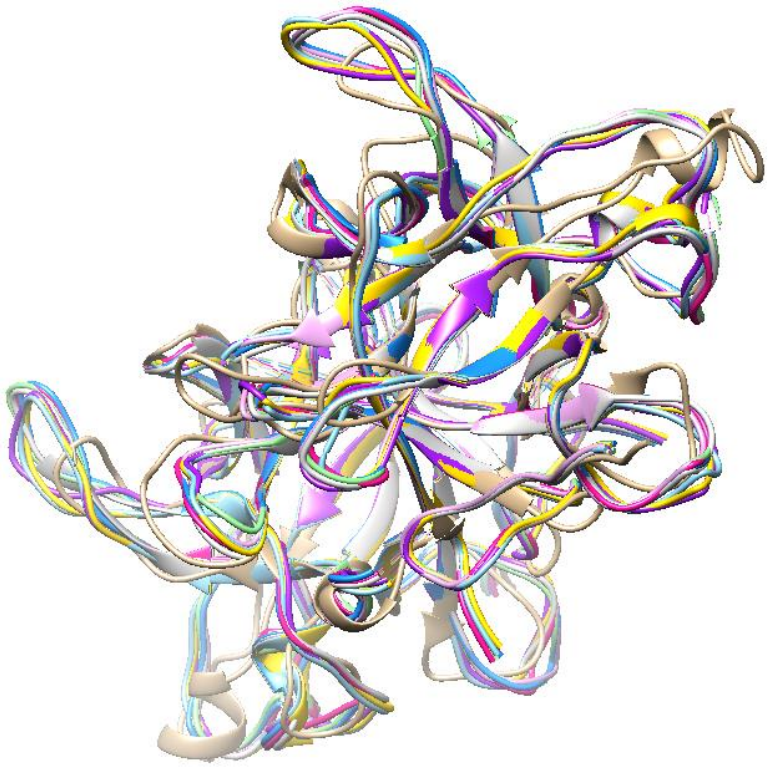
5DVH - 240 Frames a diferentes resoluciones



m271_modelo_final_
35xciento_5DVH__240F_1_60A_
35xciento_5DVH__240F_1_65A_
35xciento_5DVH__240F_1_70A_
35xciento_5DVH__240F_1_75A_
35xciento_5DVH__240F_1_80A_
35xciento_5DVH__240F_1_85A_
35xciento_5DVH__240F_1_90A_
35xciento_5DVH__240F_1_95A_
35xciento_5DVH__240F__2.00A_

En términos de los valores apareados de RMSD, hay un promedio de 1.514, con valores máximos y mínimos entre 1.600 y 1.330 y considerando 159 y 151 número de carbono alfa respectivamente.

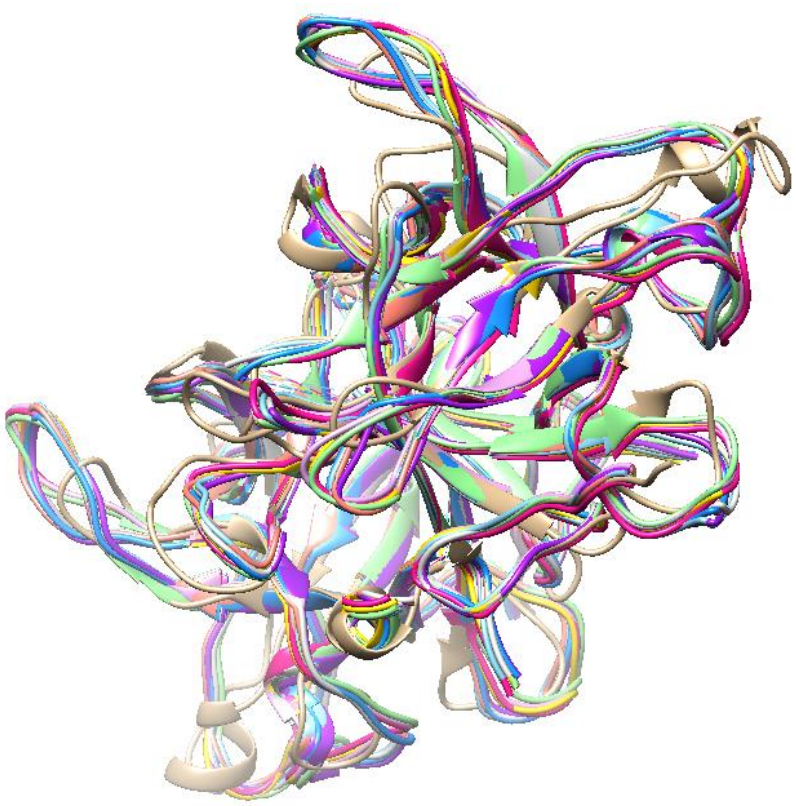
5DVH - 120 Frames a diferentes resoluciones



|m271_modelo_final_
|35xciento_5DVH_120F_1_60A_
|35xciento_5DVH_120F_1_65A_
|35xciento_5DVH_120F_1_70A_
|35xciento_5DVH_120F_1_75A_
|35xciento_5DVH_120F_1_80A_
|35xciento_5DVH_120F_1_85A_
|35xciento_5DVH_120F_1_90A_
|35xciento_5DVH_120F_1_95A_
|35xciento_5DVH_120F_2_00A_

En términos de los valores apareados de RMSD, hay un promedio de 1.456, con valores máximos y mínimos entre 1.529 y 1.382 y considerando 156 y 153 número de carbono alfa respectivamente.

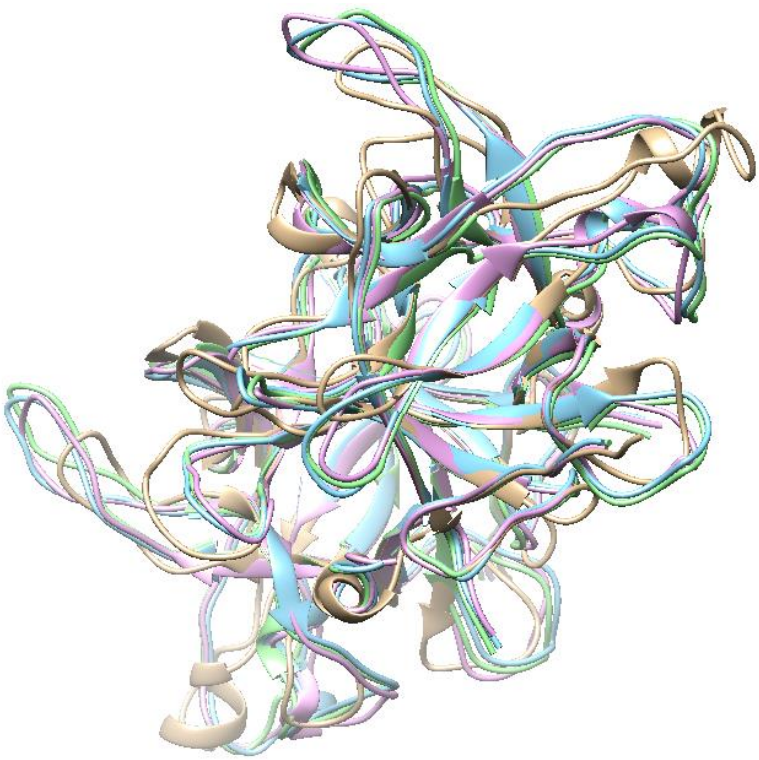
5DVH - 60 Frames a diferentes resoluciones



m271_modelo_final_
35xciento_5DVH_sinLig_60F_1_60_
35xciento_5DVH_sinLig_60F_1_65_
35xciento_5DVH_sinLig_60F_1_70_
35xciento_5DVH_sinLig_60F_1_75_
35xciento_5DVH_sinLig_60F_1_80_
35xciento_5DVH_sinLig_60F_1_85_
35xciento_5DVH_sinLig_60F_1_90_
35xciento_5DVH_sinLig_60F_1_95_

En términos de los valores apareados de RMSD, hay un promedio de 1.516, con valores máximos y mínimos entre 1.801 y 1.384 y considerando 156 y 152 número de carbono alfa respectivamente.

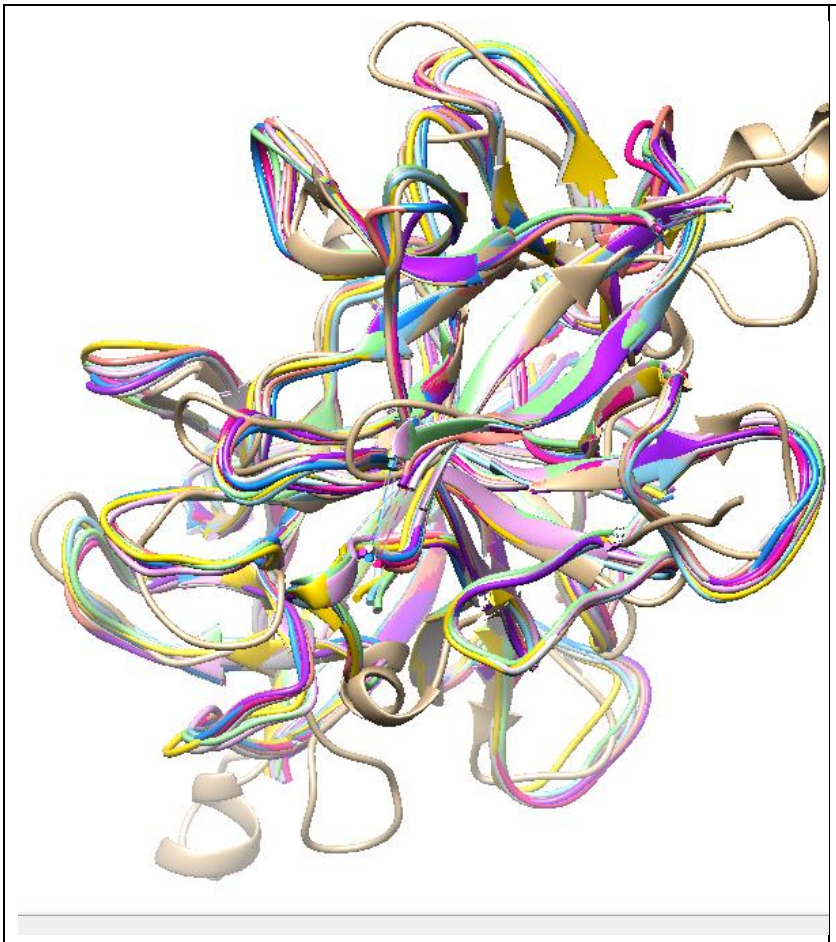
5DVH - 30 Frames a diferentes resoluciones

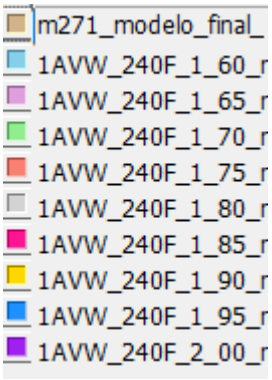


m271_modelo_final_
35xciento_5DVH__30F_1_85_
35xciento_5DVH__30F_1_90_
35xciento_5DVH__30F_1_95_
35xciento_5DVH__30F_2_00_

En términos de los valores apareados de RMSD, hay un promedio de 1.585, con valores máximos y mínimos entre 1.677 y 1.552 y considerando 157 y 154 número de carbono alfa respectivamente.

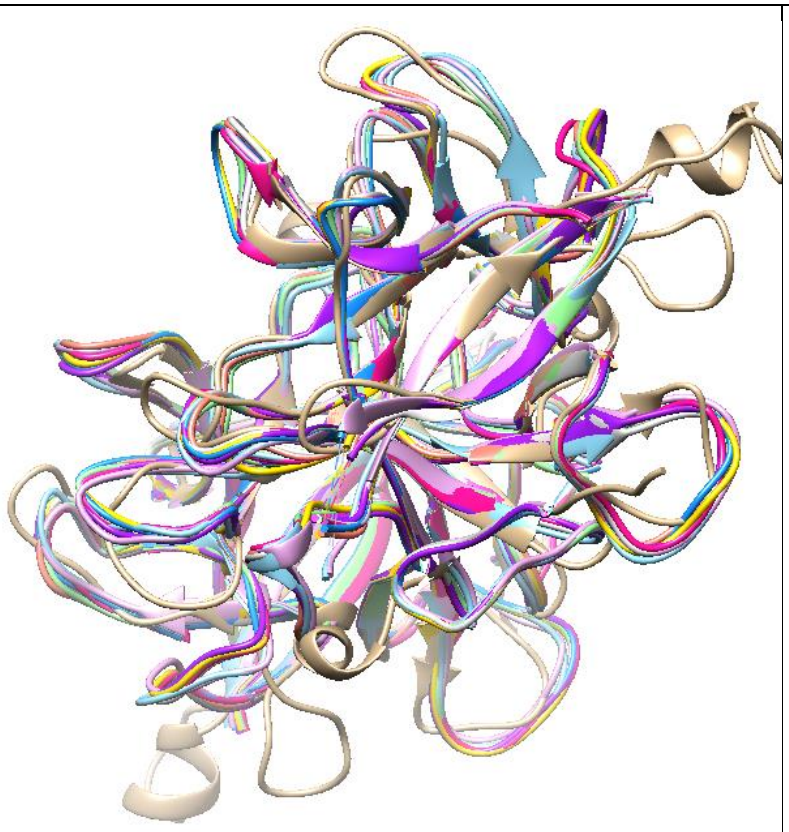
1AVW - 240 Frames a diferentes resoluciones






En términos de los valores apareados de RMSD, hay un promedio de 1.260, con valores máximos y mínimos entre 1.453 y 1.172 y considerando 148 y 147 número de carbono alfa respectivamente.

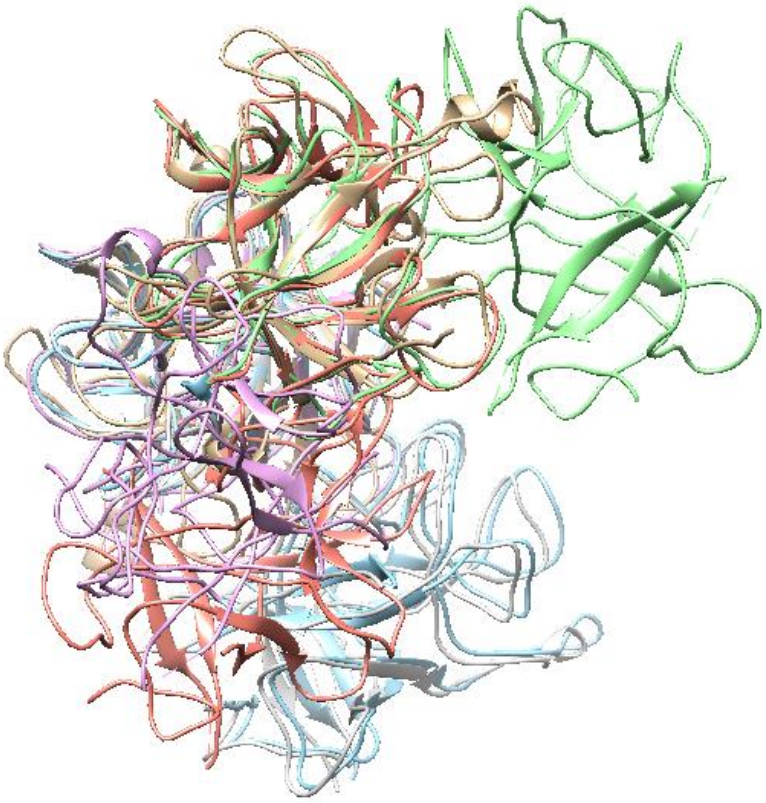
1AVW - 120 Frames a diferentes resoluciones



m271_modelo_final_
1AVW_120F_1_60_
1AVW_120F_1_65_
1AVW_120F_1_70_
1AVW_120F_1_75_
1AVW_120F_1_80_
1AVW_120F_1_85_
1AVW_120F_1_90_
1AVW_120F_1_95_
1AVW_120F_2_00_

En términos de los valores apareados de RMSD, hay un promedio de 1.278, con valores máximos y mínimos entre 1.340 y 1.188 y considerando 146 y 147 número de carbono alfa respectivamente.

1AVW - 60 Frames a diferentes resoluciones



m271_modelo_final_
1AVW_60F_1_80_
1AVW_60F_1_85_
1AVW_60F_1_90_
1AVW_60F_1_95_
1AVW_60F_2_00_

En términos de los valores apareados de RMSD, hay un promedio de 2.419, con valores máximos y mínimos entre 4.695 y 1.707 y considerando 38 y 150 número de carbono alfa respectivamente.

En estos casos no hay respuesta que permita reducir el valor de los R de afinamiento.

9. BIBLIOGRAFÍA


1. Campuzano, A. O. Estudio estructural del asa L9 en la proteína M271 de Solanum tuberosum, un inhibidor de proteasas de la familia Kunitz-STI. (U.N.A.M., 2019).

2. Gromiha, M. M. & Gromiha, M. M. Chapter 1 proteins. in *Protein Bioinformatics* 1–27 (Academic Press, 2010). doi:10.1016/B978-8-1312-2297-3.50001-1

3. Meyder, A. *et al.* Structural Bioinformatics StructureProfiler : An all-in-one Tool for 3D Protein Structure Profiling. 2017–2018 (2018). doi:10.1093/bioinformatics/bty692/5075170

4. Xu, Y. Computational Methods for Protein Structure Prediction. (2015). Available at: https://slideplayer.com/slide/4889164/.

5. Wei, X., Li, Z., Li, S., Peng, X. & Zhao, Q. Protein structure determination using Riemannian approach. *bioRxiv Bioinforma.* (2019). doi:10.1101/599761

6. Biostructure, C. Comparison of Crystallography, NMR and EM. (2017). Available at: https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm.

7. Berman, H. M. & Westbrook, John, Gary Gilliland, T.N.Bhat, Helge Weissig, Ilya N. Shindyalov, P. E. B. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

8. Kendrew, J. C., Dintzis, G. B. H. M., Parrish, R.G. & Wyckoff, H. A three-dimensional of the myoglobin molecule obtained by X-Ray analysis. *Nature* **181**, (1958).

9. Campos-Acevedo, A. A., Díaz-Vilchis, A., Sotelo-Mundo, R. R. & Rudiño-Piñera, E. First attempts to crystallize a non-homogeneous sample of thioredoxin from Litopenaeus vannamei: What to do when you have diffraction data of a protein that is not the target? *Biochem. Biophys. Reports* **8**, 284–289 (2016).

10. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr. Sect. FStructural Biol. Commun.* **70**, 2–20 (2014).

11. Krauss, I. R., Merlino, A., Vergara, A. & Sica, F. An overview of biological macromolecule crystallization. *Int. J. Mol. Sci.* **14**, 11643–11691 (2013).

12. Dessau, M. A. & Modis, Y. Protein crystallization for X-ray crystallography. *J. Vis. Exp.* **9**, 1–6 (2011).

13. Dauter, Z. *Collection of X-ray diffraction data from macromolecular crystals*. *Methods Mol Biol.* **1607**, (2017).

14. Powell, H. R. X-ray data processing. *Biosci. Rep.* **37**, (2017).

15. Evans, P. R. An introduction to data reduction: Space-group determination,

scaling and intensity statistics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 282–292 (2011).

16. Elspeth Garman, Robin L. Owen, S. D. *Macromolecular Crystallography Protocols: Volume 2: Structure Determination*. **2**, (2007).

17. Sheldrick, G. M. *et al.* Part 16 . Direct Methods. **F**, 413–432 (2012).

18. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

19. Rudiño-Piñera, E. Estudios Estructurales sobre el mecanismo de activación alostérica de la glucosamina-6-fosfato desaminasa de Escherichia coli. (Univerisdad Nacional Autónoma de México, 2001).

20. Weiss, M. S. & Hilgenfeld, R. On the use of the merging R factor as a quality indicator for X-ray data. *J. Appl. Crystallogr.* **30**, 203–205 (1997).

21. Andrea, T. *Selected data quality indicators - Compiled on special request from the CSHL class of 2014*. (2014).

22. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **34**, 60–68 (2015).

23. Weiss, M. S. Global indicators of X-ray data quality. *Appl. Crystallogr.* **34**, 130–135 (2001).

24. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1204–1214 (2013).

25. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.* **280**, 5705–5736 (2013).

26. Rhodes, G. Judging the Quality of Macromolecular Models - A Glossary of Terms from Crystallography, NMR, and Homology Modeling. Available at: https://spdbv.vital-it.ch/TheMolecularLevel/ModQual/#Completeness.

27. Kleywegt, G. J. 21 .1 Validation of protein crystal structures. *Int. Tables Crystallogr.* **F**, 497–506 (2006).

28. Phaserwiki. FAQ. (2016). Available at: https://www.phaser.cimr.cam.ac.uk/index.php/FAQ#:~:text=LLG stands for Log Likelihood Gain.&text=The LLG is the difference,distribution of the same atoms.

29. APTECH. Beginner's Guide To Maximum Likelihood Estimation. (2020). Available at: https://www.aptech.com/blog/beginners-guide-to-maximum-likelihood-estimation-in-gauss/.

30. Rycroft, P. Structural studies of the Dickeya dadantii type II secretion system protein GspB. (Queen Mary University of London, 2015).

31. Tolstikova, A. Development of diffraction analysis methods for serial

crystallography Dissertation. (Universitat Hamburg, 2020).

32. Goswami, N., Li, J. & Xie, J. Protected Metal Clusters: From Fundamentals to Applications. in *Frontiers of Nanoscience* **9**, 297–345 (2015).

33. Felcy, F., Korostele, A. & Chapman, M. S. Bias in cross-validated free R factors: mitigation of the effects of non-crystallographic symmetry. *Acta Crystallogr. a Sect. D Biol. Crystallogr.* **4449**, 227–238 (2006).

34. Wang, J. Estimation of the quality of refined protein crystal structures. *Protein Sci.* **24**, 661–669 (2015).

35. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science (80-. ).* **336**, 1030–1033 (2012).

36. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1–21 (2008).

37. Sliz, P. X-ray Crystallography. 24 (2008). Available at: https://es.scribd.com/document/219142066/x-ray-Crystallography. (Accessed: 18th March 2020)

38. Jeffrey, P. Using HKL3000R for Data Collection at Princeton University. (2017). Available at: http://xray0.princeton.edu/~phil/Facility/hkl3k.html. (Accessed: 21st March 2020)

39. Juers, D. Protein Crystallography. (2015). Available at: https://www.whitman.edu/Documents/Academics/Protein_Crystallography_Part_III.pdf. (Accessed: 18th March 2020)

40. Sawaya, M. R. *et al.* Supporting Information. *PNAS* **111**, 12769–12774 (2014).

41. Phaserwiki. Molecular Replacement. (2019). Available at: https://www.phaser.cimr.cam.ac.uk/index.php/Molecular_Replacement. (Accessed: 18th March 2020)

42. Phenix. Frequently asked questions about molecular replacement. Available at: https://www.phenix-online.org/documentation/faqs/mr.html.

43. McCoy, A. Molecular Replacement. *University of Cambridg* (2014). Available at: https://www.ccp4.ac.uk/schools/DLS-2014/course_material/day07/Airlie_McCoy_Phaser_MR.pdf. (Accessed: 22nd March 2020)

44. Lebedev, A. Molecular Replacement. (2014). Available at: https://slideplayer.com/slide/7347285/. (Accessed: 22nd March 2020)

45. OnlineDictionaryOfCrystallography. R factor. (2017). Available at: https://dictionary.iucr.org/R_factor. (Accessed: 22nd March 2020)

46. Liu, Z. J. *et al.* A multi-dataset data-collection strategy produces better diffraction data. *Acta Crystallogr. Sect. A Found. Crystallogr.* **67**, 544–549 (2011).

47. Walker, J. M. *Nucleic Acid Crystallography*. (2016).

48. McPherson, A. & Larson, S. B. A guide to the crystallographic analysis of icosahedral viruses. *Crystallogr. Rev.* **21**, 3–56 (2015).

49. Diederichs, K. CC* - Linking crystallographic model and data quality. (2015). Available at: http://www.xray.cz/setkani/abst2015/diederichs.htm.

50. Wang, J., Brudvig, G. W., Batista, V. S. & Moore, P. B. On the relationship between cumulative correlation coefficients and the quality of crystallographic data sets. *Protein Sci.* **26**, 2410–2416 (2017).

51. Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 454–467 (2012).

52. Evans, P. & McCoy, A. An introduction to molecular replacement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1–10 (2007).

53. Banaszak, L. J. *Foundations of Structural Biology*. *Foundations of Structural Biology* (2000).

54. Song, H. K. & Suh, S. W. Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator 1 1Edited by R. Huber. *J. Mol. Biol.* **275**, 347–363 (1998).

55. PDBe. 5DVH. Available at: https://www.ebi.ac.uk/pdbe/entry/pdb/5dvh. (Accessed: 22nd July 2021)

56. PDBe. 3TC2. (2021). Available at: https://www.ebi.ac.uk/pdbe/entry/pdb/3tc2. (Accessed: 22nd July 2021)

57. PDBe. 5DVH. (2021). Available at: https://www.ebi.ac.uk/pdbe/entry/pdb/5dzu. (Accessed: 22nd July 2021)

58. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385–395 (2004).

59. Leslie, A. G. W. & Powell, H. R. Processing diffraction data with mosflm. *Evol. Methods Macromol. Crystallogr.* 41–51 (2007). doi:10.1007/978-1-4020-6316-9_4

60. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. - Sect. D Biol. Crystallogr.* **62**, 72–82 (2006).

61. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 355–367 (2011).

62. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).

63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of

Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).

64. Afonine, P. V, Headd, J. J., Terwilliger, T. C. & Adams, P. D. PHENIX News - Real space refine. *Comput. Crystallogr. Newsl.* **4**, 43–44 (2013).

65. RStudio Team. RStudio: Integrated Development for R. RStudio. (2020).

66. Matsuda, T. *Future Directions in Biocatalysis*. (2017).

67. Guerra, Y., Valiente, P. A., Pons, T., Berry, C. & Rudiño-Piñera, E. Structures of a bi-functional Kunitz-type STI family inhibitor of serine and aspartic proteases: Could the aspartic protease inhibition have evolved from a canonical serine protease-binding loop? *J. Struct. Biol.* (2016). doi:10.1016/j.jsb.2016.06.014

68. Guo, J., Erskine, P. T., Coker, A. R., Wood, S. P. & Cooper, J. B. Structure of a Kunitz-type potato cathepsin D inhibitor. *J. Struct. Biol.* (2015). doi:10.1016/j.jsb.2015.10.020