UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Instituto de Biotecnología

Evaluación comparativa de modelos de lenguaje de proteínas para clasificar betalactamasas y predecir su actividad catalítica

TESIS
QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA
Miguel Ángel González Arias

TUTOR PRINCIPAL
Dr. Lorenzo Patrick Segovia Forcella
Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITÉ TUTOR
Dr. José Arcadio Farías Rico
Centro de Ciencias Genómicas, UNAM
Dr. Francisco Xavier Soberón Mainero
Instituto de Biotecnología, UNAM

Cuernavaca, Morelos. Febrero, 2024

Este trabajo se realizó en el Laboratorio 12, perteneciente al Departamento de Ingeniería Celular y Biocatálisis del Instituto de Biotecnología de la UNAM, bajo la asesoría de Alejandro Garciarubio y Lorenzo Segovia.

La realización de este proyecto estuvo financiada por parte de la Coordinación General de Estudios de Posgrado de la UNAM en el marco del proyecto "Beca Extraordinaria".

"Miro el registro geológico como una historia del mundo imperfectamente […] escrita en dialectos cambiantes;

de esta historia poseemos solamente el último volumen,

que sólo refiere a dos o tres países.

De este volumen,

solamente aquí y allá se ha conservado algún breve capítulo,

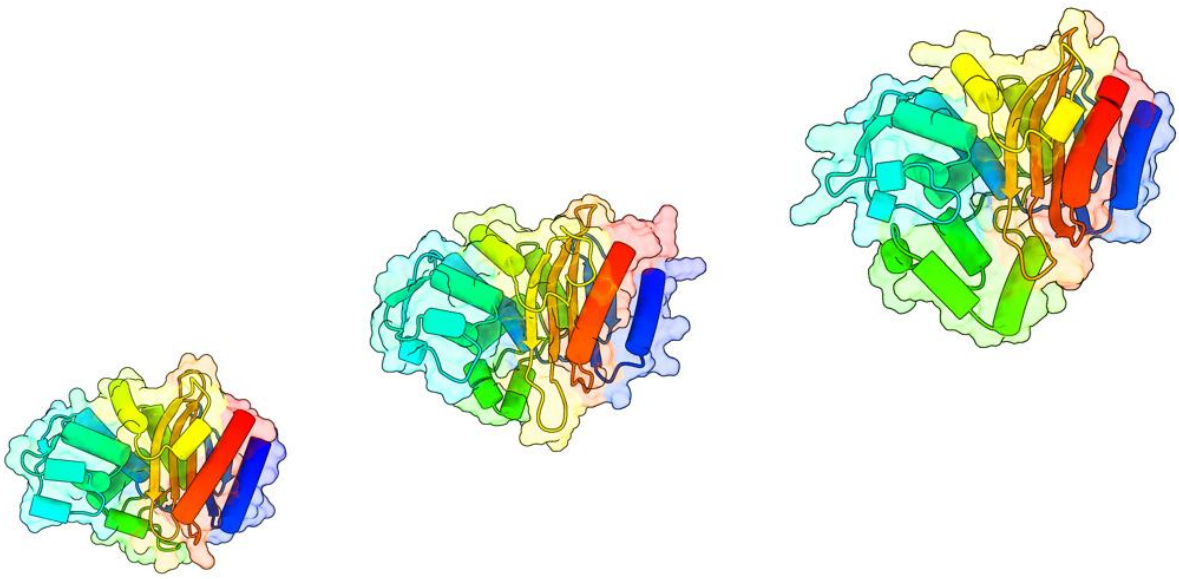
y de cada página,

solamente aquí y allá algunas líneas".

Charles Darwin

El origen de las especies

Capítulo: De la imperfección de los datos geológicos



Agradecimientos

Académicos

A mi familia y amigos por apoyarme en todo lo que hago.

A Alejandro Garciarrubio por su detallada asesoría teórica y técnica en todo el proyecto.

A Alejandro Garciarrubio y Lorenzo Segovia por su asesoría y paciencia.

A José Arcadio por las oportunidades ofrecidas para mi crecimiento personal y académico.

A los miembros de mi comité tutoral y comité de revisión por la retroalimentación a mi trabajo.

A Blanca Ramos por su apoyo técnico en el desarrollo de este proyecto.

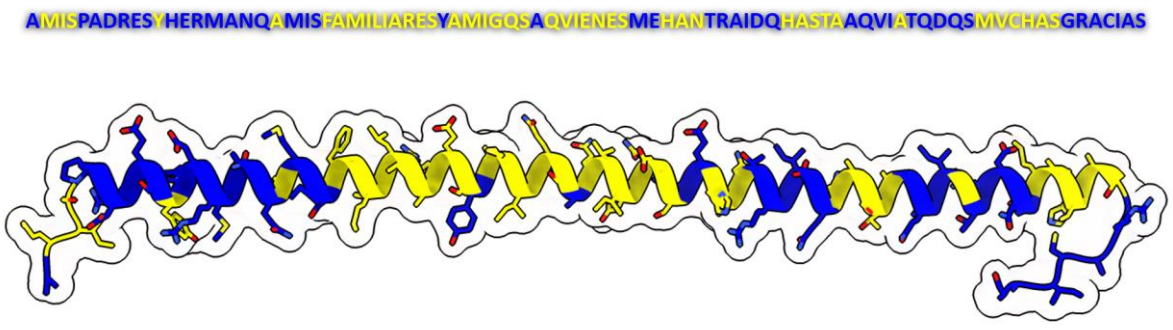
A la UNAM y a los mexicanos por haberme becado durante mi maestría.

A Walter Santos, Diana Montes, Xaira Rivera y Alejandro Alarcón por su asesoría técnica y revisiones.

A las personas detrás de los canales de YouTube ProgramaciónATS, DotCSV, StatQuest, AprendeIA, DataProfessor, Kimberly Fessel y todos los divulgadores de IA y ciencia de datos.

A Aldo Pérez, Angélica Domínguez, Armando Ávila, Laura Martínez, Rafael Sánchez, Salvador Guillen, los miembros del grupo “Tardes de bioinformática” y científicos en Twitter por su retroalimentación.

Personales



Dedicatoria

A Katy … mi muy mejor amiga 

Índice

Resumen

Las betalactamasas son enzimas que degradan antibióticos betalactámicos, los cuales se encuentran entre los más usados a nivel mundial. Existen dos esquemas de clasificación de betalactamasas: la clasificación molecular las divide en cuatro clases principales basadas en sus mecanismos catalíticos y homología estructural, y la clasificación funcional, que las agrupa en 16 categorías según su capacidad para degradar antibióticos. Pese a su utilidad, ambos esquemas han sido objeto de debate, haciéndolos interesantes a investigar mediante enfoques basados en datos. Recientemente, los modelos de lenguaje, un tipo de redes neuronales profundas, han mostrado un excelente rendimiento en el modelado de datos complejos, como la estructura y función de las proteínas. Durante su entrenamiento, estos modelos aprenden a codificar proteínas en vectores numéricos llamados *embeddings*, sin embargo, aún no es claro qué tipo de información biológica contienen.

En este trabajo evaluamos las dos clasificaciones de betalactamasas usando modelos de lenguaje de proteínas. Para ello, tomamos secuencias de bases de datos curadas de betalactamasas y obtuvimos sus *embeddings* utilizando siete modelos. Con los *embeddings*, comparamos la capacidad de los modelos para identificar los grupos de la clasificación molecular usando tres algoritmos de reducción de dimensionalidad, dos métricas de distancia y *kmeans*. Además, analizamos la función catalítica de las betalactamasas en relación con las estructuras de sus sustratos y entrenamos 21 regresores de distintos antibióticos betalactámicos para predecir el nivel de resistencia de una betalactamasa.

Los resultados sugieren que los modelos de lenguaje de proteínas como ESM-1b, son capaces de identificar las clases, subclases, familias y subfamilias de betalactamasas, así como otras propiedades como el número de residuos y fracción de estructura secundaria, haciéndolos herramientas útiles que ofrecen una nueva perspectiva en la detección de grupos. Particularmente, los *embeddings* de ESM-1b nos permitieron identificar secuencias representativas dos posibles nuevas subclases de betalactamasas: A3 y C2. Al analizar la estructura y función de los antibióticos sugerimos que las penicilinas pueden ser un sustrato más fácil de degradar respecto al resto de antibióticos betalactámicos debido a su considerable similitud estructural. Sin embargo y dado la escasez de datos, solo logramos obtener un regresor para Cefoxitina capaz de generar predicciones consistentes con las actividades catalíticas registradas por experimentos *in vitro*.

Hasta dónde sabemos, este es el primer trabajo en usar modelos de lenguaje de proteínas para modelar las dos clasificaciones de betalactamasas, y esperamos que sirva como referencia para trabajos futuros. Además, contribuye a la comprensión actual del tipo de información que detectan estos modelos y esperamos que las estrategias aquí exploradas puedan ser aplicadas a otros plegamientos proteicos en la detección de grupos.

Abstract

Betalactamases are enzymes that degrade betalactam antibiotics, which are among the most widely used worldwide. There are two classification schemes for betalactamases: the molecular classification divides them into four main classes based on their catalytic mechanisms and structural homology, and the functional classification groups them into 16 categories based on their ability to degrade antibiotics. Despite their utility, both schemes have been the subject of debate, making them interesting to investigate through data-driven approaches. Recently, language models, a type of deep neural network, have shown excellent performance in modeling complex data, such as the structure and function of proteins. During training, these models learn to encode proteins into numerical vectors called embeddings, however, it is still unclear what kind of biological information they contain.

In this work, we evaluate the two classifications of betalactamases using protein language models. We took sequences from curated betalactamase databases and obtained their embeddings using seven models. With the embeddings, we compared the models' ability to identify the groups of the molecular classification using three dimensionality reduction algorithms, two distance metrics, and kmeans. Additionally, we analyzed the catalytic function of betalactamases in relation to the structures of their substrates and trained 21 regressors for different betalactam antibiotics to predict the resistance level of a beta-lactamase.

The results suggest that protein language models like ESM-1b can identify the classes, subclasses, families, and subfamilies of betalactamases, as well as other properties such as the number of residues and secondary structure fraction, making them useful tools that offer a new perspective to detect groups. Particularly, the embeddings of ESM-1b allowed us to identify representative sequences for two possible new subclasses of betalactamases: A3 and C2. When analyzing the structure and function of antibiotics, we suggest that penicillins may be an easier substrate to degrade compared to other betalactam antibiotics due to their considerable structural similarity. However, due to the scarcity of data, we only trained a regressor for Cefoxitin capable of generating predictions consistent with catalytic activities archived by in vitro experiments.

To the best of our knowledge, this is the first work that has used protein language models to model both classifications of betalactamases, and we hope it serves as a reference for future work. Additionally, it contributes to the current understanding of the type of information these models detect, and we hope the strategies explored here can be applied to other protein folds in group detection analysis.

Resumen ejecutivo

Este resumen ejecutivo ofrece una visión general del análisis, datos utilizados y resultados clave de cada parte, pero, sin profundizar en los detalles específicos del texto principal.

Parte 1.1.

Evalúe la capacidad de siete modelos de lenguaje de proteínas para detectar las clases de betalactamasas de acuerdo con la clasificación molecular. El método de clasificación consistió en obtener por cada modelo, los *embeddings* de cada una de las proteínas, proyectarlos a dos dimensiones, y comparar visualmente su agrupación. Además, comparé estos resultados contra otros métodos no basados en reducción de la dimensionalidad.

- Se analizaron 25,809 secuencias de betalactamasas manualmente curadas, dentro de las que se consideran la mayoría de las familias enzimáticas reconocidas en cada clase, secuencias a las que no se les ha asignado una familia y secuencias de referencia (cuatro ancestros y ocho consensos).
- Las serin y metalobetalactamasas se analizaron por separado al no ser homólogas entre sí.
- Los modelos de lenguaje de proteínas fueron: ESM, ESM-1b, Prot-T5-BFD, Prot-T5XL-U50, XLNet, CARP y Bepler.
- Los algoritmos de reducción de la dimensionalidad fueron: PCA, tSNE y UMAP.
- Los métodos no basados en reducción de la dimensionalidad fueron: distancia euclidiana, similitud coseno y *kmeans*.
- Las clasificaciones conocidas fueron: las clases A, C y D de serinbetalactamasas y los grupos ME (subclase B3) y MB (subclase B1 y B2) de metalobetalactamasas.

Resultado: tSNE y UMAP permiten identificar mejor a los grupos de betalactamasas respecto a PCA debido a las transformación no lineales que implementan. Usando los modelos ESM-1b y ProtT5-XL-U50 es identificar claramente a los grupos de betalactamasas de forma tal que los miembros de cada grupo están cercanos entre sí y lejos de las otras clases en sus representaciones de baja dimensión, contrario a los resultados de otros modelos como XLNet y Bepler que son incapaces de distinguir a los grupos de betalactamasas. Se eligió la representación de tSNE de ESM-1b en el resto del trabajo para realizar análisis más detallados en experimentos posteriores dado que ESM-1b logra identificar claramente a los grupos de betalactamasas y es posible conocer la fidelidad de las representaciones de tSNE.

Parte 1.2.

Para saber qué información biológica pesa más en la representación de tSNE de ESM-1b, mapeé distintas propiedades cuantitativas y cualitativas derivadas de sus secuencias. Además, comparé la organización de las representaciones de tSNE al usar solo secuencias representativas (agrupadas a un 90% de identidad de secuencia), así como añadiendo secuencias de dos posibles nuevas subclases de serin y metalobetalactamasas. Finalmente, comparé la similitud a nivel de secuencia y estructura entre un subconjunto representativo de los grupos identificados de la clase A.

- Las propiedades cuantitativas fueron: longitud, masa molecular, aromaticidad, inestabilidad, hidropatía, punto isoeléctrico, entropía y estructura secundaria (hélice, giro y beta plegada).
- Las propiedades cualitativas fueron: taxonomía (Filo, Clase, Orden, Familia, Género y Especie), 290 familias enzimáticas, secuencias ancestrales y firmas, 44 subfamilias clase D y seis grupos filogenéticos de la clase A.
- Las posibles nuevas subclases corresponden a una secuencia llamada LRA-5 (subclase A3) y 201 secuencias VarG (metalobetalactamasas).

Resultado: tSNE agrupa correctamente a las betalactamasas de acuerdo con su familia y subfamilia enzimática. Sin embargo, no parece haber una buena agrupación al considerar la taxonomía (salvo a nivel de especie) y tampoco con los seis grupos filogenéticos en el caso de la clase A. Las propiedades cuantitativas con mayor peso fueron la longitud, el masa molecular, hélice, beta plegada, aromaticidad, inestabilidad y entropía. La identificación de los grupos de betalactamasas no se ve afectada al considerar solo secuencias representativas; sin embargo, la organización si se modifica respecto a la representación con todas las secuencias. Finalmente, los resultados soportan a VarG como un grupo distinto al resto de metalobetalactamasas. Además, se detectaron secuencias representativas de una posible subclase C2 y A3. Este último grupo se distingue a nivel estructural por tres alfa hélices en la región del *loop* omega y un mayor número de residuos respecto a A1 y A2.

Parte 2.1.

Usando bases de datos curadas para betalactamasas, construí un conjunto de datos de concentraciones mínimas inhibitorias de 21 antibióticos betalactámicos que sirvieron como referencia y analicé la relación que tienen entre si al ser degradados por varias betalactamasas. Con un conjunto de 50 antibióticos betalactámicos, analicé su similitud estructural. Finalmente, para un subconjunto de 16 antibióticos, analicé la relación funcional y estructural que tienen entre sí.

- Después de una limpieza y normalización de los datos, obtuve 2,383 datos de concentraciones mínimas inhibitorias que provienen de varios artículos manualmente curados para 21 antibióticos.
- La métrica de similitud funcional fue la correlación de Spearman entre los valores de concentraciones mínimas inhibitorias. La métrica de similitud estructural entre 50 antibióticos fue la similitud de Tanimoto.

- Consideré 16 antibióticos no asociados a inhibidores para evaluar la relación entre las similitud estructural y funcional al calcular la correlación de Pearson entre los datos.

Resultado: los datos de concentraciones mínimas inhibitorias reflejan correctamente las capacidades catalíticas de las clases de betalactamasas reportadas en la literatura, aunque se sugiere que la especificidad de los inhibidores debe ser reconsiderada. Las penicilinas son más susceptibles a ser degradadas por las betalactamasas, posiblemente debido la considerable similitud estructural entre ellas respecto al resto de antibióticos. Consistente con esto, las penicilinas son degradadas en buena manera por todas las clases de betalactamasas.

Parte 2.2.

Usé datos de concentraciones mínimas inhibitorias de serinbetalactamasas para entrenar un modelo de regresión por cada uno de los 21 conjuntos de antibióticos. Si los modelos obtuvieron buenas métricas de desempeño, se usaron para predecir la capacidad de una betalactamasa para degradar distintos antibióticos.

- Determiné el mejor modelo de lenguaje de proteínas al comparar los siete modelos previamente señalados y se añadieron dos representaciones de baja dimensionalidad con PCA de ESM-1b y ProtT5-XL-U50.
- Se compararon 42 algoritmos de regresión para determinar cuál es el mejor.
- Dentro de los antibióticos analizados se encuentran: 11 cefalosporinas, siete penicilinas, dos carbapenemas y una monobactama.

Resultado: los modelos de lenguaje de proteínas con mejor desempeño en la tarea de regresión fueron ProtT5-BFD y ESM. El mejor algoritmo de regresión fue *Nu-Support Vector Machine* y se usó un *kernel Radial Basis Function* para lidiar con la alta dimensionalidad. Sin embargo, solo obtuve un buen modelo de regresión para Cefoxitina basado en cinco métricas de desempeño. Este modelo es capaz de generalizar a familias fuera de los datos de entrenamiento. Además, si se consideran otros regresiones de menor calidad, es posible construir un perfil funcional para las 25,809 betalactamasas el cual distingue relativamente bien a las clases de serinbetalactamasas.

Nota: esta tesis considera varios anglicismos indicados en itálicas y referidos en pies de página. Conservé estos anglicismos debido a que no tienen una traducción equivalente al español, y además, son términos más fáciles de encontrar en motores de búsqueda. De igual forma, todas las figuras de esta tesis se encuentran en inglés debido a que el código y datos usados para su creación también fueron escritos en inglés (por ejemplo, los nombres de los antibióticos).

Introducción

BETALACTAMASAS

DESCRIPCIÓN E IMPORTANCIA

Las betalactamasas son enzimas que degradan antibióticos betalactámicos, los cuales se distinguen por tener una amida cíclica de cuatro miembros llamada anillo betalactámico[1] y por ser los antibióticos más consumidos a nivel global[2,3]. Dado su diversidad estructural, estos antibióticos se organizan en cuatro clases: penicilinas, cefalosporinas, carbapenemas y monobactamas[4].

Las betalactamasas suelen encontrarse en el periplasma y son principalmente producidas por bacterias. Estas enzimas han sido investigadas por más de 80 años[5] debido a que son uno de los principales mecanismos de resistencia a antibióticos en varios patógenos, y su expresión se ha ligado a la aparición de epidemias en distintos países[6].

Las betalactamasas son un modelo de estudio para comprender la estructura, función y evolución de las proteínas[7]. Se estima que estas enzimas han existido por más de 2,000 millones de años[8] y se caracterizan por ser promiscuas[9], ya que pueden degradar distintos antibióticos betalactámicos.

Actualmente, existen dos clasificaciones de betalactamasas: Clasificación molecular[10] que las organiza en cuatro clases (A, B, C y D) que se distinguen por sus mecanismos catalíticos (basados en serinas o en metales), residuos altamente conservados e implicados en la catálisis y por homología estructural y de secuencia (Fig. 1). Clasificación funcional[11] que las organiza en 16 grupos basados en su capacidad de degradar antibióticos betalactámicos e inhibidores. Su nomenclatura está basada en la clasificación molecular y sugiere enzimas representativas de cada grupo[12] (Fig. 2).

ESQUEMAS DE CLASIFICACIÓN

Debido a la alta divergencia a nivel de secuencia, la clasificación molecular se basa en otra alternativa: la homología estructural[13,14]. Por ejemplo, las betalactamasas CfxA y CARB-2 de la clase A son homólogas estructurales con una similitud de secuencia de 16%[15]. Las betalactamasas de las clases A, C y D son homólogas estructurales[16] y en conjunto, se les llama serinbetalactamasas por su mecanismo catalítico basado en serinas. Sin embargo, los microambientes catalíticos de las tres clases de serinbetalactamasas son distintos[17–20] y evolucionaron de forma independiente a partir de diferentes ancestros dentro de la superfamilia *PBP-like*[A] hace más de 2,000 millones de años[9,21,22]. Cada clase de serinbetalactamasas puede dividirse en subgrupos: la clase A se subdivide en dos subclases[23] (A1 y A2), la familia OXA dentro de la clase D en 44 subfamilias[24] y, se ha propuesto que la clase C también puede subdividirse en dos subclases[25] así como la existencia de una subclase A3[26].

Las betalactamasas cuyo mecanismo catalítico se basa en las propiedades redox de distintos metales coordinados a distintos aminoácidos, se llaman metalobetalactamasas, y no son homólogas con las serinbetalactamasas[27]. Inicialmente, todas las metalobetalactamasas se clasificaron en la clase B y se propusieron tres subclases: B1, B2 y B3. Sin embargo, en 2003 se demostró mediante análisis de la estructura de sus proteínas que son dos grupos distintos llamados ME (subclase B3) y MB (subclases B1 y B2)[14]. Este hallazgo generó discusión[28] en la comunidad respecto a la inclusión de una quinta clase en la clasificación molecular[29]. Aunque se estima que el grupo ME evolucionó hace ≈2,000 millones de años y el grupo MB hace ≈1,000 millones de años, aún no es claro si estos dos grupos comparten un ancestro en común[30,31].

Al describir una nueva betalactamasa se suele evaluar su capacidad de degradar diferentes antibióticos betalactámicos e inhibidores con datos de concentraciones mínimas inhibitorias (*MICs*[B]) o parámetros de cinética enzimática. Sin embargo, la recopilación de estos datos no había sido estandarizada hasta que en 2022 Bradford *et al*.[32] fijaron una serie de directrices para describir la capacidad catalítica de una betalactamasa de manera precisa y sistemática.

Aunque las betalactamasas hidrolizan varios antibióticos betalactámicos, cada clase tiene preferencias específicas determinadas por la relación geométrica y fisicoquímica entre la estructura proteica y la estructura del antibiótico[33]. Esta relación ha sido moldeada por millones de años de evolución[8,9] y por el reciente uso intensivo de antibióticos[6,17]. De hecho, las betalactamasas son consideradas enzimas tan eficientes que solo se ven limitadas por la difusión del antibiótico[34–36]. La clase A es buena contra penicilinas y cefalosporinas[15,23], la clase B contra carbapenemas[27,37], la clase C contra cefalosporinas[25,38] y la clase D contra carbapenemas y penicilinas[17,24] (Fig. 2). Sin embargo, estas observaciones no son tajantes[39] y se ha propuesto que la actividad catalítica contra un antibiótico puede evolucionar de forma convergente dentro de cada clase[40].

[A] El término "*PBP-like*" es la abreviación inglesa de *Penicillin-Binding Proteins like superfamily* (Pfam ID: CL0013).
[B] El término "*MIC*" es la abreviación inglesa de *Minimum inhibitory concentration*.

Pese a su utilidad, la clasificación funcional ha sido muy discutida porque es demasiado compleja[41,42], incluye asignaciones subjetivas[43,44], y sobre todo, porque se ha construido solo con betalactamasas de alto interés clínico[45], dejando de lado betalactamasas de varios grupos de bacterias Gram negativas y positivas[23,46,47]. Dado lo anterior, recientemente se han actualizado los esquemas de nomenclatura para todas las clases[32,38], como previamente se había hecho para las clases A[48] y B[37]. Dichos esquemas son necesarios para organizar la diversidad molecular y funcional.

Las betalactamasas clase A son las más diversas de todas y actualmente se dividen en dos subclases, A1 y A2, que se distinguen por aminoácidos conservados en sus secuencias y arreglos estructurales[9,15,16,23]. La subclase A2, separada de A1 por una amplia distancia filogenética, está representada principalmente por secuencias en cromosomas del filo Bacteroidota, y en menor medida por la clase Gammaproteobacteria, la cual se cree tiene origen por transferencia horizontal[15]. Para una mejor organización de la clase A, Philippon *et al*.[15] proponen una clasificación en seis grupos. El grupo A incluye los taxa de la subclase A2. El grupo B contiene taxa de la clase Alphaproteobacteria que suelen ser simbiontes de raíces de plantas, microbios fotosintéticos, acuáticos y fitopatógenos. Los grupos C y E son taxa de la clase Gammaproteobacteria y en menor medida Alpha y Betaproteobacteria. El grupo C se subdivide en varios grupos considerados como betalactamasas de espectro limitado, con familias enzimáticas como TEM, SHV, CARB y OKP de los grupos 2b y 2c de la clasificación funcional (Fig. 2). El grupo D es muy diverso y tiene 20 subgrupos en los que se encuentran algunos patógenos de humanos como *Clostridium* o *Mycobacterium*. El grupo E se compone de varios grupos de enterobacterias con betalactamasas de amplio espectro capaces de degradar oximino-cefalosporinas (grupos 2be y 2f de la clasificación funcional). Finalmente, el grupo F tiene varios grupos cuyos taxa de las clases Alpha, Gamma, y Betaproteobacteria que tienden a ser microorganismos de vida libre y no ser patógenos de humanos.

Visto lo anterior, es claro que las betalactamasas son muy diversas a nivel filogenético, de secuencia de aminoácidos, estructuras proteicas y actividad catalítica. Por tales motivos, una revisión de los esquemas de clasificación de betalactamasas utilizando enfoques basados en datos resulta interesante de explorar y podría ser de gran utilidad dado el debate de sus clasificaciones.

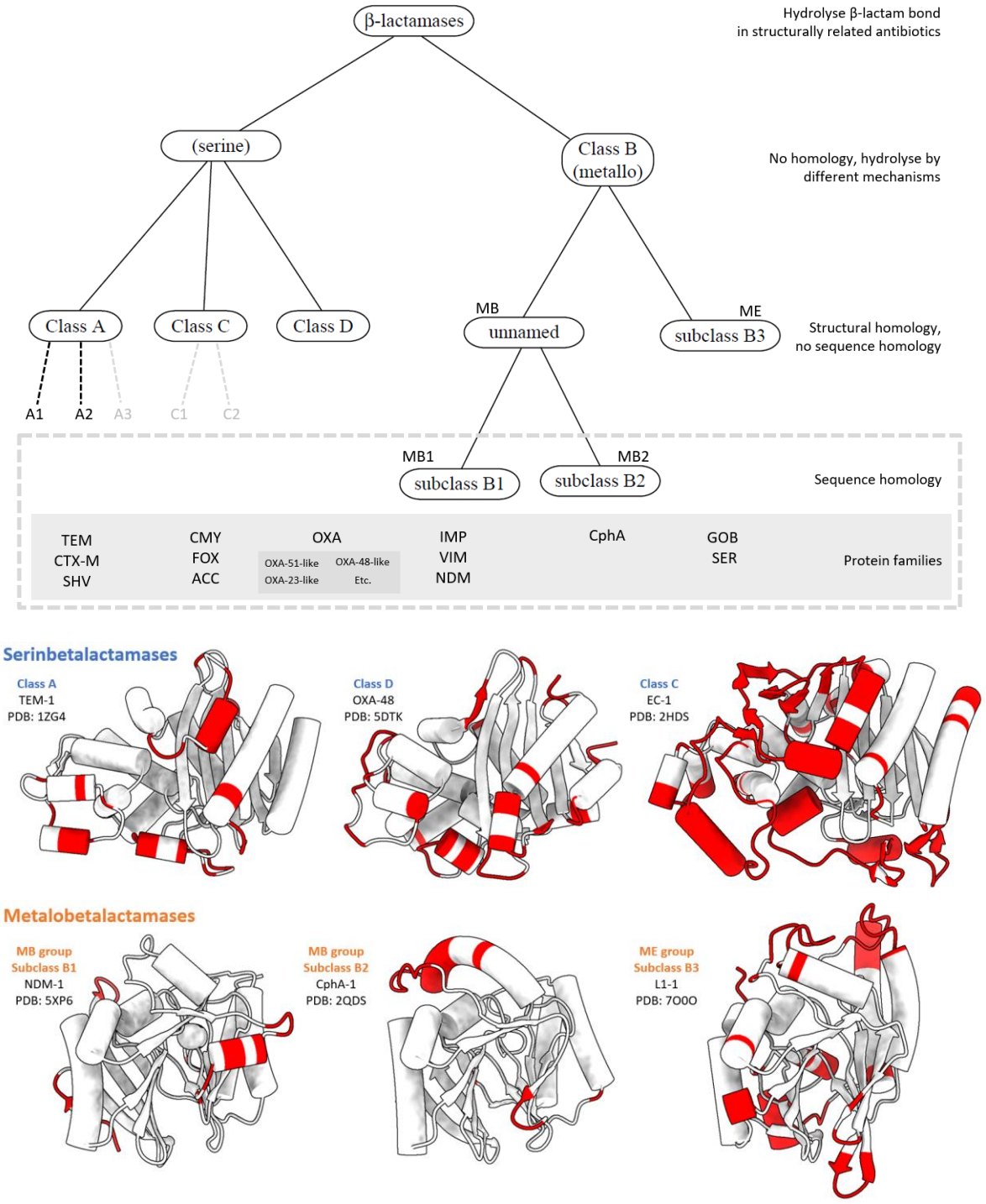


Figura 1. Clasificación molecular de betalactamasas. Panel superior. Se indican dentro de líneas punteadas los grupos donde se suele detectar homología a nivel de secuencia, y en gris sólido algunas de las familias enzimáticas más estudiadas dentro de cada clase. Actualmente se reconocen dos subclases en la clase A[15,23]. En la clase D se reconocen 44 subfamilias dentro de la familia OXA[24]. En la clase C se han sugerido dos subclases[25]. Esquema modificado de la referencia [29]. Panel inferior. Se muestran las estructuras de las serin y metalobetalactamasas, señalando en blanco las regiones en común y en rojo las regiones con mayores diferencias estructurales (*RMSD*, *Root-mean-square deviation*) al ser alineadas con la versión de *TM-align* implementada en la *Protein Data Bank* (*PDB:* https://www.rcsb.org/alignment). Por cada estructura se indica su clase, subclase, familia, variante y el respectivo *PDB ID*. Se seleccionó como representante en cada caso a la estructura cristalográfica más frecuente y con mejor resolución según la *Betalactamase Databse*[49].

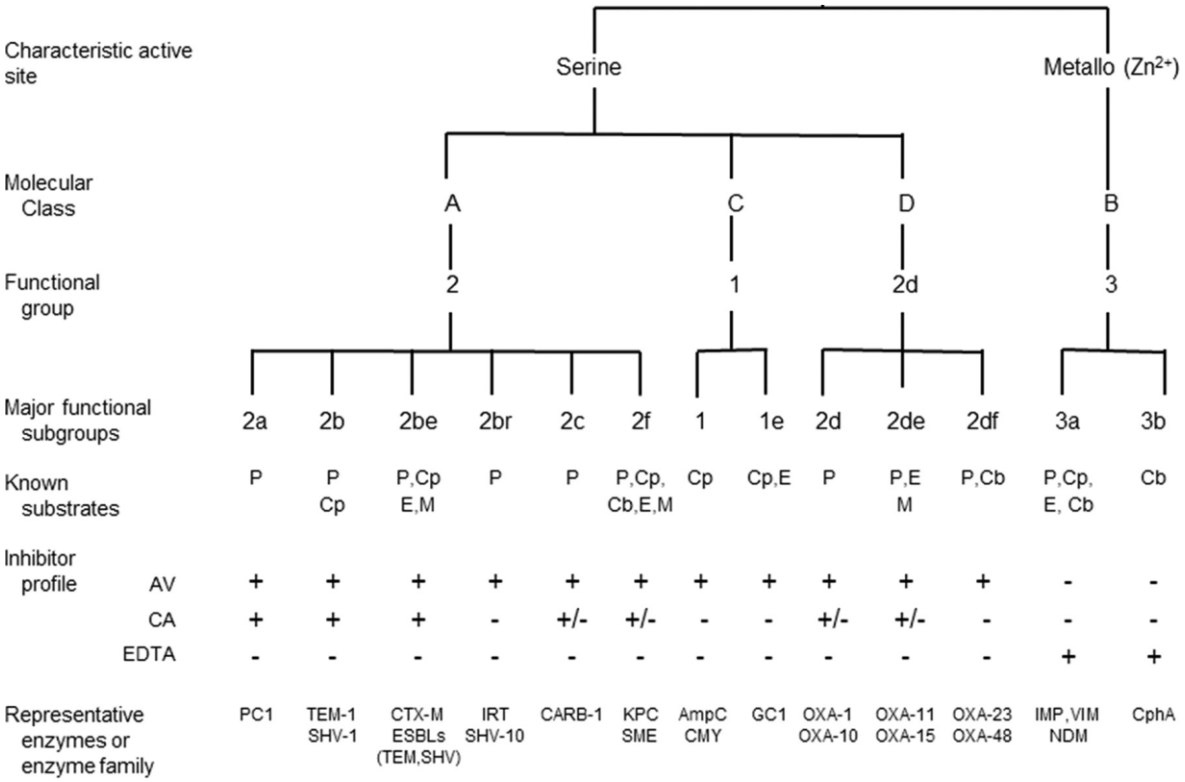


Figura 2. Clasificación funcional de betalactamasas. Se indica con un símbolo "+" si el grupo de enzimas se inhibe, con "-" si no se inhibe y con "+/-" un valor intermedio. Se recomienda revisar las tablas 2.1 y 2.2 de la ref. 12 para más detalles de los 16 grupos funcionales. Abreviaciones: AV, Avibactam; CA, Ácido clavulánico; Cb, carbapenemas; Cp, cefalosporinas; E, cefalosporinas de espectro extendido; M, monobactamas; P, penicilinas. Esquema tomado de la referencia [45].

MODELOS DE LENGUAJE

APRENDIZAJE AUTOMÁTICO Y APRENDIZAJE PROFUNDO

El aprendizaje automático se refiere a un conjunto de algoritmos que aprenden patrones de forma autónoma a partir de datos en un proceso llamado entrenamiento[50]. Estos algoritmos pueden clasificarse según el tipo de datos que procesan en aprendizaje supervisado o no supervisado. El aprendizaje supervisado usa datos etiquetados, es decir, datos a los que se les puede asignar un valor numérico o categórico. Posteriormente, el algoritmo descubre la relación entre los valores de entrada y sus respectivas etiquetas, permitiendo predecir las etiquetas que corresponden a datos que no fueron usados en el entrenamiento. Existen dos enfoques de aprendizaje supervisado, la regresión y la clasificación, que consisten en la predicción de valores numéricos y categóricos, respectivamente[50]. El aprendizaje no supervisado usa datos no etiquetados para generar abstracciones sin indicarle el tipo de resultado esperado, de tal forma que el algoritmo encuentra una estructura interna en los datos, por ejemplo, a forma de una organización secuencial, relacional o jerárquica[50].

En el aprendizaje automático es común trabajar con miles o millones de datos representados por un gran número de variables. A este tipo de datos se les llama datos de alta dimensionalidad debido a que cada dato está representado por un gran número de variables[51]. Sin embargo, a medida que aumenta el número de variables, se dificulta su análisis a nivel teórico y computacional[52]. Unos de los algoritmos de aprendizaje no supervisado usados para analizar datos de alta dimensionalidad son los métodos de reducción de dimensionalidad como *Uniform Manifold Approximation and Projection* (UMAP), *t-Distributed Stochastic Neighbor Embedding* (tSNE) o *Principal component analysis* (PCA). Estos algoritmos transforman los datos originales para crear una representación equivalente de baja dimensión que facilita su análisis y visualización[53]. Sin embargo, durante este proceso se distorsionan los datos y ocurre una pérdida de información; tal como ocurre al representar la superficie tridimensional de la Tierra en un mapa bidimensional que no preserva sus propiedades originales[54].

Los datos de alta dimensión también pueden analizarse con algoritmos de agrupación, como el algoritmo *kmeans*. Estos algoritmos agrupan datos similares en grupos con base en la distancia que existe entre ellos en un espacio vectorial; sin embargo, la forma en que se calcula la distancia entre los datos depende de cada algoritmo[55]. La distancia euclidiana y la similitud coseno son dos de las métricas de distancia más usadas en aprendizaje automático. La distancia euclidiana se considera como la métrica estándar para estimar la distancia entre vectores en gran variedad de análisis, mientras que la similitud coseno es considerada el estándar para analizar datos del área de procesamiento de lenguaje natural[56,57]. Además, es importante tener claro que los métodos de reducción de dimensionalidad no son métodos de agrupación, ya que el objetivo de estos algoritmos es simplificar la representación de los datos, mientras que los algoritmos de agrupación asignan datos a grupos con base en su similitud[53].

Otra rama del aprendizaje automático es el aprendizaje profundo[58], el cual se refiere al conjunto de algoritmos basados en redes neuronales profundas, las cuales se componen de unidades mínimas de procesamiento interconectadas llamadas neuronas[59]. Una neurona es una función no lineal que toma los valores de entrada, los pondera con relación al resto de neuronas y propaga su resultado al resto de neuronas en la red[58]. En estas redes las neuronas se organizan en capas secuenciales, las cuales pueden estar interconectadas entre sí o variar en sus conexiones. A la organización específica de neuronas y el conjunto de operaciones que realizan se le llama arquitectura[60]. Cada arquitectura favorece una forma de aprendizaje especifica que permite modelar mejor ciertas estructuras de datos, a lo cual se le conoce como sesgo inductivo[60]. Por ejemplo, las redes neuronales convolucionales modelan estructuras locales, las redes neuronales de grafos modelan estructuras relacionales y las redes neuronales *transformer*[C] modelan estructuras secuenciales[60].

Desde su aparición en 2017[61], los *transformers* han revolucionado la inteligencia artificial. Actualmente, existen muchas variantes de *transformers*[62], y dado que usualmente se usan para modelar texto, se les incluye dentro de un tipo de algoritmos llamados modelos de lenguaje. Los modelos de lenguaje aprenden una distribución de probabilidad sobre secuencias de palabras que les permite predecir cuál es la palabra más probable dado un contexto. Sin embargo, también han sido usados para modelar distintos datos[62]. Cuando se usan para modelar secuencias de proteínas, hablamos de modelos de lenguaje de proteínas (PLM[D])[63–66]. Los PLM son actualmente una de las herramientas bioinformáticas más interesantes por sus capacidades predictivas, siendo parte fundamental de algoritmos como AlphaFold2[67] y otros más que han demostrado resultados muy prometedores en tareas distintas a la predicción estructural[68–71].

[C] El término "*transformer*" es un anglicismo que se usa para referirse a una arquitectura basada en redes neuronales profundas que usan un mecanismo de atención para modelar la distribución de una secuencia de objetos, los cuales usualmente son cadenas de texto o, proteínas en este caso (ver Figura 3).
[D] La abreviación "PLM" deriva del inglés *Protein Language Models*.

MODELOS DE LENGUAJE DE PROTEÍNAS

Los PLMs pueden ser entrenados usando dos estrategias de modelado principales: la autorregresiva y la autocodificación[72]. En la estrategia autorregresiva el PLM se entrena para predecir el siguiente aminoácido más probable dado una cadena de aminoácidos, lo cual es útil para crear modelos generativos como ProtGPT2[73] o ProGen[74]. En la estrategia de autocodificación, el PLM se entrena para predecir el aminoácido más probable en regiones enmascaradas al azar de la secuencia, lo cual es útil para crear modelos que aprenden a codificar a las proteínas en vectores numéricos llamados *embeddings*[E], los cuales contienen información biológica[75] y son útiles para entrenar otros modelos.

El entrenamiento de un PLM autocodificador inicia con la colecta y limpieza de millones de secuencias de proteínas. Posteriormente, se enmascaran aleatoriamente los residuos de las secuencias. Las secuencias enmascaradas se representan con vectores numéricos binarios (tipo *one-hot*[F]) y se procesan con una operación matemática llamada atención[61], la cual permite identificar las dependencias entre residuos más relevantes. Los residuos con más atención entre sí son aquellos que están en contacto en la estructura o que son importantes para la función biológica, por ejemplo, de unión al sustrato o catálisis[75]. Considerando la atención entre residuos, el modelo predice cuáles son los aminoácidos más probables para la región enmascarada. Finalmente, se comparan las predicciones con las secuencias originales usando una función matemática de evaluación que reajusta los parámetros del modelo según las predicciones sean correctas o no (Fig. 3). Cada PLM autocodificador usa distintas estrategias para realizar todos estos pasos[62].

Los *embeddings* son valores numéricos resultantes de la interacción entre la secuencia de entrada y los parámetros aprendidos por cada capa del modelo. Los *embeddings* pueden generarse a partir de cualquier capa, pero la convención es tomarlos de la última capa pues es donde suele haber más información codificada[70,75–77]. Al procesar una proteína se crea un *embedding* por cada uno de sus residuos, los cuales pueden ser usados en tareas de predicción a nivel de residuo[78]. Para representar una proteína completa, la convención es usar una operación llamada *average pooling*[G] que usa los *embeddings* a nivel de residuo para computar una representación global de una proteína. Sin embargo, existen varias estrategias alternas para crear *embeddings* a nivel de proteína[79–81].

Cuando se usan *embeddings* para entrenar algoritmos se usa una estrategia llamada aprendizaje por transferencia[56,79,82], donde el algoritmo aprovecha la información codificada en los *embeddings* para realizar predicciones. Aplicando esta estrategia a la ingeniería de proteínas se ha establecido un nuevo paradigma llamado "ingeniería basada en pocos datos"[83–85], bastando decenas de datos etiquetados[78,86] o incluso ninguno[87] para obtener buenas predicciones en múltiples tareas de

[E] El término "*embedding*" es un anglicismo que se usa para referirse a la representación numérica de un objeto en un espacio vectorial que un modelo ha aprendido durante su entrenamiento. Dicho espacio vectorial captura el significado y relaciones entre objetos, los cuales son usualmente cadenas de texto o proteínas en este caso (Figura 3).
[F] El término "*one-hot*" es un anglicismo que se usa para referirse a una representación vectorial binaria donde cada aminoácido se convierte en un vector de ceros y unos, donde solo una posición contiene un "1" para indicar la presencia de ese aminoácido en la secuencia.
[G] El término "*average pooling*" es un anglicismo que se usa para referirse al proceso de convertir una secuencia de *embeddings* en una representación vectorial global usando la media aritmética.

predicción a nivel de residuo o a nivel de proteína; por ejemplo, al predecir los efectos de variantes sin sentido en el genoma humano[88] o los residuos en contacto en una proteina[89].

Debido a que los PLM tienen arquitecturas y esquemas de entrenamiento distintos, no son estrictamente comparables entre sí. Sin embargo, para tener una idea de su capacidad, se suele comparar su número de parámetros[90–92]. Los parámetros son valores numéricos que un modelo usa para aprender a representar la información. Generalmente, a medida que aumenta el número de parámetros mejora el desempeño de un modelo en tareas predictivas y emergen otras nuevas, un fenómeno conocido como "leyes de escala"[91,93,94]. Sin embargo, esto ha sido debatido[95], pues el desempeño de un modelo depende de su arquitectura así como de la cantidad, calidad y diversidad de los datos con que se entrena[96]. Además, un mayor número de parámetros requiere de una mayor capacidad de cómputo y complica el proceso de entrenamiento, el cual puede volverse inestable[97]. Estos conceptos aplican tanto a modelos entrenados con cadenas de texto así como modelos entrenado con cadenas de aminoácidos como los PLM[75,76,86,91,98,99].

Varias arquitecturas han sido usadas para entrenar PLM (Tab. 1). Bepler y Berger entrenaron uno de los primeros PLM usando una red neuronal recurrente de memoria de corto-largo plazo junto con información derivada de secuencias y estructuras proteicas[100]. Posteriormente, varios grupos usaron *transformers*[70,76,98,101], los cuales han demostrado tener un mejor desempeño respecto al resto de arquitecturas. Actualmente los *transformers* son considerados el estado del arte, siendo los modelos ESM-1b[101], ProtT5-XLU50[98] y ESM-2[91] los PLM más robustos. Aunque también se ha propuesto que las redes neuronales convolucionales pueden ser competitivas respecto a los *transformers*[77].

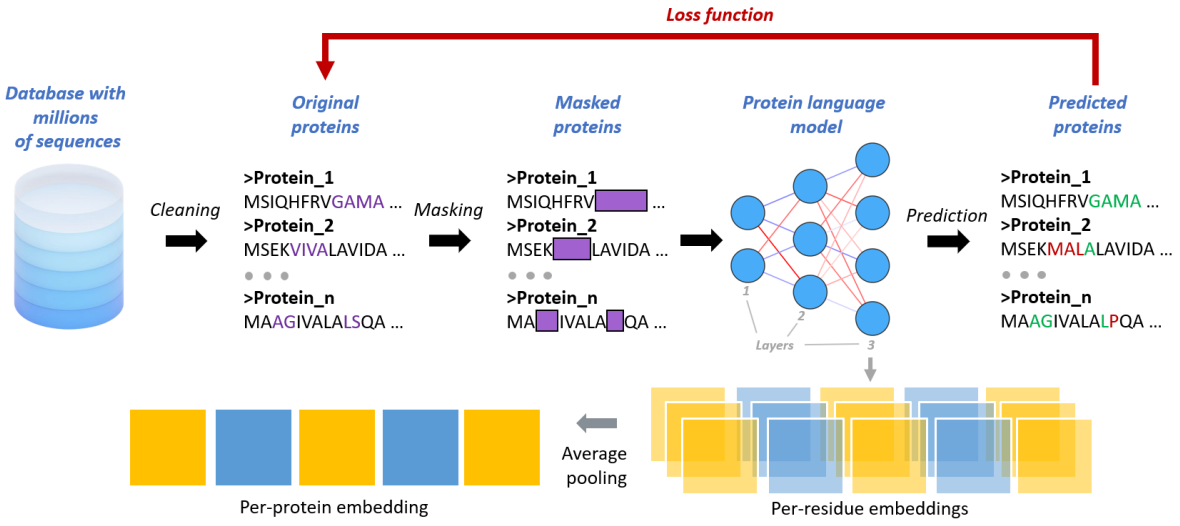


Figura 3. Entrenamiento de los modelos de lenguaje de proteínas autocodificadores. En la parte superior se muestra el proceso de entrenamiento y en la parte inferior la generación de *embeddings* a nivel de residuo y proteína. Las áreas enmascaradas se destacan en morado, las predicciones correctas en verde y las incorrectas en rojo.

Tabla 1. Modelos de lenguaje de proteínas autocodificadores usados en el presente trabajo. Se indica entre paréntesis la variante de la arquitectura *transformer*, así como el número de secuencias usadas en el entrenamiento de los modelos. Las betalactamasas suelen ser secuencias abundantes las bases de datos, sin embargo, el número exacto de serin y metalobetalactamasas con las que se entrenaron los distintos modelos se desconoce. Abreviaciones; M, millones; CNN, red neuronal convolucional; -, desconocido; biLSTM, *Bidirectional Long short-term memory*.

Modelo	Número de capas	Número de parámetros	Arquitectura	Base de datos de entrenamiento (Número de secuencias)	Referencia
Prot-T5-BFD	24	11,000M	*Transformer* (T5)	BFD (~2,122M)	98
Prot-T5XL-U50	24	3,000M	*Transformer* (T5)	BFD y UniRef50 (~2,122M y ~45M)	98
ESM	34	669.2M	*Transformer* (RoBERTa)	UniRef50/S (~27.1M)	101
ESM-1b	33	652.4M	*Transformer* (RoBERTa)	UniRef50/S (~27.1M)	101
XLNet	30	409M	*Transformer* (XL)	UniRef100 (~216M)	98
CARP	33	640M	CNN	UniRef50 (~41.5M)	77
Bepler	3	-	biLSTM	Pfam (~21.8M)	100

Pese a que los PLM han demostrado ser herramientas útiles, aún no es claro cómo operan y se les considera como "cajas negras" de información en comparación con los algoritmos basados en principios biológicos (Fig. 4)[75]. Si los resultados derivados de PLM resultan consistentes con los esquemas de clasificación de betalactamasas, se podría promover su uso en la identificación de grupos de proteínas y su respectiva función biológica. Particularmente, esperamos que, si un PLM ha logrado codificar suficiente información en sus *embeddings*, los resultados de los análisis que usen estas representaciones sean consistentes con las clasificaciones de las betalactamasas.

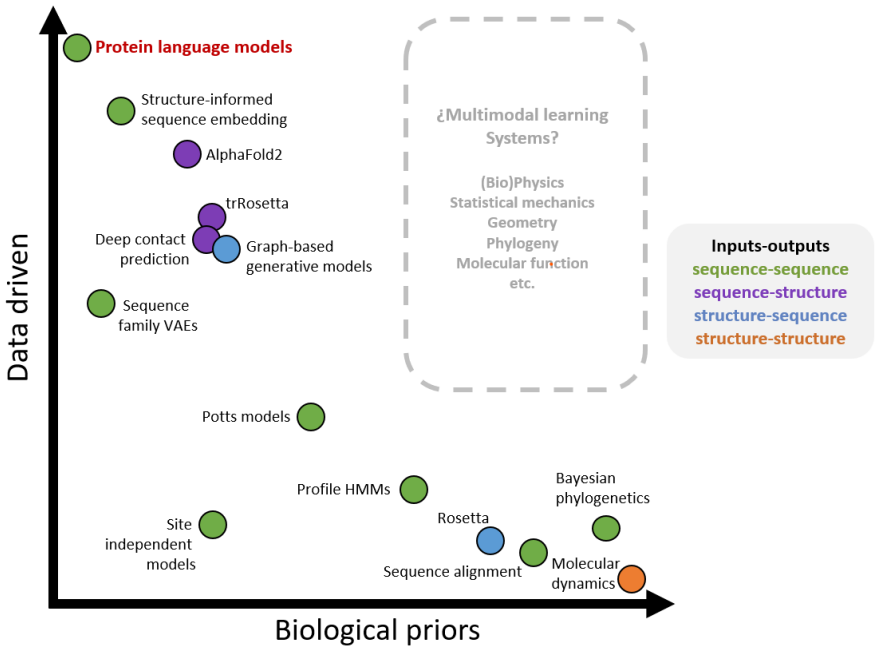


Figura 4. Algoritmos basados en principios biológicos y algoritmos de aprendizaje automático basados en datos. Se señala en rojo a los modelos de lenguajes de proteínas y con puntos de colores el tipo de datos de entrada y salida de los algoritmos. Se indica en un recuadro punteado un espacio hipotético de futuros modelos que integren datos de múltiples fuentes, lo cual se conoce como aprendizaje multimodal[74]. Esquema modificado de la ref. [64].

ANTECEDENTES

Debido al gran interés en las betalactamasas, existen varias bases de datos curadas[49,102,103] así como conjuntos de datos que han sido tomados como estándares para evaluar algoritmos de aprendizaje automático[69–71,104]. Sin embargo, la mayoría de los trabajos solo analizan datos de secuencias representativas, y especialmente, datos relacionados con la betalactamasa llamada TEM-1 dentro de la clase A (Fig. 2).

El trabajo de Figliuzzi *et al*.[105] es uno de los predecesores de los trabajos contemporáneos basados en aprendizaje automático. Los autores modelaron betalactamasas usando información derivada de la covariación de residuos de un alineamiento de secuencias para predecir las MICs de amoxicilina de la TEM-1. Gray *et al*.[106] fueron de los primeros equipos en usar aprendizaje automático para modelar betalactamasas junto con otras proteínas. Específicamente, entrenaron un regresor con datos de experimentos de mutagénesis de la TEM-1 para predecir resistencia contra ampicilina. En 2018, Riesselman *et al.*[107] fueron los primeros en modelar la TEM-1 usando redes neuronales, y su modelo demostró superar al resto de algoritmos publicados hasta la fecha. Alley *et al*.[84] fueron los primeros en usar los *embeddings* de un PLM para modelar la TEM-1. Trabajos más detallados del mismo grupo[85] demostraron que los PLM ofrecen una alta capacidad de generalización a proteínas fuera del conjunto de datos de entrenamiento, y además, demostraron que 24 datos etiquetados son suficientes para obtener buenas predicciones de MICs de la TEM-1 contra ampicilina. Rives *et al*.[101] fueron los primeros en modelar la TEM-1 usando PLM basados en *transformers*, demostrando que esta arquitectura supera en capacidades predictivas al resto de arquitecturas usadas hasta ahora.

La principal ventaja que ofrecen los *transformers* en comparación al resto de arquitecturas es su gran capacidad de generalización a conjuntos fuera de los datos de entrenamiento. Por ejemplo, Verkuil *et al*.[108] demostraron que el *transformer* ESM-2 es capaz de generalizar y asistir al diseño de plegamientos *de novo*, siendo que ESM-2 solo se entrenó con secuencias de proteínas naturales. Otro trabajo con el *transformer* ESM-1b sugiere que este modelo es tan sensible que puede detectar variaciones de un solo residuo y agrupa correctamente unas proteínas llamadas lecitinas[109]. Incluso, PLM que no están basados en *transformers* como el de Bepler y Berger han demostrado generalizar bien entre especies de distintas categorías taxonómicas al predecir interacciones entre proteínas[110].

Aunque los PLM han demostrado ser una herramienta útil, no existe ningún trabajo que los haya usado para identificar los grupos y subgrupos de la clasificación molecular de betalactamasas, así como para predecir la actividad catalítica de betalactamasas no representativas dentro de la clasificación funcional.

PLANTEAMIENTO DEL PROYECTO

Nuestro conocimiento actual sobre betalactamasas se ha construido con algoritmos basados en principios biológicos y solo con betalactamasas representativas. Pese a estar basados únicamente en datos, los PLM han demostrado capturar múltiples señales biológicas. Hasta ahora no se han usado PLM para estudiar la clasificación molecular y funcional de betalactamasas, lo cual sería interesante para verificar si este tipo de algoritmos generan resultados consistentes con lo que sabemos sobre betalactamasas, y posiblemente, encontrar patrones que arrojen nuevas ideas.

En este trabajo evalúe los dos esquemas de clasificación de las betalactamasas usando PLM. Para ello, usé bases de datos curadas junto con los *embeddings* de siete PLM previamente reportados para construir una estrategia de modelado que permita identificar el tipo de información que estos algoritmos están detectando. Además, entrené regresores usando MICs para predecir la actividad catalítica de serinbetalactamasas no representativas contra distintos antibióticos betalactámicos.

Esta estrategia podría mejorar la comprensión del funcionamiento de los PLM y aportar nuevas ideas al conocimiento actual de las betalactamasas. Además, podría ayudar a identificar betalactamasas sin anotaciones dentro de la clasificación molecular o funcional. Igualmente, puede servir para delimitar grupos de betalactamasas con los cuales realizar análisis detallados de sus firmas a nivel de secuencia y estructura en relación con los antibióticos que degradan o con los grupos con los que se asocian. Finalmente, esta estrategia puede servir como referencia para futuros trabajos basados aprendizaje automático que busquen predecir distintas propiedades de las betalactamasas que permitan un monitoreo automatizado y específico, como lo puede ser la predicción a nivel de clase y familia enzimática[111]. Dichos sistemas podrían ser aplicados a betalactamasas de distintos biomas o linajes evolutivos para evaluar el riesgo de resistencia múltiple contra antibióticos betalactámicos.

Este proyecto enfrentó tres principales retos:

1. Falta de estándares de análisis: no existía una estrategia firmemente establecida para evaluar PLM en la identificación de secuencias.
2. Datos muy escasos: los datos disponibles de MICs de betalactamasas no habían sido colectados de forma estandarizada y sistemática de tal forma que permitieran contar con un conjunto abundante de datos[40].
3. Primer trabajo en su tipo: no existía un trabajo previo que evaluara todas las clases de betalactamasas usando distintos PLM para valorar si la información codificada en los *embeddings* es suficiente para predecir datos de MICs[1,112].

HIPÓTESIS

1. Si los *embeddings* de los PLM contienen suficiente información sobre betalactamasas, se podrá identificar claramente a las clases, subclases, familias y subfamilias de serin y metalobetalctamasas (Fig. 1).
2. Si los *embeddings* de los PLM contienen suficiente información sobre betalactamasas, se podrán predecir los valores de la actividad catalítica de betalactamasas con una buena correlación respecto a los datos derivados de experimentos *in vitro* (Fig. 2).

OBJETIVOS

En este trabajo se tuvieron dos objetivos generales y cuatro específicos para cada uno de ellos:

1. Crear una estrategia para modelar betalactamasas usando PLM:
    a. Descargar, ordenar y etiquetar secuencias curadas de betalactamasas[49].
    b. Enriquecer el conjunto anterior con otras betalactamasas de referencia[8,113–115].
    c. Usar distintos PLM preentrenados para generar los *embeddings* a nivel de proteína por cada una de las secuencias de betalactamasas[116] (Tab. 1 y Fig. 3).
    d. Comparar algoritmos que permitan visualizar la organización de betalactamasas y evaluar su habilidad para distinguir los grupos de la clasificación molecular (Fig. 1).
2. Entrenar regresores usando los *embeddings* de distintos PLM para predecir la actividad catalítica de betalactamasas:
    a. Descargar, ordenar y etiquetar datos curados de MICs de betalactamasas[102].
    b. Analizar la relación funcional y estructural entre betalactamasas y antibióticos.
    c. Entrenar regresores usando MICs de betalactamasas representativas.
    d. Si los regresores son buenos, usarlos para predecir la actividad catalítica de las secuencias del primer objetivo y evaluar su consistencia con la literatura.

JUSTIFICACIÓN

Las betalactamasas son enzimas importantes porque generan resistencia contra antibióticos betalactámicos, los cuales representan más de la mitad del consumo global de antibióticos[2,3]. Sin embargo, sólo conocemos la actividad catalítica de un puñado de enzimas representativas de interés clínico[12,45]. Por dichos motivos, explorar nuevas estrategias de identificación y anotación funcional de grupos de betalactamasas usando PLM podría ofrecer un mejor entendimiento de este grupo de enzimas, así como del desempeño estos nuevos algoritmos basados únicamente en datos[63,64].

Métodos y resultados

PARTE 1. PATRONES DE ORGANIZACIÓN

PROCESAMIENTO DE LA BASE DE DATOS

Esta primer parte consiste en crear una estrategia que permita evaluar que tan bien distinguen los PLM a los distintos grupos de betalactamasas al usar distintas propiedades cualitativas y cuantitativas derivadas de sus secuencias de aminoácidos, lo cual ayudará a ver qué tipo información biológica detectan y dar una mejor interpretabilidad a este tipo de modelos.

Para ello, accedí, descargué (02 de Marzo de 2022), ordené y etiqueté todas las secuencias en la *Betalactamase Database*[49] (BLDB). La BLDB organiza las betalactamasas en cuatro clases (A, B, C y D) y distingue a la clase B en tres subclases (B1, B2 y B3). La BLDB nombra a los archivos usando el formato Clase-Familia-Variante e incluye esta información los encabezados de las secuencias. Las secuencias nombradas como AFAM-[número] son secuencias cuya clasificación solo llega a nivel de clase y sus encabezados suelen tener información relativa a muestras ambientales, especie bacteriana o mecanismo catalítico. Es decir, la BLDB se compone de betalactamasas asignadas a una familia conocida (como las de familia TEM), así como de betalactamasas de origen metagenómico, y no son secuencias creadas con algún modelo generativo[73,74]. El siguiente es un ejemplo de los datos para la betalactamasa TEM-1 de la clase A:

> Nombre de archivo: *A-TEM-1-prot.fasta*
> Encabezado:
> *>gi|30230644|gb|AAP20891.1|TEM-1| class A broad-spectrum beta-lactamase TEM-1*

Descarté siete secuencias cuyo formato no está relacionado a una clase de betalactamasas. También removí tres secuencias sin de información, obteniendo un total de 25,809 secuencias (Fig. 5). A la fecha de colecta de datos, la BLDB reconocía un total de 290 familias en las cuatro clases de betalactamasas. Para etiquetarlas, usé la información de sus encabezados y todas las secuencias sin registro con alguna familia reconocida por la BLDB fueron etiquetadas como “Otras” (Fig. Sup. 1).

Para enriquecer estos datos generé una “secuencia firma” por cada una de las cuatro clases de betalactamasas y tres subclases de metalobetalactamasas, las cuales representan los residuos más conservados de sus respectivos grupos. Para ello usé un procedimiento basado en la referencia[117]. Brevemente: por cada clase y subclase, descarté las secuencias fuera del rango de ±30% del valor de la mediana de la longitud de secuencia del grupo y posteriormente usé Cd-hit[118] para agrupar las betalactamasas a un 90% de identidad de secuencia para reducir la redundancia y obtener secuencias

representativas. Finalmente, usé las secuencias representativas para construir un alineamiento con MAFFTv7[119], usé BuddySuite[120] para remover los *gaps*[H] del alineamiento y construir una secuencia firma con un enfoque de pesos (comando: *alignBuddy -con weighted*).

Además, añadí otras secuencias de referencia que derivan de un alineamiento de 75 betalactamasas clase A que corresponden a: una secuencia consenso[113] (construida con un enfoque frecuentista), tres secuencias ancestrales de distintos linajes evolutivos[8,114] y una mutante de un ancestro que asemeja la capacidad catalítica especialista en penicilinas de la TEM-1[115].

Para obtener más información de las secuencias usé Biopython[121], SeqKit[122] y ProtLearn[123] para estimar distintas propiedades fisicoquímicas y estadísticas (Tab. 2). Además, anoté la taxonomía de las secuencias usando el protocolo de la referencia[124]. Brevemente: la anotación se realiza por similitud de secuencia contra las proteínas de la *Genome Taxonomy Database* (RS207)[125]. Descargué las proteínas de los genomas representativos de bacterias y arqueas y con ellas creé una base de datos de Diamond2[126] contra la que alineé mis betalactamasas (usando parámetros predeterminados, *i.e*. 25 registros). Finalmente, la taxonomía de cada secuencia se anota usando el algoritmo de último ancestro en común considerando los registros identificados con Diamond2.

Tabla 2. Propiedades cualitativas y cuantitativas de las secuencias de betalactamasas.

Propiedad	Breve descripción	Ref.
Longitud	Conteo del número de residuos de una proteína	122
Masa molecular	Sumatoria del masa molecular (en Daltons) de cada uno de los residuos de la proteína	121
Aromaticidad	Fracción de aminoácidos con propiedades aromáticas (F, W, Y) en la secuencia	121
Inestabilidad	Inestabilidad de la proteína estimada a partir de la frecuencia de dipéptidos de baja y alta estabilidad observados en proteínas estables e inestables. Valores >40 significan que la proteína es inestable (tiene una vida media corta).	121
Hidropatía (*AKA Gravy*)	Se estima como la sumatoria de la hidropatía de cada residuo dividía por la longitud de la secuencia.	121
Punto isoeléctrico	Estimación del pH en el cual la proteína no tiene carga eléctrica neta	121
Entropía de secuencia	Estimación de la diversidad de la composición de secuencia. Valores bajos significan una menor diversidad de aminoácidos en la composición. Una secuencia con mínima entropía es una secuencia compuesta por un solo tipo de residuo, mientras que una secuencia con máxima entropía tiene todos los residuos posibles en proporciones iguales.	123
Estructura secundaria	Estimaciones realizadas a partir de la fracción de residuos asociados a regiones hélice (V, I, Y, F, W, L), beta plegada (E, M, A, L) y giro (N, P, G, S).	121
Taxonomía	Anotación a nivel de Filo, Clase, Orden, Familia, Género y Especie por contra la *Genome Taxonomy Database*.	124
Conjuntos de referencia	Ancestros y secuencias firma.	8,113–115
Clasificación molecular de betalactamasas	Etiquetas correspondientes a las clases (A, C y D), subclases (A1, A2, B1, B2, B3. C1, C2), familias (290 distintas) y subfamilias (44 dentro de la clase D).	49

[H] El término "*gap*" es un anglicismo que se usa para referirse a una penalización que se introduce para maximizar el número de coincidencias en un alineamiento múltiple de secuencias, el cual se suele señalar con el símbolo "-".

ANÁLISIS DE LOS *EMBEDDINGS*

Se analizaron individualmente las serin y las metalobetalactamasas debido a que no son grupos homólogos entre sí. Por cada secuencia de las betalactamasas generé sus *embeddings* a nivel de proteína (Fig. 3) usando siete PLM (Tab. 1) con la librería bio_embeddings[116] y con el *script*[l] *extract.py* para el modelo CARP provisto en su repositorio en GitHub.

A diferencia de los datos multivariados tradicionales donde cada una de las variables suelen ser independientes, lineales o interpretables[52,53], la información contenida en los *embeddings* de un modelo de lenguaje no es independiente, es no lineal y no tiene una interpretación clara; aunque se ha encontrado que algunas neuronas parecen haberse especializado en la detección propiedades de las proteínas como regiones de estructura secundaria o accesibilidad al solvente[84]. Por lo tanto, no se suele aplicar pruebas estadísticas comunes para analizar este tipo de datos no lineales, dado que estas pruebas son modelos lineales[127]. En su lugar, para analizar el tipo y organización de la información codificada en los *embeddings* es común realizar análisis de reducción de la dimensionalidad y mapear propiedades cualitativas o cuantitativas[64,84,85,128]. Sin embargo, estos análisis son limitados pues dado su estocasticidad, solo ayudan a hacer observaciones cualitativas, son propensos a mal interpretarse[53,129] y su solidez y reproducibilidad no son fáciles de evaluar[130–133]; por dichos motivos, este tipo de análisis debe de ser usado para plantear hipótesis y no para realizar conclusiones[129]. Otros métodos para analizar *embeddings* es medir su similitud usando métricas como la distancia Euclidiana y similitud coseno[79] o evaluando su desempeño en tareas de agrupación[134], clasificación[68,69] o regresión[135]. También es posible combinar varios de los métodos antes mencionados[136], usar análisis de mapeo de atención[75,137] o usar métodos basados en geometría analitica[56] y topología algebraica[138]; sin embargo, estos últimos métodos son complejos y su aplicación es incipiente en modelos de lenguaje de proteínas. Por estos motivos, analicé los *embeddings* usando tres estrategias distintas: algoritmos de reducción de la dimensionalidad, métricas de similitud y algoritmos de agrupación no supervisados (Fig. 5).

Para representar la información de los *embeddings* en dos dimensiones, usé tres algoritmos: PCA que usa transformaciones lineales[139], y tSNE[140] y UMAP[141] que usan transformaciones no lineales. A diferencia de PCA, los resultados de tSNE y UMAP dependen mucho del valor de sus parámetros clave. tSNE depende de un parámetro (*perplexity*) y UMAP depende de tres parámetros (*n_neighbors*, *spread* y *min_dist*)[142]. tSNE tiende a preservar las distancias locales, mientras que UMAP tiende a preservar las distancias locales y globales, aunque esto sigue siendo muy debatido[130,143–146]. Comúnmente, las representaciones de baja de dimensión de PCA pueden ser evaluadas mediante la cantidad de varianza modelada. tSNE puede ser evaluado con la divergencia de Kullback-Leibler; esta métrica toma valores entre cero a infinito positivo y representa la discrepancia entre la distribución de los datos en la alta dimensión respecto a la distribución de los datos en la baja dimensión. Bajos valores de divergencia indican una mejor fidelidad de las representaciones de baja dimensión[147–150].

[l] El término "*script*" es un anglicismo que se usa para referirse a un fragmento de código aislado y autocontenido de un lenguaje de programación que puede ser ejecutado por múltiples intérpretes.

Por su parte, la implementación de UMAP en Python no cuenta con una forma de evaluar la fidelidad de las representaciones en la baja dimensión[J]; aunque recientemente se han propuesto formas alternas de hacer esto basadas en enfoques probabilísticos[130,151]. Por dichos motivos evalué el efecto de la *perplexity* y número de iteraciones en tSNE para identificar las representaciones más fieles a los datos originales de alta dimensión de cada PLM (Fig. Sup. 2-4). En el caso de UMAP donde no se puede evaluar directamente la fidelidad de las representaciones, seleccioné una configuración de parámetros similares a las de tSNE (Fig. Sup. 5-8). Una vez evaluada la estabilidad de las representaciones y determinado el conjunto de parámetros que dan mejores resultados, realicé gráficos de dispersión con las dos dimensiones a las cuales mapeé las propiedades cualitativas y cuantitativas (Tab. 2) para evaluar sus patrones de organización.

El criterio para evaluar las representaciones de baja dimensión se basó en que los miembros de cada clase deben estar cercanos entre sí y lejos de las otras clases, una propiedad cualitativa conocida como homofilia[133]. En otras palabras, si los *embeddings* de los PLM contienen suficiente información, se espera que los PLM identifiquen claramente a las clases, subclases, familias y subfamilias de betalactamasas en regiones específicas. Igualmente se espera que las secuencias firma se agrupen dentro de sus respectivas (sub)clases, pues son una especie de "promedio" de la diversidad de aminoácidos[152,153]. La diferencia entre una secuencia firma y una ancestral es la forma en que se estiman. Mientras que la secuencia consenso se puede entender como un conteo del aminoácido más probable en una posición del alineamiento, los ancestros derivan de una filogenia que toma en consideración la longitud de las ramas[154,155]. Previamente, Risso *et al*. han sugerido que una secuencia consenso es una buena aproximación de un ancestro[113,114]. Si esto es cierto, podemos esperar que los ancestros y consensos de la clase A se agrupen en un espacio en común.

Para evaluar la capacidad de identificación de las clases de betalactamasas con algoritmos no basados en reducción de la dimensionalidad, muestreé 100 secuencias al azar por cada (sub)clase de betalactamasas y usé los *embeddings* para computar sus distancias por pares usando dos métricas de similitud[156]: distancia Euclidiana y similitud coseno. Si los *embeddings* de los PLM contienen suficiente información sobre betalactamasas, podemos esperar que los *embeddings* de cada (sub)clase sean más similares entre ellos que con los *embeddings* de otras (sub)clases. Estos mismos *embeddings* muestreados al azar fueron usados para identificar grupos con el algoritmo *Kmeans*[50], el cual es un algoritmo de agrupación no supervisada que asigna datos a *k* grupos previamente indicados. Si los *embeddings* contienen suficiente información sobre betalactamasas, podemos esperar que al determinar un valor de $K = 2$ se distingan a los grupos ME y MB en las metalobetalactamasas dado sus diferencias a nivel de estructura proteica[14] (Fig. 1); y en caso de las serinbetalactamasas, podemos esperar que se distinga a las clases A y D de la clase C, debido a que las clases A y D tienen una mayor similitud entre si a nivel filogenético y estructrual[13] (Fig. 1). Si el valor de $K = 3$ podemos esperar que se distingan las tres subclases de metalobetalctamasas (B1, B2 y B3) así como las tres clases de serinbetalactamasas (A, C y D) (Fig. 1). Ambos valores de k fueron evaluados mediante la métrica de agrupamiento *adjusted rand index* (ARI) contra las clases reales.

[J] https://github.com/lmcinnes/umap/issues/100

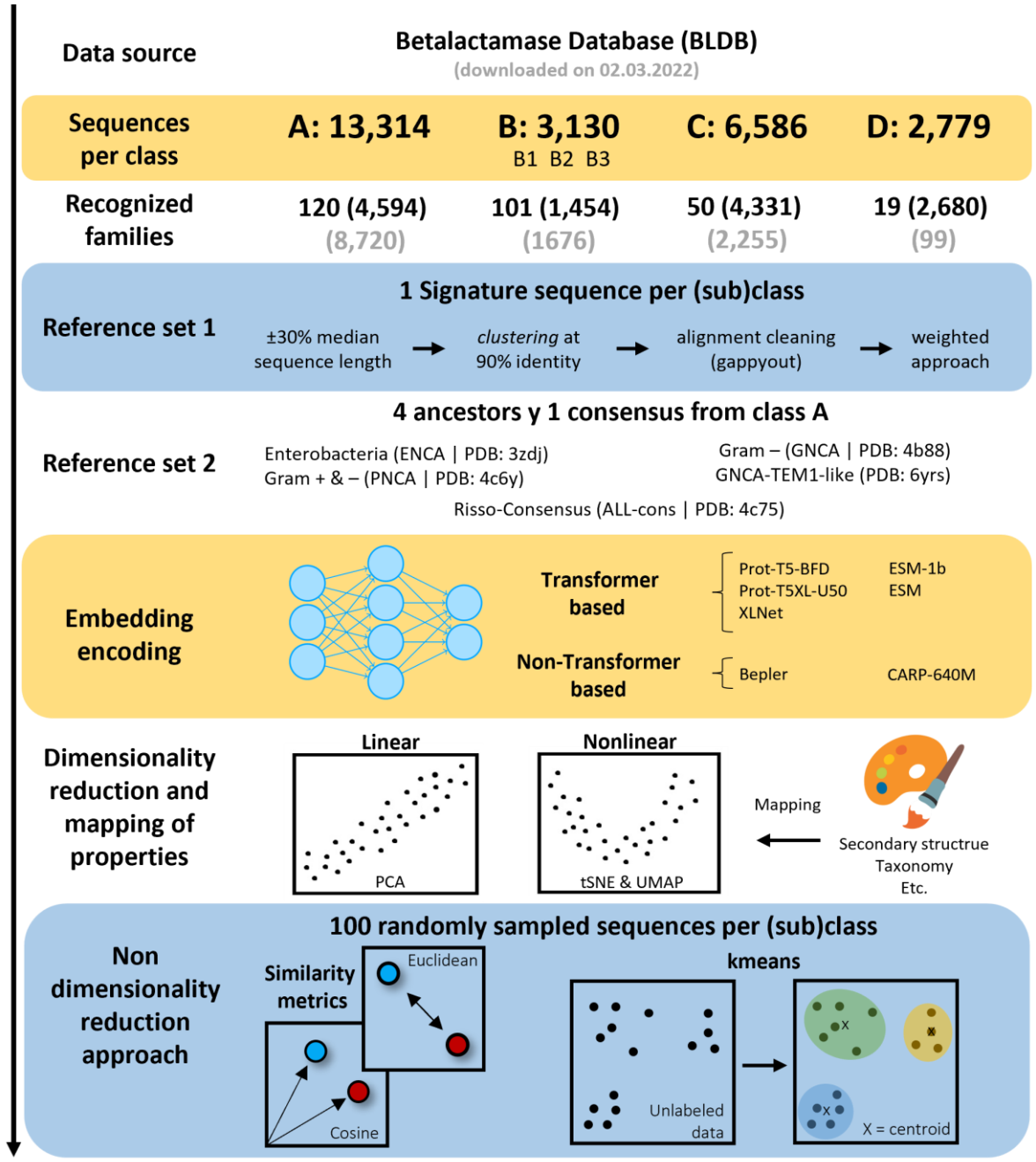


Figura 5. Estrategia de modelado de las secuencias de betalactamasas. Se analizaron 25,809 secuencias curadas por la BLDB[49], la cual reconoce 290 familias enzimáticas. Todas las secuencias que no mapearon a estas familias se etiquetaron como "Otras" (Fig. sup. 1). Se indica entre paréntesis el número de secuencias reconocidas y en color gris al número de secuencias etiquetadas como Otras. Como referencias, se estimaron secuencias firma por cada (sub)clase de betalactamasas (7 secuencias), y se incluyeron secuencias de ancestros y un consenso que derivan un alineamiento de 75 betalactamasas clase A estimadas por Risso *et al.*[8,113–115] (5 secuencias). Además, se incluyeron 201 secuencias de una putativa nueva subclase de metalobetalactamasas llamada VarG[157] y la betalactamasa LRA-5, la cual se ha sugerido como representante de una putativa nueva subclase A3[26,158]. Para el total de 26,023 secuencias se computaron sus *embeddings* a nivel de proteína de la forma convencional (Fig. 3) usando siete modelos de lenguaje de proteínas (Tab. 1). Para visualizar la organización derivada de los *embeddings* se usaron tres algoritmos de reducción de la dimensionalidad para crear dos dimensiones en las cuales mapear distintas propiedades cualitativas y cuantitativas (Tab. 2). Finalmente, se realizó un muestreo al azar de 100 secuencias por cada (sub)clase y se analizaron sus *embeddings* usando dos métricas de similitud (distancia Euclidiana y similitud coseno) y *kmeans* para evaluar la capacidad de detección de las clases de betalactamasas.

METALOBETALACTAMASAS

A continuación, se presentan los resultados de los métodos de reducción de la dimensionalidad. La finalidad es encontrar un algoritmo que identifique a los grupos de betalactamasas y no se trata de una comparación exhaustiva de este tipo de algoritmos; pues para llegar a observaciones generales, es necesario incluir más de los tres algoritmos aquí usados, así como distintos conjuntos de datos de proteínas. Para evaluar de forma general el desempeño de los PLM consideré su capacidad de separar a los grupos ME (subclase B3) y MB (subclases B1 y B2) de metalobetalactamasas, y las clases A, C y D de serinbetalactamasas (Fig. 1).

Los resultados de PCA sugieren que ESM-1b y Prot-T5XL-U50 producen los mejores resultados pues son los únicos que separan a los grupos ME y MB. Por el contrario, el modelo Bepler muestra el peor desempeño al no separar a estos grupos y mapearlos en una misa región (Fig. 6).

Al evaluar el efecto de diferentes valores de *perplexity* en la divergencia de Kullback-Leibler, noté que todos los PLM tienen menores divergencias cuando se utiliza el valor más alto considerado (*i.e. perplexity* = 400), razón por la cual elegí las representaciones obtenidas con dicho valor. Lo anterior sugiere que todos los PLM parecen haber sido representados de forma fiel en la baja dimensión, sin embargo, no todos los PLM logran separar a los grupos ME y MB (Fig. sup. 2). Con tSNE la mayoría de los PLM separan a los grupos MB y ME excepto por los modelos Bepler y XLNet que los ubican en un espacio en común, en contrate con modelos como ESM-1b o Prot-T5XL-U50 que los distinguen en regiones claras y más separadas entre si (Fig. 7).

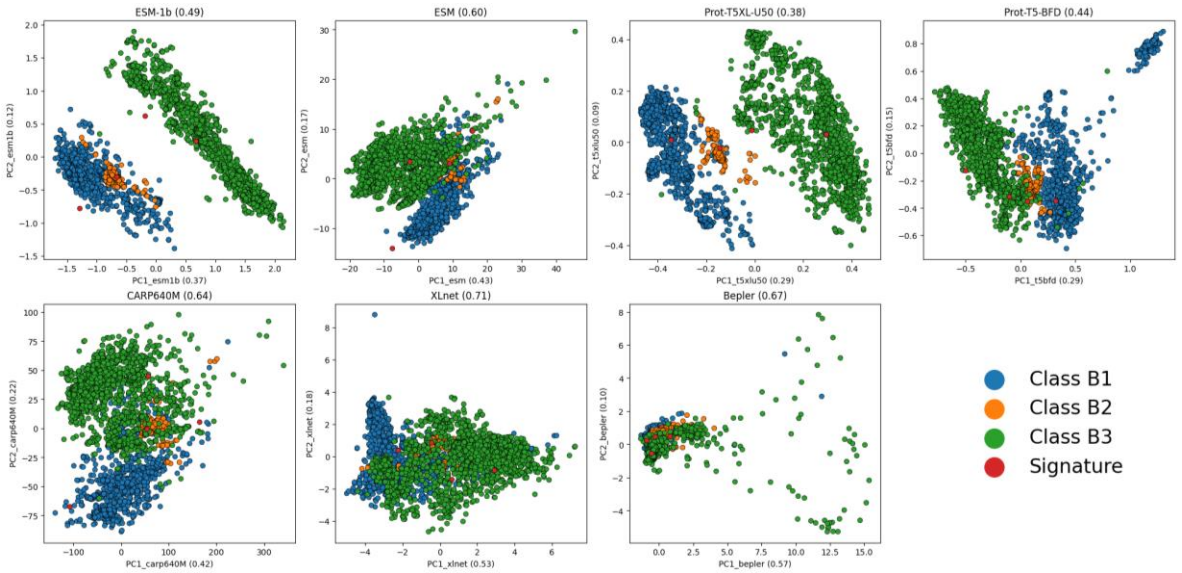


Figura 6. Análisis de componentes principales con metalobetalactamasas. Se indica en el encabezado el modelo de lenguaje de proteínas y entre paréntesis la varianza total representada por los componentes. En cada eje se muestra entre paréntesis la varianza representada por el componente.

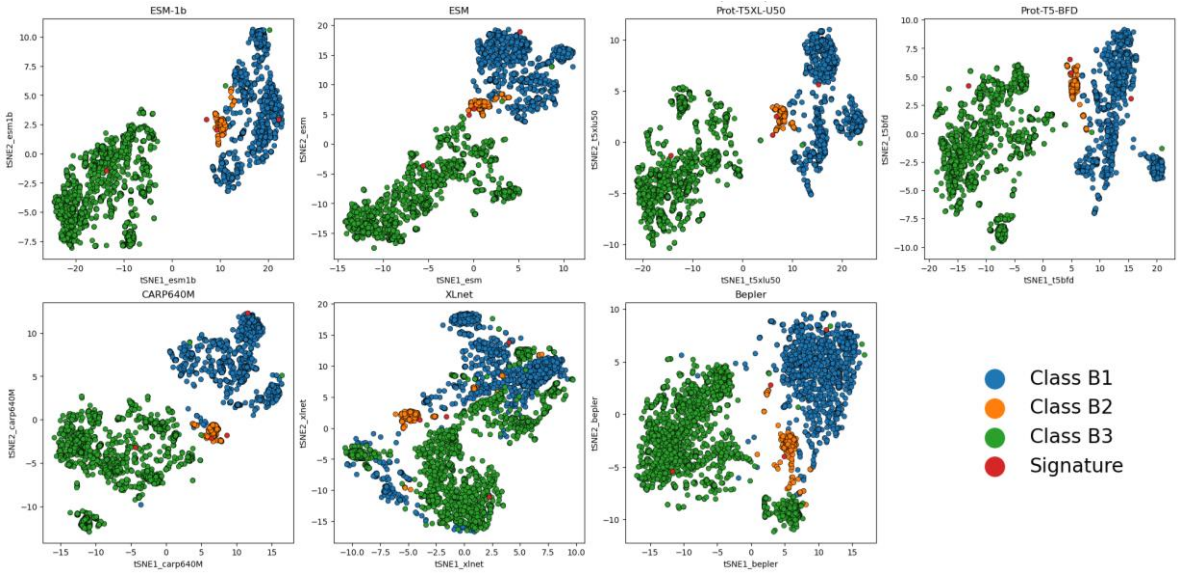


Figura 7. Análisis de reducción de la dimensionalidad con tSNE con metalobetalactamasas. Se indica en cada encabezado el modelo de lenguaje de proteínas correspondiente. Parámetros: perplexity = 400; iteraciones = 1500.

Al igual que tSNE, con UMAP la mayoría de los PLM separan a los grupos MB y ME excepto por los modelos Bepler y XLNet que los agrupan en un espacio en común, en contrate con modelos como ESM-1b o Prot-T5XL-U50 que los distingue en regiones claras y separadas entre si (Fig. 8).

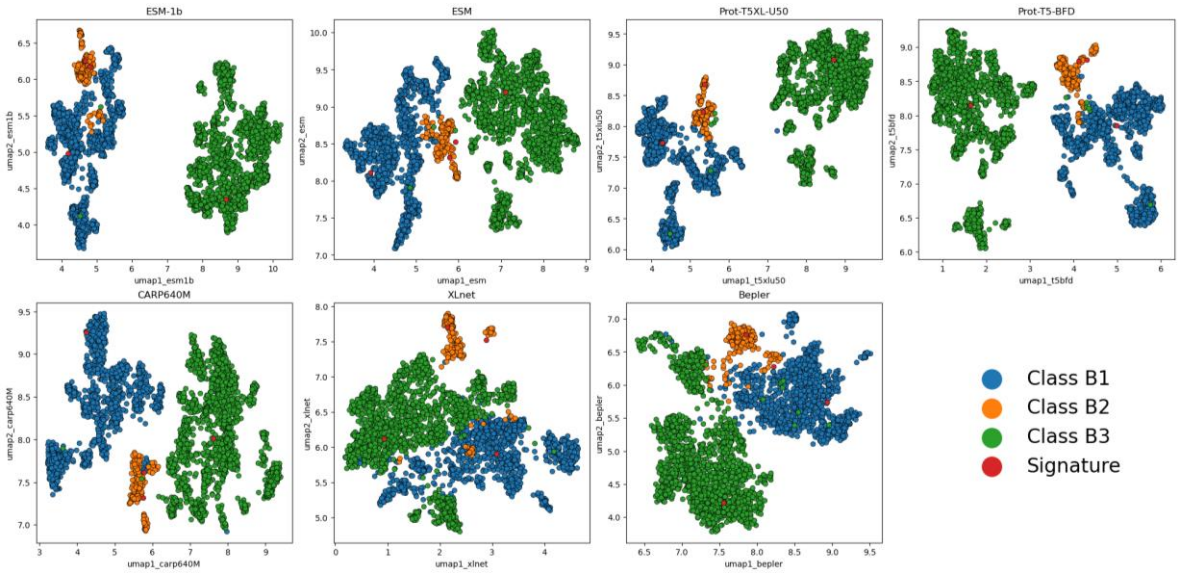


Figura 8. Análisis de reducción de la dimensionalidad con UMAP con metalobetalactamasas. Se indica en cada encabezado el modelo de lenguaje de proteínas correspondiente. Parámetros usados: n_neighbors = 400; mindist y spread = 0.2.

Tomando en conjunto los resultados, sugiero que tSNE y UMAP generan mejores representaciones respecto a PCA. Con PCA solo dos de los siete PLM separan a los grupos ME y MB, mientras que con tSNE y UMAP la mayoría de los modelos separa dichos grupos, excepto por los modelos XLNet y Bepler. En relación con el resto de PLM, ESM-1 y Prot-T5XL-U50 lograron separar a los grupos ME y MB en regiones bastante claras lo cual parece consistente con trabajos previos que sugieren que estos dos PLM cuentan con un buen desempeño respecto a otros modelos[56,79,159,160]. Vale la pena resaltar que aunque las betalactamasas de las subclases B1 y B2 suelen compartir una identidad de secuencia tan baja como del 20%, aún se les considera como homologas estructurales[9], lo cual se estaría reflejando en la agrupación de estas subclases.

Dado que conozco la fidelidad de las representaciones de tSNE (Fig. sup. 2) y que estas dependen de menos parámetros respecto a UMAP, decidí tomar los resultados de tSNE para los siguientes análisis. Igualmente me centré en los resultados con ESM-1b al haber separado claramente a los grupos ME y MB. Sin embargo, otros PLM como Prot-T5XL-U50 son igualmente interesantes de explorar.

En la representación de tSNE de ESM-1b dos secuencias de la subclase B3 no se agruparon con el resto de sus secuencias (Fig. 9A). Ambas secuencias corresponden a familias no reconocidas. Al inspeccionar sus encabezados encontré que una de ellas se llama BLEG-1, la cual fue recientemente descrita dentro de la superfamilia con plegamiento similar a metalohidrolasas/oxidorreductasas, pero aún no es claro si corresponde al grupo de metalobetalactamasas o glioxalasas tipo II[161]. Pese a esta ambigüedad, la inclusión de la secuencia BLEG-1 es un error en la BLDB, pues su plegamiento predicho con ESMFold[91,162] no corresponde al de la superfamilia y no tiene relación con la única estructura cristalográfica de BLEG-1 (Fig. sup. 9A). La segunda secuencia anómala[K] se llama VarG; un estudio reciente ha propuesto VarG podría representar una nueva (sub)clase de metalobetalactamasas similar al grupo MB[157], razón por la cual se habría agrupado en la vecindad de este grupo. Todas las secuencias firma se agruparon en sus regiones esperadas excepto por la secuencia firma de la clase B que se ubica en la vecindad del grupo MB (Fig. 9A). Debido a la propuesta de que un consenso es una buena aproximación a un ancestro se esperaba que esta secuencia firma se agrupara cerca del grupo ME, pues es ≈1,000 millones de años más antiguo que el grupo MB[14]. La ubicación de esta secuencia firma puede deberse al procedimiento y secuencias que usé para estimarla.

Al mapear las familias de metalobetalactamasas observé una buena agrupación de sus miembros (Fig. 9B). Del total de 3,130 secuencias, 53.6% (1,676 secuencias) fueron etiquetadas como “Otras”. Por otro lado, las 10 familias más abundantes representan un 31.6% (1,052 secuencias) del total. Lo anterior ilustra la poca diversidad que representan las familias reconocidas y como solo un puñado de estas familias cuentan con un alto número de variantes. Tomando los resultados en conjunto, sugiero que ESM-1b tiene una buena capacidad de distinguir secuencias de metalobetalactamasas a nivel de subclases y familias enzimáticas de metalobetalactamasas.

[K] Uso el término "secuencia anómala" para referirme a aquellas detectadas por tSNE fuera de sus regiones esperadas. Estas secuencias no corresponden a un plegamiento de betalactamasas o son betalactamasas fusionadas con otros plegamientos. En el primer caso, la BLDB las curó erróneamente, mientras que, en el segundo, su integración en la BLDB es ambigua, ya que parte de la secuencia es una betalactamasa. Por ello, opté por usar "anomalía" en lugar de "error" para incluir ambos escenarios.

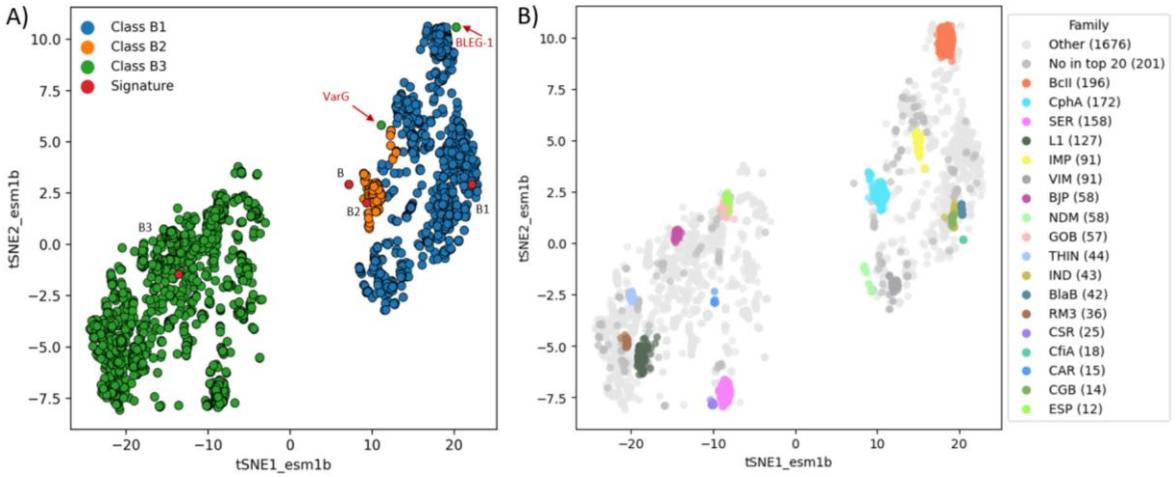


Figura 9. Mapeo de las familias enzimáticas de metalobetalactamasas en las representaciones de tSNE de ESM-1b. (A) Ubicación de las secuencias anómalas y firmas. Se señalan con letras la ubicación relativa de las secuencias firma y con flechas rojas las secuencias anómalas BLEG-1 (ID: AIC95013.1) y VarG (ID: ACQ60975.1). (B) Mapeo de las familias enzimáticas. Por cada familia, se muestra entre paréntesis su número de secuencias. Se muestran de un color distintos a las 18 familias más abundantes, en gris claro las familias identificadas como "Otras" y en gris oscuro las secuencias que no se encuentran dentro de las 18 familias más abundantes.

Para evaluar cómo podría verse reflejado en las representaciones de tSNE de ESM-1b la propuesta de VarG como una nueva subclase[9] de metalobetalactamasas, descargué las 201 secuencias VarG reportadas por Lin *et al.*[157] y computé una nueva representación con tSNE (Fig. 10A). Observé que todas las secuencias VarG se agruparon en un espacio en común y distinto a los grupos ME y MB, sugiriendo que las secuencias VarG cuentan con diferencias respecto a estos grupos. Sin embargo, también se agrupó en esta región la secuencia PNGM-1 que representaría un posible error puesto se asocia a la subclase B3. Lo anterior también sugiere que es necesario un número mínimo de secuencias de un grupo para que este pueda ser detectado por ESM-1b y tSNE, pues de lo contrario sólo se detectará como anomalía.

Para saber si las representaciones de tSNE se ven muy influenciadas por el número de secuencias de un mismo grupo (*i.e*. proteínas que comparten >90% de similitud de secuencia), computé una nueva representación con ESM-1b y tSNE usando sólo las secuencias representativas (Fig. 5). Observé una clara separación de los grupos ME y MB, lo cual sugiere que la organización global no es marcadamente afectada por la cantidad de secuencias de alta similitud (Fig. 10B). En conjunto, los resultados sugieren que esta estrategia puede tener variaciones ligeras a nivel local en función de la cantidad de secuencias de alta similitud, pero representa bien la organización global de las metalobetalactamasas.

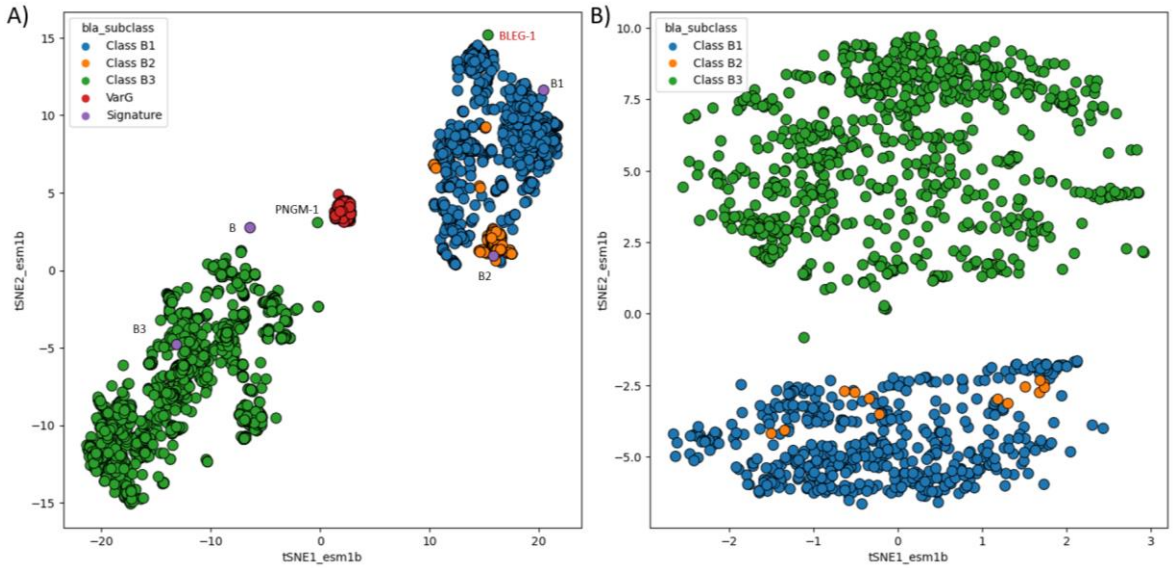


Figura 10. Representaciones de tSNE de ESM-1b al considerar al grupo VarG o secuencias representativas. (A) Organización de las metalobetalactamasas al incluir las secuencias VarG. Se señala con letras la ubicación relativa de las secuencias firma, BLEG-1 (ID: AIC95013.1) y PNGM-1 (ID: AWN09461.1). Divergencia de Kullback-Leibler = 0.25. (B) Organización de las metalobetalactamasas representativas. Después de los filtros de longitud e identidad de secuencia a 90% aplicados para la generación de las secuencias firma (Fig. 5), se obtuvieron un total de 487 secuencias representativas de la subclase B1, 13 de la subclase B2 y 781 de la subclase B3. Divergencia de Kullback-Leibler = 0.17. Ambas representaciones usaron los parámetros perplexity = 400 y 1500 iteraciones.

Al mapear la taxonomía de las secuencias observé que no existe una clara organización en función de las categorías taxonómicas, y solo es posible detectar grupos relativamente claros a nivel de género y especie (Fig. 11). El grupo ME (1,722 secuencias) está principalmente representado por el filo Proteobacteria (1,263 secuencias o ≈73.34%) y en menor medida por Bacteroidota (176 secuencias o ≈10.22%). El grupo MB (1,408 secuencias) está representado por el filo Proteobacteria (578 secuencias o ≈41.05%), Bacteroidota (475 secuencias o ≈33.73%) y Firmicutes (291 secuencias equivalente al 20.66%). El hecho de que las metalobetalactamasas se hayan separado principalmente por subclases y no por su categoría taxonómica sugiere que la taxonomía no representa un factor clave en su organización en las representaciones de baja dimensión de tSNE de ESM-1b.

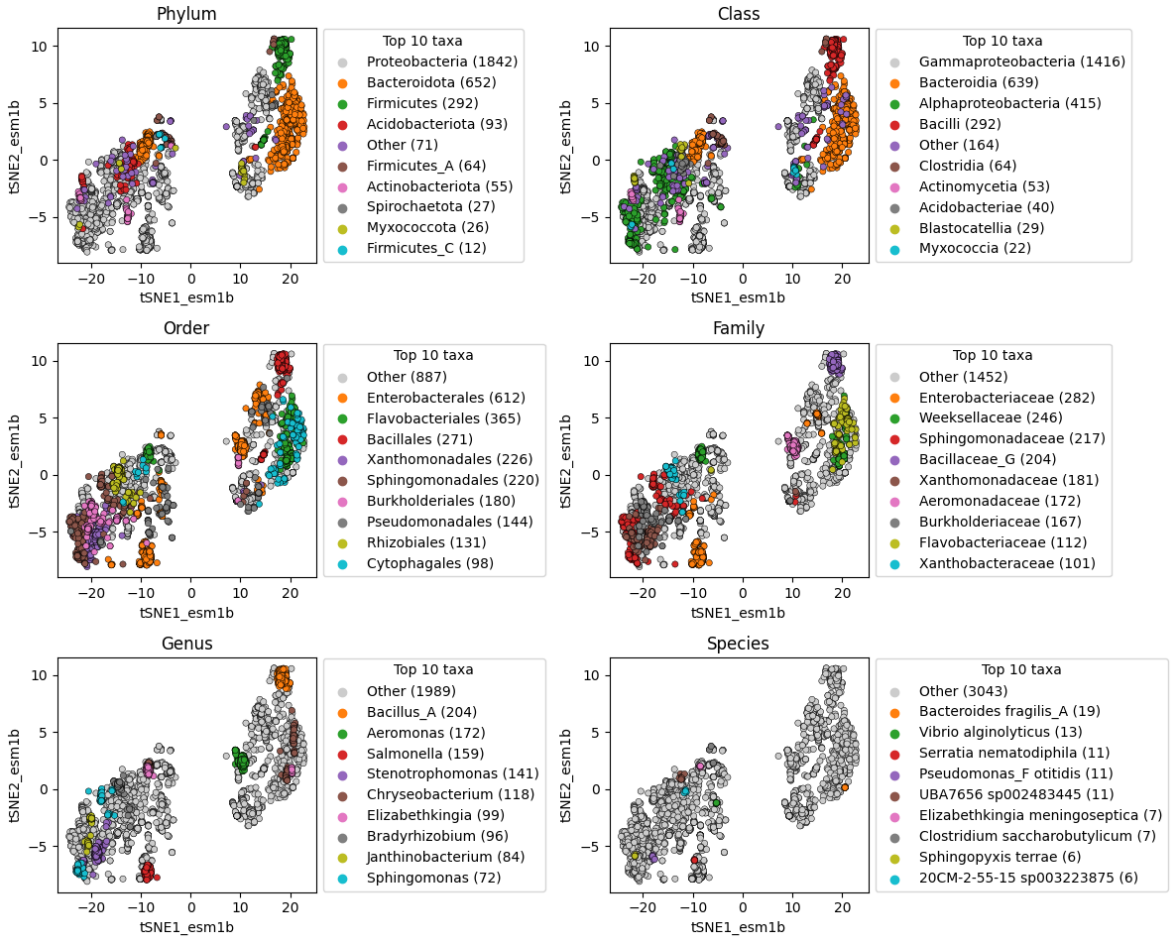


Figura 11. Mapeo taxonómico de las secuencias de metalobetalactamasas en las representaciones de tSNE de ESM-1b. Se indica la categoría taxonómica en título de cada panel y en su respectiva leyenda se enlista en orden decreciente los nueve taxa más abundantes, señalando entre paréntesis el número de secuencias asociadas. Todos los taxa que no se encuentran dentro de los nueve taxa más abundantes se indican como "Otras".

Para evaluar que otras propiedades se asocian a la organización de las metalobetalactamasas, mapeé distintas propiedades fisicoquímicas/estadísticas y computé la correlación de Spearman que existe entre sus valores y las dimensiones de tSNE (Fig. 12). Es importante notar que, dado que las representaciones de tSNE son estocásticas y dependientes de la configuración de los parámetros usados, los valores de correlación son difíciles de reproducir al usar otras configuraciones. La longitud, masa molecular, inestabilidad y fracción de residuos propensos a formar regiones hélice y beta plegada son las que tienen las correlaciones más altas ($\rho > 0.50$ con alguna de las dimensiones de tSNE). De hecho, la longitud y masa molecular son redundantes entre sí al tener una alta correlación ($\rho = 0.94$; Fig. sup. 10). La subclase B3 tiene secuencias más grandes (promedio, desviación estándar = 297.48 ± 19.24 residuos; n = 1,693), mientras que las subclases B1 (252.54 ± 12.20 residuos; n = 1,205) y B2 (247.17 ± 17.99 residuos; n = 201) son más cortas (Fig. sup. 11). La subclase B3 tiene una menor fracción de aminoácidos propensos a formar regiones hélice (0.29% ± 0.03%) respecto a la subclase B1 (0.33% ± 0.02%) y B2 (0.33% ± 0.01%). Por el contrario, la subclase B3 tiene una mayor

fracción de aminoácidos propensos a formar regiones beta plegada (0.28% ± 0.04%) respecto a la subclase B1 (0.24% ± 0.04%) y B2 (0.25% ± 0.02%). La subclase B3 tiene secuencias que se estiman ser más inestables (33.96 ± 7.86) respecto a las subclases B1 (27.24 ± 7.21) y B2 (24.74 ± 4.46), siendo particularmente inestables las secuencias de la familia Sphingobacteriaceae así como de la familia enzimática SER de *Salmonella* (Fig. 9B y 11). El resto de las propiedades muestra una correlación baja o nula que no reflejan un claro patrón de organización en las metalobetalactamasas.

Tomando los resultados en conjunto sugiero que las características estructurales de los grupos ME y MB son la principal propiedad que separa a las metalobetalactamasas. Por otro lado, la taxonomía parece no tener una influencia en la organización de las metalobetalactamasas. Y finalmente, solo algunas propiedades fisicoquímicas como el número de residuos, inestabilidad y fracción de aminoácidos propensos a formar hélices y beta plegada podrían generar cierta influencia en la organización de las representaciones de tSNE de ESM-1b de las metalobetalactamasas.

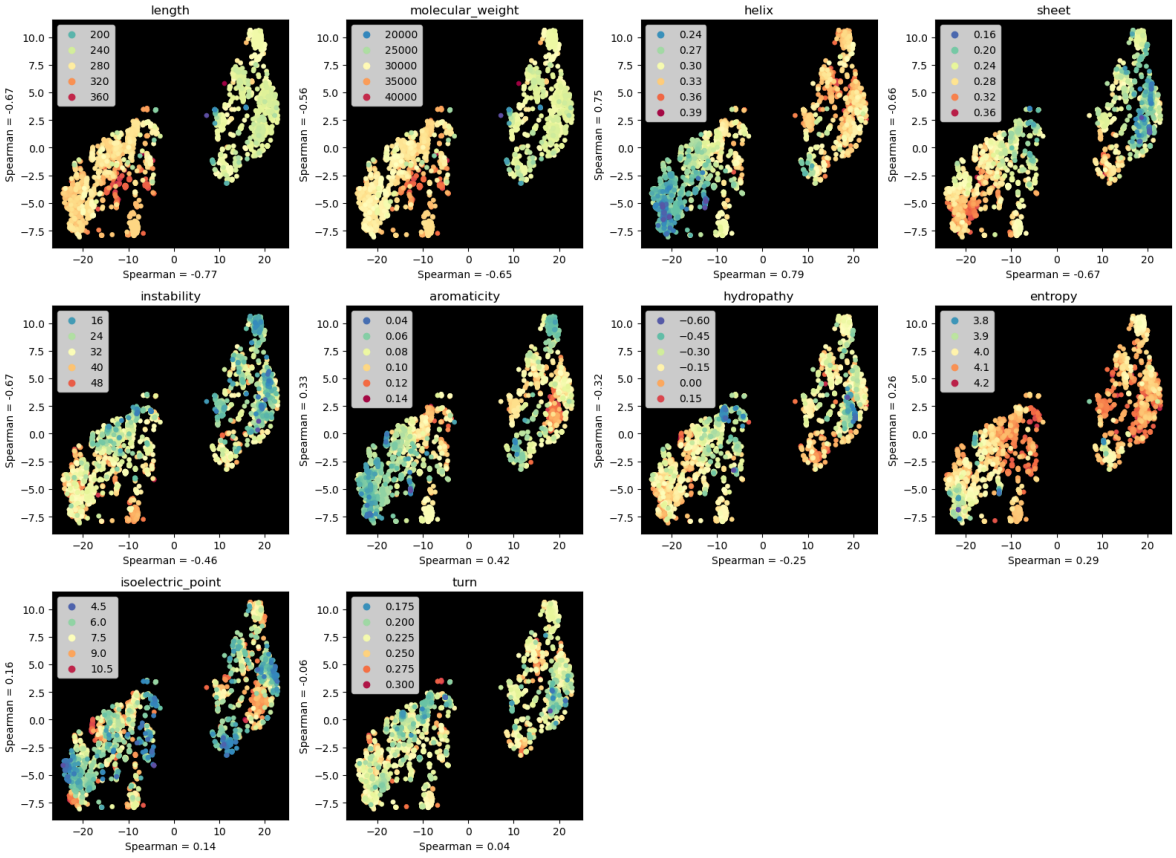


Figura 12. Propiedades cuantitativas de las metalobetalactamasas. Se indica en el encabezado de cada panel la propiedad correspondiente y con distintos marcadores a cada una de las subclases de metalobetalactamasas. Se omite el nombre de las dos dimensiones de tSNE y en su lugar se muestran los valores de correlación de Spearman que tiene dicho eje respecto a los valores de la propiedad en cuestión. Por ejemplo, la longitud de secuencia tiene una correlación de ρ = -0.76 con la primer dimensión de tSNE (tSNE1_esm1b) y de ρ = -0.66 con la segunda dimensión (tSNE2_esm1b). Para una mejor visualización, se removieron los 31 coordenadas correspondientes a secuencias cuya longitud no cae dentro de un rango de ±30% del valor de la mediana de cada subclase. Esto ayuda a remover valores extremos que sesgan el rango de visualización. Para más detalles sobre las propiedades consultar la Tabla 2 y las Figuras suplementarias 10 y 11.

Al estimar las métricas de distancia de las 100 secuencias muestreadas al azar por cada subclase, observé que los *embeddings* de ESM-1b pueden distinguir a los grupos ME y MB, sin embargo, la subclase B1 no es fácilmente distinguible respecto a la subclase B2. (Fig. 13). Como comparación, los *embeddings* del modelo Bepler no pueden distinguir a los grupos ME y MB, lo cual también se ve parcialmente reflejado en el análisis de PCA con este modelo (Fig. 6).

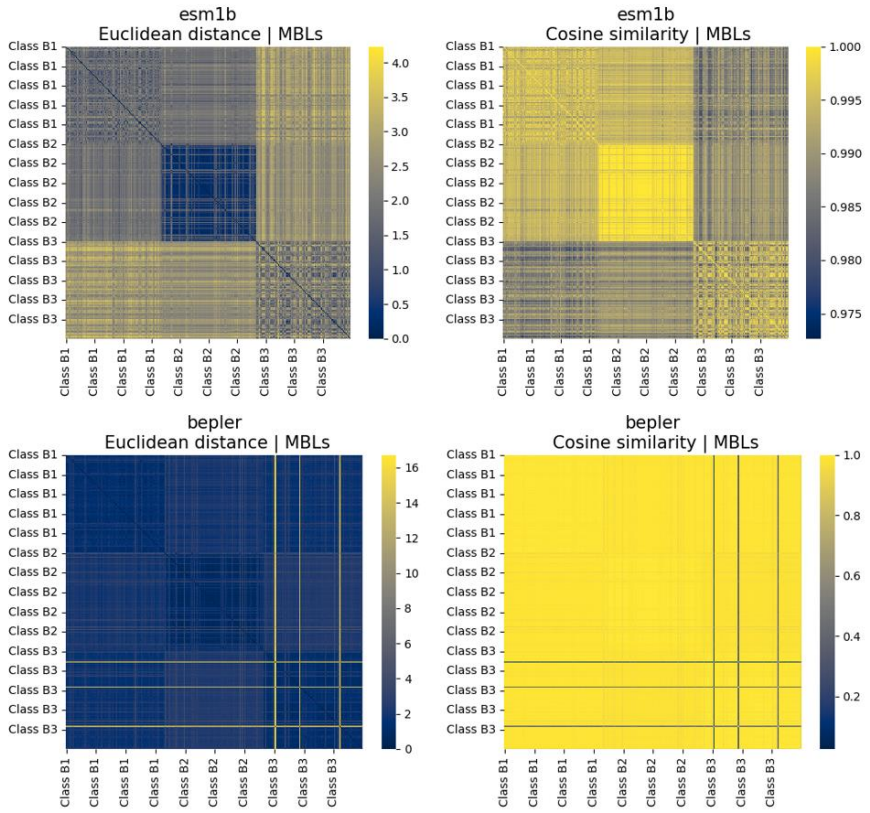


Figura 13. Identificación de grupos de metalobetalactamasas usando métricas de distancia. Se indica en el encabezado de cada panel la métrica de distancia y el modelo de lenguaje de proteínas. La distancia Euclidiana toma valores de cero a infinito positivo, siendo valores cercanos a cero indicativos de una alta similitud. La similitud coseno toma valores entre 0 y 1, siendo valores cercanos a 1 indicativos de alta similitud. Para consultar el resto de las distancias por pares consultar el *jupyter notebook* "*Embedding_distance*" en el repositorio de esta tesis.

Posteriormente, evalúe la capacidad de identificación de grupos de *kmeans*. Este algoritmo usa la distancia Euclidiana como métrica de similitud junto con otras estrategias para identificar grupos de forma no supervisada y usa el parámetro k para indicar el número de grupos que deben encontrarse. Cuando k = 2, observé que *kmeans* distingue correctamente a los grupos ME y MB usando los *embeddings* de ESM-1b (ARI = 0.98), mientras que con el modelo Bepler no es capaz de distinguirlos (ARI = 0.02; Fig. 14). Cuando k = 3, las tres subclases de metalobetalactamasas son distinguibles de manera moderada mediante los *embeddings* de ESM-1b (ARI = 0.67). Por el contrario, usando el modelo Bepler solo se distinguen la subclase B2 del resto (ARI = 0.47).

Los resultados con las métricas de distancia y *kmeans* sugieren que la información presente en los *embeddings* de ESM-1b es suficiente para que algoritmos simples como estos puedan detectar grupos de proteínas como las subclases de metalobetalactamasas. Tomando los resultados en conjunto, sugiero que, independientemente de si se usa un algoritmo basado en reducción de la dimensionalidad o no, los *embeddings* del modelo ESM-1b cuentan con la suficiente información para distinguir los grupos de metalobetalactamasas.

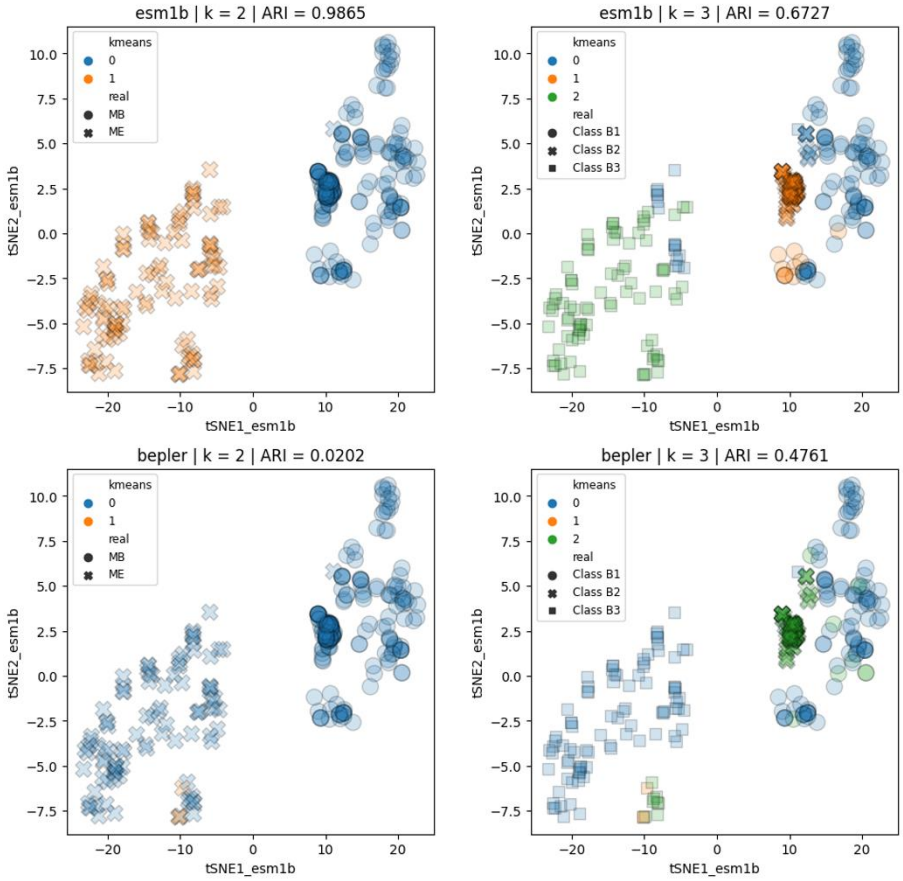


Figura 14. Organización de las metalobetalactamasas usando *kmeans*. Se indica en el encabezado de cada panel el modelo de lenguaje de proteínas, el valor de k de *kmeans* y el valor de *adjusted rand index* (ARI) obtenido al comparar los grupos de kmeans contra las clases reales. Se indica en la leyenda los grupos inferidos por *kmeans* y con distintos marcadores a los grupos de metalobetalactamasas (real). Para visualizar los grupos identificados por *kmeans*, se mapearon los resultados a la representación de tSNE de ESM-1b, sin embargo, la inferencia con kmeans se realizó con los *embeddings* de alta dimensión y no con las representaciones de baja dimensión de tSNE. Para este análisis se usaron los valores por defecto de los parámetros de kmeans implementados en la librería Scikit-Learn v1.3.2.

SERINBETALACTAMASAS

A diferencia de las metalobetalactamasas, el análisis con PCA de las serinbetalactamasas sugiere que ningún PLM genera buenos resultados, pues no logran separar a las tres clases (Fig. 15).

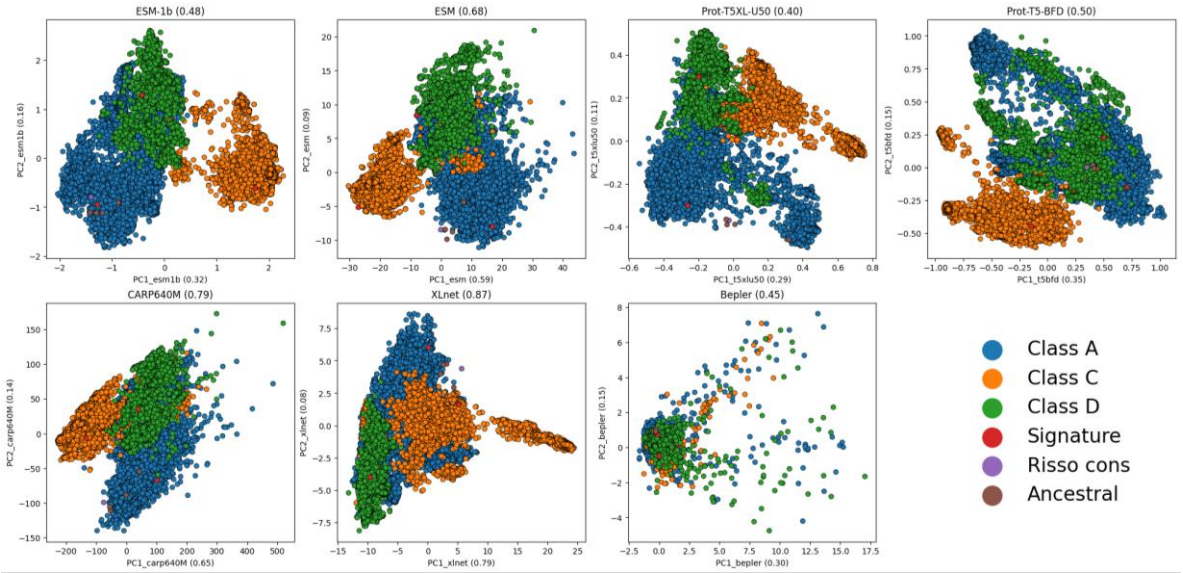


Figura 15. Análisis de componentes principales con serinbetalactamasas. Se indica en el encabezado el modelo de lenguaje de proteínas y entre paréntesis la varianza total representada por los componentes. En cada eje se muestra entre paréntesis la varianza representada por el componente.

Los valores de divergencia de Kullback-Leibler sugieren que todos los PLM fueron representados de forma precisa en la baja dimensión, siguiendo el mismo patrón de las metalobetalactamasas al obtener menores divergencias conforme aumenta el valor de *perplexity* (Fig. sup. 2). Con tSNE la mayoría de los PLM separa las tres clases de serinbetalactamasas excepto por los modelos Bepler, XLNet y CARP que mapean en un mismo espacio a las clases A y D (Fig. 16), lo cual podría deberse a la estrecha relación evolutiva entre estas clases que son consideradas como "hermanas"[13]. En contraste, los modelos ESM-1b, ESM, Prot-T5XL-U50 y Prot-T5-BFD son capaces de distinguir las tres clases de serinbetalactamasas, sin embargo, dividen a las clases A y C en varios grupos. Por ejemplo, ESM-1b divide a la clase A en tres grupos y la clase C en dos grupos.

Al igual que tSNE, con UMAP la mayoría de los PLM separan a las tres clases de serinbetalactamasas excepto por los modelos Bepler, XLNet y CARP que mapean en un mismo espacio a las clases A y D (Fig. 17). Igualmente, los modelos ESM-1b, ESM, Prot-T5XL-U50 y Prot-T5-BFD separan a las tres clases de serinbetalactamasas y dividen estas clases en varios grupos. Por ejemplo, ESM-1b divide a la clase A en tres grupos.

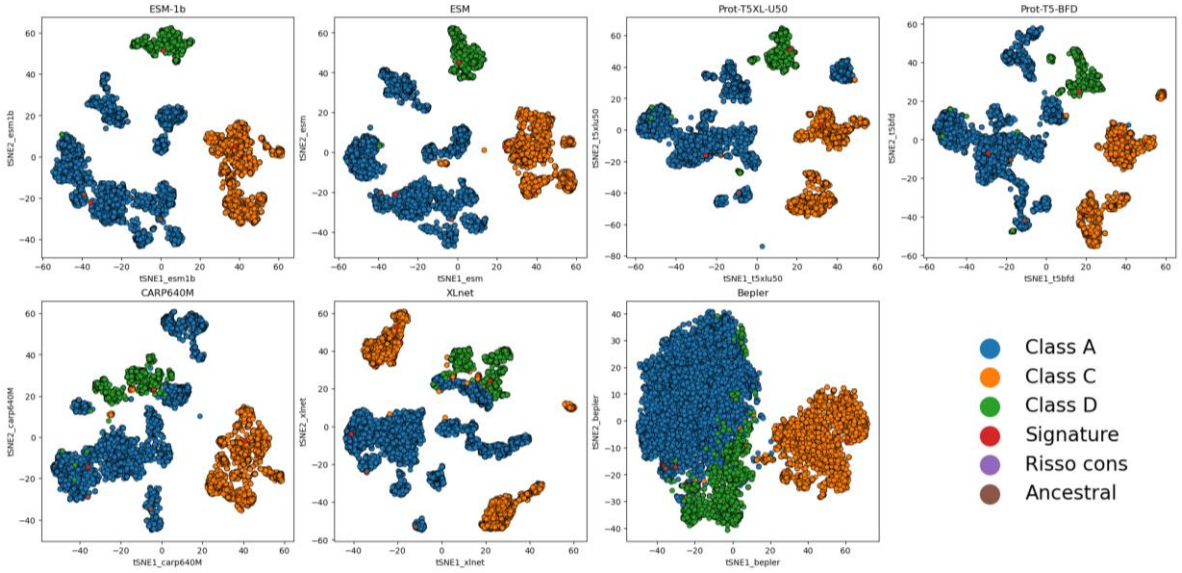


Figura 16. Análisis de reducción de la dimensionalidad con tSNE con serinbetalactamasas. Se indica en cada encabezado el modelo de lenguaje de proteínas correspondiente. Parámetros usados: perplexity = 400; iteraciones = 1500.

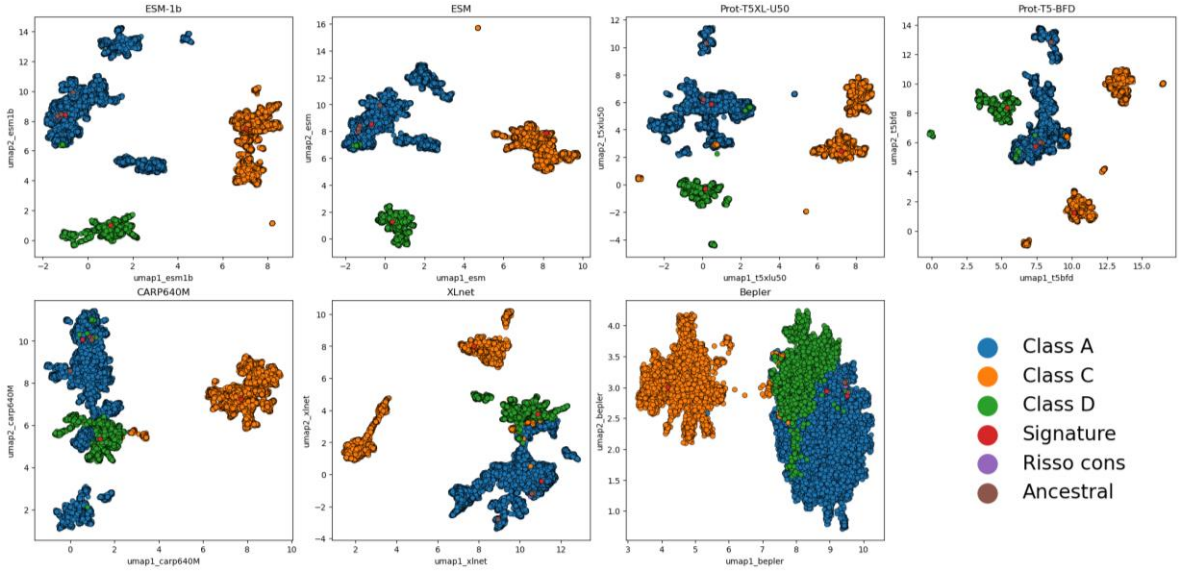


Figura 17. Análisis de reducción de la dimensionalidad con UMAP con serinbetalactamasas. Se indica en cada encabezado el modelo de lenguaje de proteínas correspondiente. Parámetros usados: n_neighbors = 400; mindist y spread = 0.2.

Tomando los resultados en conjunto sugiero que tSNE y UMAP generan mejores representaciones respecto a PCA. Con PCA ninguno de los PLM separa a las tres clases de serinbetalactamasas, mientras que con tSNE y UMAP solo los modelos CARP, XLNet y Bepler no separan las clases A y D, posiblemente, debido a su estrecha relación evolutiva. En contraste los modelos ESM-1b, ESM, Prot-T5XL-U50 y Prot-T5-BFD si distinguen dichos grupos, aunque los dividen en varios grupos.

Al igual que con las metalobetalactamasas, decidí centrarme en la representación de tSNE de ESM-1b de las serinbetalactamasas dado que conozco la fidelidad de sus representaciones y dado que logró distinguir a las clases A, C y D. Sin embargo, el modelo Prot-T5XL-U50 parece igualmente interesante al explorar detalles de sus grupos y sugiero que estos dos PLM pueden ser de utilidad para hacer análisis con otros grupos de proteínas.

En la representación de tSNE de ESM-1b la clase A se separó en tres grandes grupos, pese a que solo se reconocen dos subclases[15,23] (Fig. 18A). La clase D se separó como un solo grupo. La clase C se separó en dos grupos, lo cual podría tener relación con la reciente propuesta de que esta clase puede dividirse en dos subclases. Philippon *et al.*[25] sugieren que la putativa subclase C2 está representada por betalactamasas filogenéticamente distantes al resto y que se asocian a géneros como *Legionella*, *Bradyrhizobium* y *Parachlamydia*. Curiosamente, estos géneros y otros más como *Acinetobacter*, *Chryseobacterium* y *Sediminibacterium* se agrupan en uno de los dos grupos formados. Lo anterior sugiere que podríamos considerar a este vecindario de secuencias para hacer análisis detallados a nivel de secuencia y estructura para valorar la existencia de una subclase C2 (Fig. sup. 12).

Al mapear las familias enzimáticas observé buenas agrupaciones de sus variantes (Fig. 18B-D). Esto es notable, pues las variantes de una familia pueden tener diferencias de identidad de secuencia considerables. Por ejemplo, CTX-M-14 y CTX-M-15 de la clase A tienen 83% de identidad de secuencia[163]. Del total de 13,314 secuencias clase A, solo 34.5% (4,594 secuencias) pertenecen a una familia reconocida, y dentro de este grupo, las 10 familias más abundantes representan un 65.2% (2,998 secuencias). Del total de 6,586 secuencias clase C, 65.8% (4,331 secuencias) pertenecen a una familia reconocida, y dentro de este grupo, las 10 familias más abundantes representan un 92% (3,985 secuencias). Notablemente, las variantes de la familia EC (2,282 secuencias) representan un 52.69% del total de familias reconocidas de la clase C. Del total de 2,779 secuencias clase D, 96.44% (2,680 secuencias) pertenecen a una familia reconocida, y dentro de este grupo, la familia OXA representa el 95.15% (2,550 secuencias). Lo anterior sugiere que las familias reconocidas en la clase A representan poca diversidad del total de secuencias. Mientras que para la clase C las familias reconocidas representan mediamente la diversidad de secuencias de toda la clase. La familia EC representa una gran cantidad de estas secuencias, por lo que valdría la pena estudiarlas para evaluar si es pertinente dividir esta familia en subgrupos como ocurre con la familia OXA de la clase D.

La familia OXA se suele clasificar en subfamilias porque es muy heterogénea[24]. Los miembros dentro de una subfamilia suelen compartir >60% de similitud de secuencia, mientras que entre subfamilias los miembros pueden mostrar una similitud de apenas 16.4%[24]. Para ver como se refleja esto en la representación de ESM-1b, mapeé las variantes asociadas a 63 subfamilias registradas en la BLDB. Observé una buena agrupación de las 18 subfamilias más abundantes (Fig. sup. 13), lo cual sugiere que ESM-1b captura este nivel de organización en sus *embeddings*. Tomando los resultados en conjunto, sugiero que ESM-1b tiene una buena capacidad de distinguir la organización de las serinbetalactamasas a nivel de clase, familia y subfamilia.

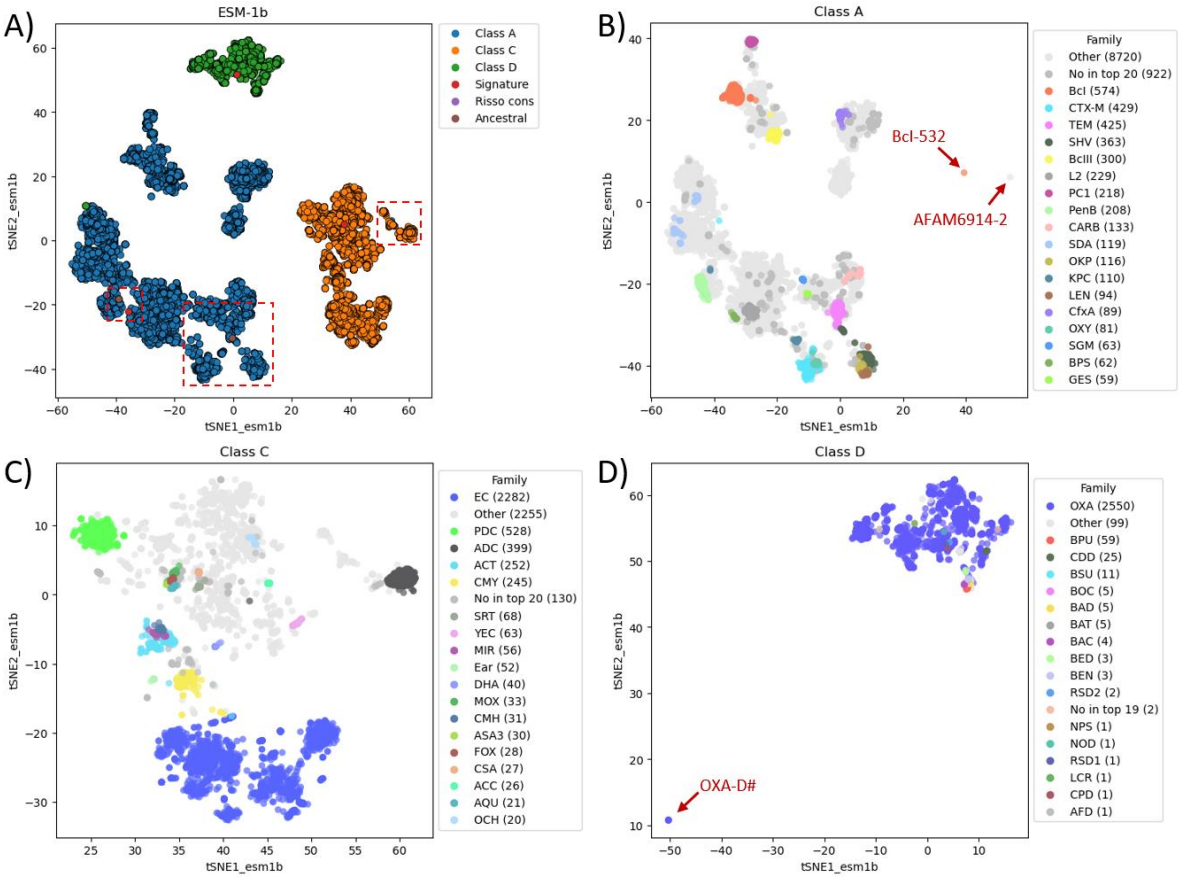


Figura 18. Organización de las serinbetalactamasas y sus familias enzimáticas. (A) Organización de las serinbetalactamasas. Se señala con recuadros rojos a las regiones donde se hizo un acercamiento en la clase A (Fig. 19) y la clase C (Fig. sup. 12). (B) Mapeo de las familias clase A. (B) Mapeo de las familias clase C. (D) Mapeo de las familias clase D. Se indica con flechas la ubicación de las secuencias anómalas y en paréntesis el número de secuencias de cada familia. Se señala con colores a las 18 familias más abundantes, en gris claro las familias identificadas como "Otras", y en gris oscuro las secuencias que no se encuentran dentro de las 18 familias más abundantes.

Unas pocas secuencias anómalas no se agruparon donde se esperaba. En la clase D no se agruparon siete secuencias de familia la OXA-D. Estas secuencias tienen entre 296 a 302 residuos y fueron depositadas por Pares *et al.*[164], quienes cristalizaron el dominio transpeptidasa de la PBP2x. La PBP2x es una proteína de unión a membrana con 675 residuos, pero Pares *et al.* removieron la parte N-terminal que permite anclarse a membrana. Los autores reportan que la estructura de la PBP2x es similar a la clase A, razón por la cual se habría agrupado con dicha clase (Fig. 18A). Sin embargo, estas secuencias no son serinbetalactamasas y son un segundo error presente en la BLDB.

La secuencia anómala BcI-532 clase A que se agrupó con la clase C. Su encabezado indica que es un canal de sodio, lo cual representa un tercer error en la BLDB (Fig. sup. 9B). BcI-532 pudo agruparse en con la clase C debido a que sus secuencias son más grandes respecto al resto de betalactamasas. Sin embargo, al tratarse de un plegamiento distinto, lo ideal sería que no se ubicara en ninguna región asociada a las serinbetalactamasas. Aunque esto sí se ve reflejado a nivel de clase, pues se distingue como un punto extremo en la clase A (Fig. 18B)

La secuencia anómala AFAM6914-2 de la clase A se agrupó con la clase C. Esta secuencia tiene 757 residuos y no tiene información en su encabezado. Debido a esto, predije su estructura con ESMFold y observé una betalactamasa dimérica. LRA-13 fue la primer betalactamasa dimérica caracterizada experimentalmente[158]. Su dominio N-terminal alinea con la clase D y el C-terminal con la clase C, y ambos presentan actividad catalítica. Para saber si AFAM6914-2 podría ser similar a LRA-13, tomé dos estructuras cristalográficas que estuvieran en la región entre las clases A (PDB: 3W4Q) y C (PDB: 6PWL) y alineé sus estructuras con ChimeraX[165] (Fig. sup. 9C-D). La estructura clase A alineó con el dominio C-terminal de AFAM6914-2 y la estructura clase C con el N-terminal. Similar al caso de LRA-13. La identificación de una betalactamasa dimérica es un caso complicado para ESM-1b y cualquier otro algoritmo. Sin embargo, AFAM6914-2 podría haberse agrupado con la clase C porque el *embedding* tendría más información de esta clase al tener más residuos respecto a la clase A. Curiosamente, otras nueve betalactamasas diméricas[166] también fueron incluidas en la BLDB y todas ellas se ubican en la clase C, posiblemente, por la mayor cantidad de información contenida en el *embedding* de esta clase (Fig. sup. 14).

Además de las betalactamasas diméricas, encontré casos de betalactamasas fusionadas con otros plegamientos (Fig. sup. 9E). Estos casos son secuencias clase A con un alto número de residuos y no son fáciles de resolver para ESM-1b. Sería deseable que, al contener información de plegamientos distintos a betalactamasas, ESM-1b fuera suficientemente sensible para detectarlos como puntos extremos. Sin embargo, existe información de betalactamasas en los *embeddings* que hace que se agrupen en la clase A. Estos casos ilustran una posible falta de sensibilidad por parte de ESM-1b.

Todas las secuencias firma se agruparon en sus regiones esperadas (Fig. 18A). Aunque existe una ligera diferencia entre la ubicación de mi secuencia firma respecto al consenso estimado por Risso *et al*.[113]. Esto puede deberse al enfoque de pesos que usé para su estimación y el número de secuencias usadas entre ambos métodos (yo = 5,449 secuencias; Risso *et al*. = 75). Pese a ello, ambas secuencias se agrupan en una vecindad considerable (Fig. 19B). La betalactamasa clase A que representa al ancestro de las Enterobacterias se agrupó con la familia TEM (Fig. 19A). Lo cual parece consistente pues las variantes de esta familia suelen aislarse de enterobacterias[167]. Además, parece tener relación la propuesta del origen de la familia TEM a partir enterobacterias hace ≈700 millones de años[8,15,23,114].

Cerca del ancestro de las Enterobacterias también se encuentran unas pocas secuencias parciales de la familia SHV (Fig. sup. 9F). La ubicación de estas secuencias puede valorarse de dos formas. En la primera, ESM-1b no es suficientemente sensible para agruparlas con el resto de las secuencias completas. En la segunda, ESM-1b muestra ser robusto porque sus *embeddings* tienen información suficiente para agruparlas en la vecindad de la familia SHV y no en cualquier otra región de la clase A. Por ejemplo, las variantes SHV tienen una longitud promedio de 281 ± 18.47 residuos, mientras que SHV-P3 tiene 103 residuos. Y aunque SHV-P3 tiene menos del 50% de la longitud promedio de su familia, y tampoco tiene todos los residuos catalíticos, se agrupa en la vecindad de la familia SHV.

Los otros ancestros se agruparon en un estrecho espacio junto con el consenso de Risso *et al.*[113] (Fig. 19B). Esto podría deberse a que los tres ancestros corresponden a taxa basales que divergieron hace más de 2,000 millones de años[8,114]. Además, parece soportar la propuesta de Risso *et al.* de que una

secuencia consenso es una buena aproximación de una secuencia ancestral[113,168]. Para saber si esta estrecha agrupación se debe a una alta similitud de secuencia, estimé el porcentaje de identidad entre los ancestros y dos secuencias cercanas a esta región (Fig. sup. 15). Observé que el consenso de Risso *et al.* tiene una identidad estimada de ≤79% con el resto de las secuencias, sugiriendo que la similitud de secuencia no es clave en su agrupación a nivel local. Curiosamente, la mutante GNCA TEM-1 like no se agrupó con la familia TEM. Esta mutante porta 21 sustituciones respecto al ancestro GNCA que hacen que su capacidad catalítica contra penicilinas sea similar a la de la TEM-1, la cual es considerada una especialista en penicilinas[167]. Esto sugiere que la organización de las serinbetalactamasas no estaría muy influenciada por la actividad catalítica.

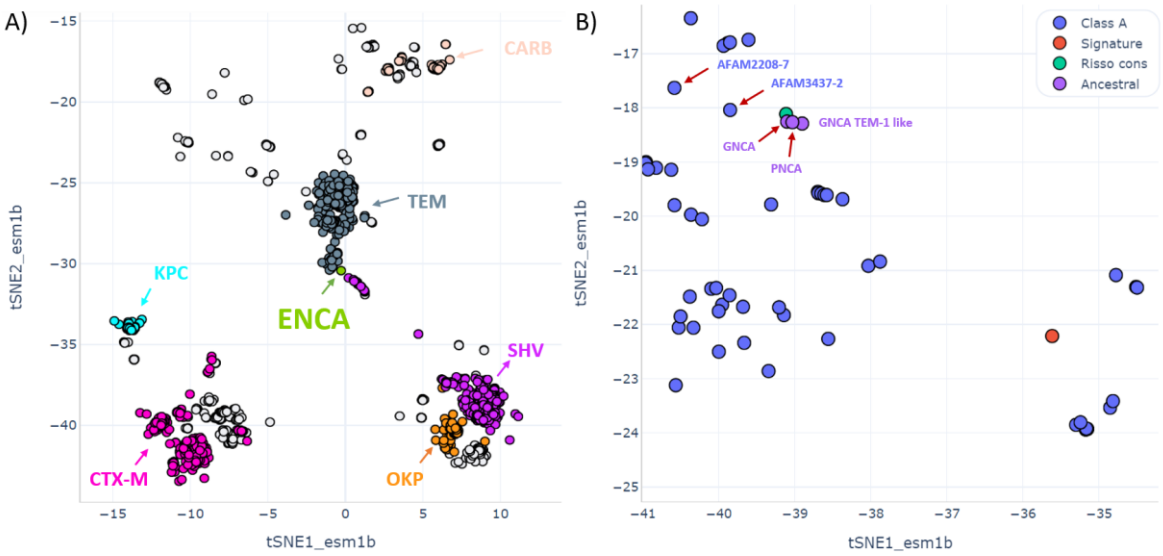


Figura 19. Ubicación del consenso y ancestros estimados por Risso. et. al. (A) Ubicación del ancestro de Enterobacterias (ENCA). Para una mejor visualización no se muestran las secuencias de familias enzimáticas etiquetadas como "Otras" y solo se muestran unas familias de referencia indicadas con colores. (B) Ubicación del consenso y los ancestros de bacterias Gram Positivas y negativas (PNCA), Gram negativas (GNCA) y su mutante que asemeja la capacidad catalítica de la TEM-1 (GNCA TEM1-like). Se señalan las dos secuencias que fueron usadas en la comparación de la identidad de secuencia (Fig. sup. 15).

Al mapear la taxonomía de las serinbetalactamasas observé que no existe una clara organización en función de las categorías taxonómicas y solo es posible detectar grupos relativamente claros a nivel de especie (Fig. 20). La Clase C está representada en un 98.88% por el filo Proteobacteria (6,512 secuencias). Notablemente, la familia Enterobacteriaceae representa el 51.61% (3,399 secuencias) del total de las secuencias. Igualmente, de todas las secuencias clase C, el género *Escherichia* representa el 34.65% (2,282 secuencias). Curiosamente, las secuencias de *Escherichia* corresponden a las de la familia enzimática EC (Fig. 18C). Lo anterior sugiere que la clase C tiene poca diversidad filogenética. La clase D está representada en un 74.34% por el filo Proteobacteria (2,066 secuencias), 7.63% por Campylobacteriota (212 secuencias) y 6.26% por Firmicutes (174 secuencias). Existen también unos géneros patógenos relativamente abundantes como *Acinetobacter* que representa un 26.08% (725 secuencias) del total de secuencias clase D, o *Burkholderia* que representa un 9.71%

(270 secuencias). De hecho, existe una buena correspondencia entre secuencias clasificadas dentro de la subfamilia OXA-51-like y las clasificadas como *A. baumannii*, como previamente se ha observado[6,24], lo cual sugiere una buena consistencia de las anotaciones taxonómicas (Fig. sup. 13A). La clase A tiene una mayor diversidad taxonómica respecto a las clases C y D que están principalmente representadas por el filo Proteobacteria. Del total de secuencias clase A, este filo representa un 46.48% (6,188 secuencias), seguido por Actinobacteriota, Firmicutes y Bacteroidota que representan el 19.81% (2,638 secuencias), 8.38% (1,116 secuencias) y 8.16% (1,087 secuencias), respectivamente.

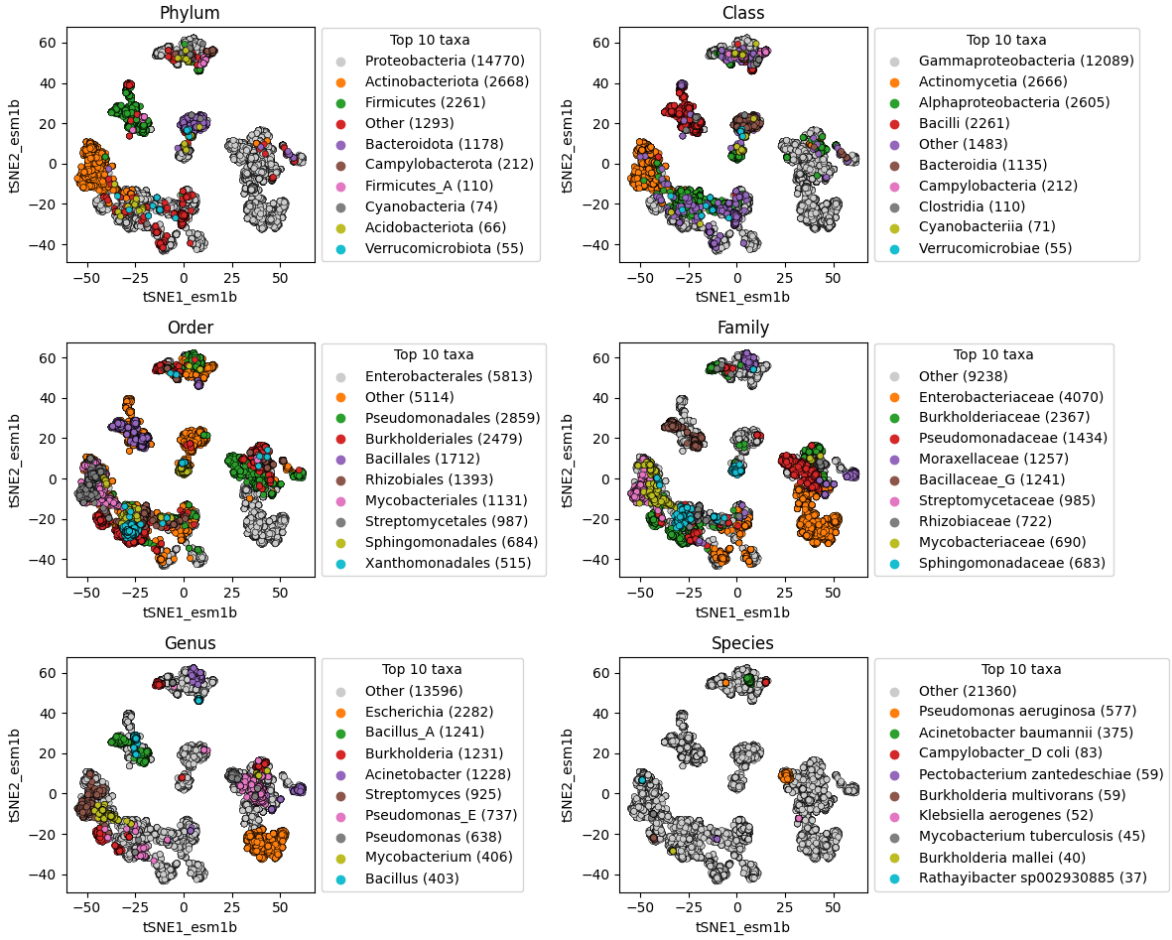


Figura 20. Mapeo taxonómico de las secuencias de serinbetalactamasas en las representaciones de tSNE de ESM-1b. Se indica la categoría taxonómica en título de cada panel y en su respectiva leyenda se enlista en orden decreciente los nueve taxa más abundantes, señalando entre paréntesis el número de secuencias asociadas. Todos los taxa que no se encuentran dentro de los nueve taxa más abundantes se indican como "Otras".

La clase A se separó en tres grupos distribuidos en varios filos, lo cual no es consistente con los dos grupos que podríamos esperar correspondieran a las subclases A1 y A2 (Fig. 20). El primer grupo (A01) tiene 8,986 secuencias principalmente de Proteobacteria y Actinobacteriota. El segundo grupo (A1F) tiene 2,344 secuencias principalmente de Firmicutes. El tercer grupo (A02) tiene 1,982

secuencias principalmente de Bacteroidota y Proteobacteria. Para comprender mejor esta organización, mapeé 83 familias enzimáticas representativas de los seis grupos filogenéticos propuestos por Philippon *et al*.[15]. En general, observé un mapeo consistente de las familias con la región de la clase taxonómica a la que pertenecen (Fig. 21A). Las familias del grupo A mapean a Bacteroidia. Las familias del grupo B mapean a Alphaproteobacteria. Las familias del grupo C mapean a Gammaproteobacteria, sin embargo, se encuentran como subgrupos asociados a la familias enzimáticas CARB, TEM, SHV/OKP/LEN y KLUY. Dichos subgrupos también han sido identificados con filogenias y de redes de similitud de secuencia[46]. Las familias del grupo D, que es conocido por ser muy diverso, mapean a Actinomycetia y Bacilli. Esto es interesante, pues estas dos clases de bacterias Gram positivas están en regiones separadas en la representación de tSNE. Las familias del grupo E, que es conocido por ser muy diverso, mapean de forma dispersa en Gammaproteobacteria, asociándose a las regiones de los órdenes Burkholderiales y Enterobacteriales. Las familias del grupo F mapean a Alpha y Gammaproteobacteria. Más específicamente, a las regiones de los órdenes Xanthomonadales, Sphingomonadales, Rhizobiales y Burkholderiales.

Curiosamente, un conjunto secuencias de la clase Alphaproteobacteria (A2P) del grupo A02 no tuvo ningún registro contra alguna familia representativa. Para tener más información de este grupo, mapeé 562 especies de bacterias (Fig. 21B) distribuidas en 285 géneros (Fig. 21C) propuestos por Philippon *et al*.[15] como posibles miembros de los seis grupos filogenéticos. Aunque los miembros de estos dos esquemas son putativos, permiten aproximar la identidad de los grupos donde las familias representativas no ofrecen mucha información. En general, observé que el mapeo con las familias representativas es consistente con las especies y géneros putativos, pues los tres esquemas muestran patrones de distribución muy similares. Particularmente, el grupo A2P cuenta con especies del grupo A y F y géneros de los grupos A, B y F principalmente asociados a la familia Sphingomonadaceae, el cual se compone por bacterias de vida libre. Esto es curioso, pues también existen secuencias de Sphingomonadaceae en el grupo A01, sin embargo, ambos grupos no se agruparon en un lugar en común. Similar al caso de los filos de bacterias Gram positivas. Tomando los resultados en conjunto, sugiero que la organización de la representación de tSNE de las serinbetalactamasas clase A no es consistente con el esquema de seis grupos filogenéticos de Philippon *et al.*[15], pues los seis grupos no se agrupan de forma discreta a regiones específicas.

Para conocer la diversidad de secuencias y estructuras de la clase A en relación con los tres grupos de la representación de tSNE (Fig. 21A), tomé al azar 20 secuencias por cada grupo, predije sus estructuras con AlphaFold2[67,162] (Fig. sup. 16) y calculé una matriz de similitud por pares con sus valores de identidad de secuencia y RMSD estimados con TM-align[169]. Al considerar la identidad de secuencia observé dos grupos (Fig. 22A), uno contiene al grupo A02 y el otro a los grupos A01 y A1F (A01-A1F). Las secuencias del grupo A01-A1F comparten una identidad ≥30%. Particularmente, las secuencias del grupo A1F comparten una identidad ≥60%. tSNE identifica grupos en función de fuerzas atractivas y repulsivas entre pares de datos[143]. Debido a que las secuencias del grupo A1F comparten una identidad ≥60%, es posible que se haya generado una mayor fuerza atractiva que las separó en un grupo distinto a A01. El grupo A01-A1F comparte una identidad de secuencia ≤30% respecto al grupo A02. Lo anterior parece ser consistente con la división de la clase A en dos subclases.

El grupo A02 se habría separado del resto en la representación de tSNE debido a las fuerzas repulsivas generadas por la diferencia a nivel de secuencia. Al considerar la similitud estructural observé los mismos dos grupos formados mediante identidad de secuencia (Fig. 22B). Sin embargo, los grupos A1 y A1F no se separan tan claro como lo hicieron previamente. Las estructuras del grupo A01-A1F muestran valores de RMSD ≤2. Particularmente, las estructuras del grupo A1F muestran valores de RMSD ≤1.5 entre sí, lo cual sugiere una similitud estructural. El grupo A01-A1F tiene valores de RMSD ≥2 respecto al grupo A02, sugiriendo una diferencia estructural entre los grupos.

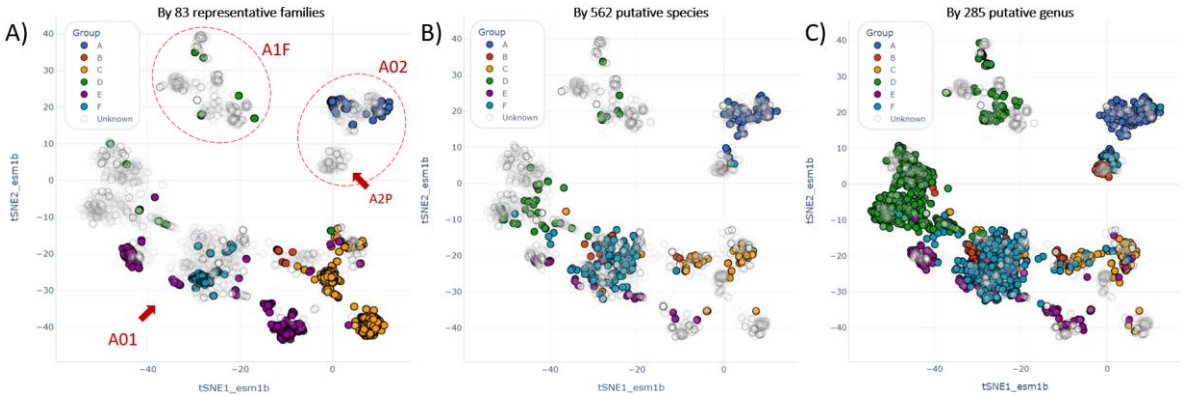


Figura 21. Organización de las serinbetalactamasas clase A de acuerdo con Philippon *et al.*[15]. (A) Familias representativas. Se mapearon 11 familias del Grupo A (CblA, CepA, CfxA, CGA, CIA, CME, LUS, PER, TLA, TLA2, VEB), 4 familias del Grupo B (BEL, GES, PME, SGM), 21 familias del Grupo C (AER, CARB, CKO, FPH, GIL, HMS, KLUY, LAP, LEN, MAL, MP, OHIO, OKP, ORN, PAL, PLES, RUB, SCO, SHV, TER, TEM), 12 familias del Grupo D (ACI, ARL, AST, BCL, BlaS, CAD, CBP, FAR, MAB, OIH, R39, ROB), 27 familias del Grupo E (BES, BIC, BPS, CRH, CRP, CTX-M, DES, ERP, FONA, FRI, HugA, IMI, KLUA, KLUC, KLUG, KPC, LUT, MIN, OXY, PenA, PenB, RAHN, RIC, SED, SFC, SME, SFO) y 8 familias del Grupo F (AXC, BKC, BOR, CzoA, GPC, L2, PAD, XCC). El grupo A corresponde a las familias de la subclase A2. Se indican con flechas y círculos punteados al grupo principal (A01), el de Firmicutes (A1F), el asociado a la subclase A2 (A02) y al de la clase Alphaproteobacteria donde ninguna familia representativa tuvo registro (A2P). (B) Especies putativas. Número de especies por grupo: A = 106, B = 32, C = 90, D = 116, E = 55, F = 163. (C) Géneros putativos. Número de géneros por grupo: A = 61, B = 20, C = 47, D = 54, E = 25, F = 78. Aquellas secuencias que no tuvieron registro con alguno de los grupos fueron etiquetadas como "Desconocido". Para una mejor visualización, se excluyeron las dos secuencias anómalas (AFAM6914-2 y BcI-532).

Curiosamente, las secuencias AFAM7362-6 y AFAM7372-1 del grupo A02 muestran una identidad de secuencia de ≈30%, y un RMSD de ≈2, respecto a los grupos A02 y A01-A1F. Estas dos secuencias pertenecen a Sphingomonadaceae dentro del grupo A2P y sus estructuras cuentan con regiones donde el plegamiento es distinto respecto al resto de estructuras de la clase A (Fig. sup. 17A). Lo anterior sugiere que estas dos betalactamasas podrían representar nueva subclase A3. Chaves-Silveria *et al.*[26] sugirieron una subclase A3 representada por una secuencia llamada LRA-5 que tuvo una baja similitud respecto al resto de betalactamasas que estudiaron. Para evaluar lo anterior, usé la secuencia de LRA-5 para computar una representación de tSNE con ESM-1b. LRA-5 se ubicó en el grupo A02, pero no específicamente dentro del grupo A2P. Además, su plegamiento es similar al de la betalactamasa PER-1 (Fig. sup. 17B), lo cual sugiere que LRA-5 podría no representar una nueva subclase A3 dado su similitud estructural con la subclase A2.

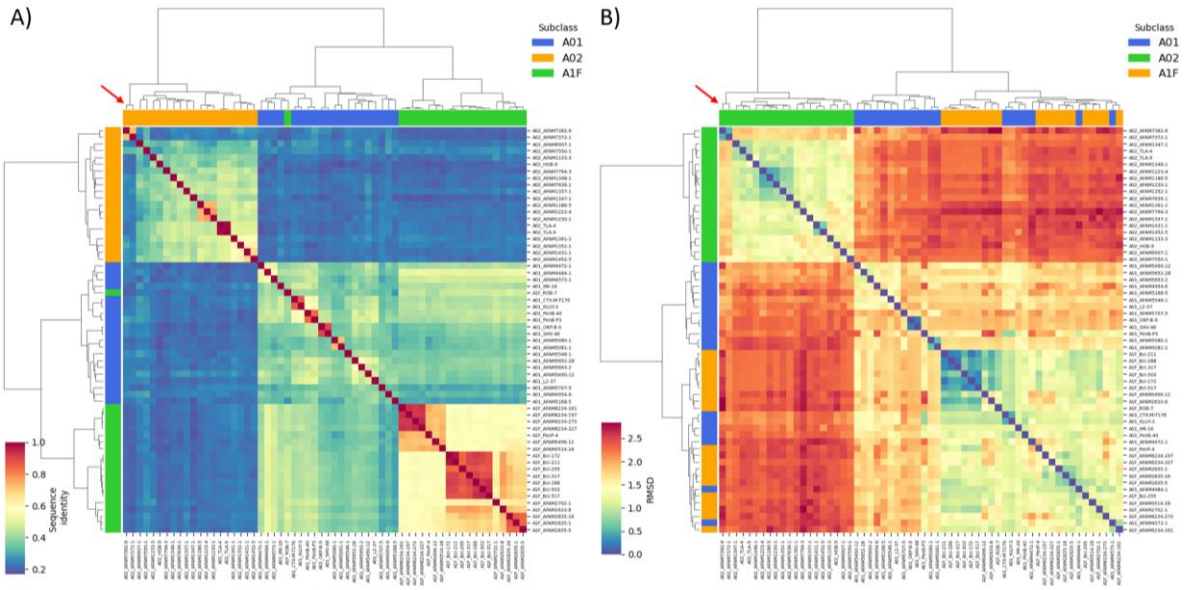


Figura 22. Similitud a nivel de secuencia y estructura de 60 betalactamasas clase A. (A) Matriz de similitud de secuencia. (B) Matriz de similitud de RMSD. Se muestra una leyenda de tres colores asociada cada uno de los tres grupos y una barra de color que indica los valores de identidad de secuencia o RMSD según el caso. Se señala con una flecha roja a las secuencias AFAM7362-6 y AFAM7372-1. Las matrices de similitud se construyeron con las estimaciones por pares de los respectivos valores y posteriormente, se visualizaron con la función clustermap (method = ward, metric = euclidean) de la librería Seaborn. Se presentan más detalles en las estructuras en las Figuras suplementarias 16, 17 y 18.

Philippon et al.[23] dividieron a la clase A con base en firmas conservadas a nivel de estructura y secuencia, siendo el plegamiento del sitio activo la más relevante. En la subclase A2, representada por PER-1, el sitio activo está expandido en relación con la subclase A1, lo cual favorece la actividad contra cefalosporinas. Las betalactamasas clase A se caracterizan por la presencia de una estructura llamada *loop* omega[L], el cual porta uno de los seis residuos catalíticos (E166). Este *loop* se encuentra cerca del sitio activo y sirve como una compuerta dinámica que permite una mayor apertura del sitio activo[170]. En PER-1, el *loop* omega adopta un plegamiento en forma de "V" debido a que tiene varias inserciones que forman una segunda alfa hélice[171] (Fig. sup. 17A). Además, varias regiones de giro cercanas al sitio activo cuentan con varias inserciones, lo cual promueve su flexibilidad.

Para evaluar estas características estructurales en los grupos A01, A1F, A02 y A2P (Fig. sup. 18A), tomé las secuencias representativas a un 90% de identidad de cada grupo, construí un alineamiento con ellas usando MAFFT, estimé una secuencia consenso con un enfoque por pesos usando BuddySuite, predije sus estructuras con AlphaFold2 y las alineé usando ChimeraX. Observé que la estructura del consenso A1F tiene unas pocas inserciones en giros respecto al consenso de A01 (Fig. 23). El consenso A02 también tiene inserciones respecto al consenso A01 y su *loop* omega adopta un plegamiento en forma de V al igual que PER-1. El consenso A2P también tiene inserciones respecto al consenso A01 y su *loop* omega está compuesto por tres hélices. Las secuencias del grupo A2P

[L] El "*loop* omega" es una región estructural importante en las serinbetalactamasas y otras proteínas. Es una región de giro que adopta una conformación similar a la letra omega del alfabeto griego. Particularmente en las betalactamasas clase A, el *loop* omega presenta una o más alfa hélices y además porta un residuo de ácido glutámico asociado a la catálisis (ver Figura 23).

también son considerablemente más largas respecto al resto (Fig. sup. 18B), lo cual es consistente con la presencia de las tres hélices del *loop* omega. Sin embargo, parece que estas características se asocian principalmente a familia Sphingomonadaceae (Fig. sup. 19). Estas observaciones se basan en predicciones estructurales, sin embargo, la caracterización de una nueva subclase requiere de análisis finos sobre la geometría del sitio activo, posible función de sitios conservados e identificación de firmas estructurales[15,23]. Por estos motivos, sería interesante obtener una estructura cristalográfica de las betalactamasas de Sphingomonadaceae, corroborar las observaciones y determinar si las diferencias a nivel de secuencia, estructura y función catalítica son suficientes para establecer una nueva subclase A3. Lo mismo también aplica a los taxa cuyas predicciones estructurales corresponden a la putativa subclase C2 (Fig. sup. 12). Tomando los resultados en conjunto, sugiero que la representación de tSNE considera las variaciones a nivel de secuencia y estructura de las betalactamasas, razón que habría hecho que la clase A se dividiera en tres grandes grupos.

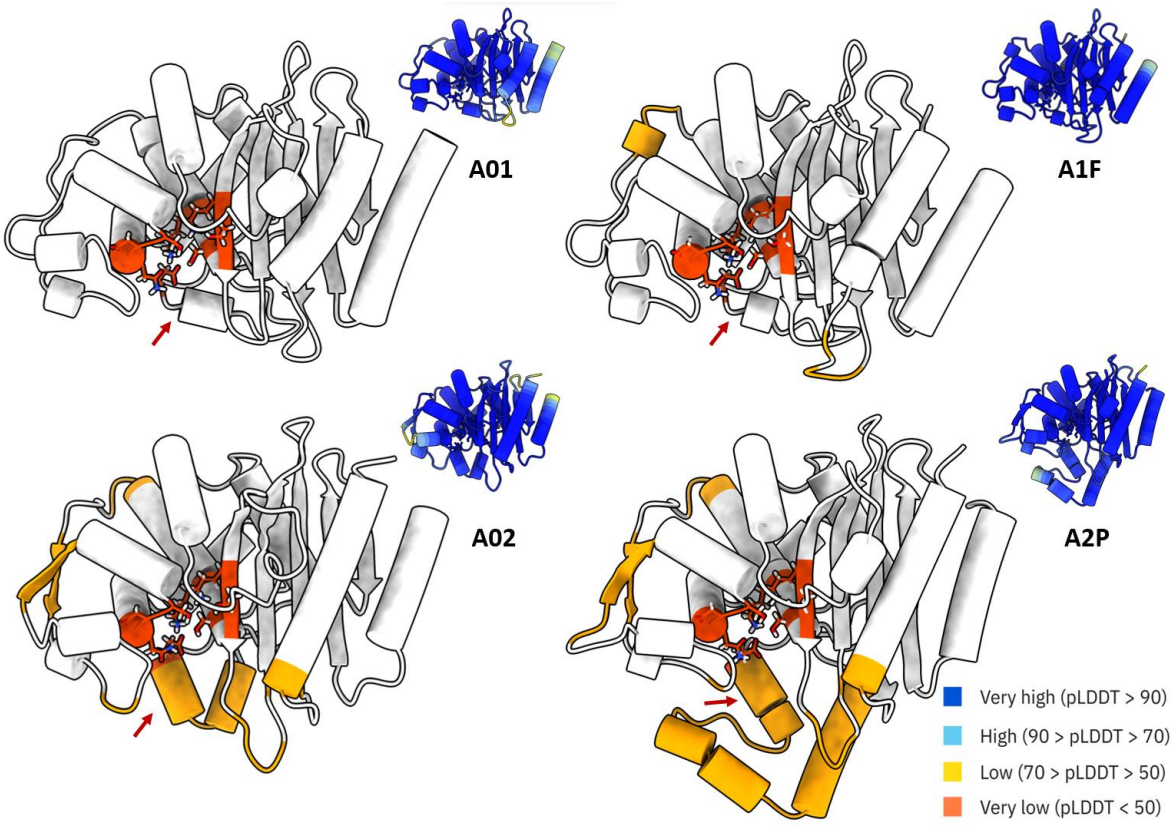


Figura 23. Diversidad estructural de la clase A. Se muestra la estructura predicha con AlphaFold2 de cada consenso. Se señala en rojo a los residuos catalíticos (S70xxK, S130DN, E166 y K233TG). Se señala con una flecha roja la ubicación del *loop* omega. Se resaltan en naranja las regiones estructurales con diferencia respecto al consenso A01. Estas regiones se identificaron con la versión de TM-align implementada en la PDB (https://www.rcsb.org/alignment). A la derecha de cada estructura se muestra la misma estructura, pero coloreada por valores de pLDDT. Las caricaturas se crearon con ChimeraX. Número de secuencias representativas por grupo: A1F = 251, A01 = 2,916, A02 = 588, A2P = 306.

Al mapear las propiedades fisicoquímicas/estadísticas observé que la fracción de residuos propensos a formar hélices y betas plegadas, fracción de residuos aromáticos, entropía y masa molecular cuentan con las correlaciones más altas respecto a las dimensiones de tSNE ($\rho > 0.50$ con alguna de ellas; Fig. 24 y Fig. sup. 10B). Nuevamente, la longitud y masa molecular son propiedades redundantes entre sí ($\rho = 0.95$; Fig. sup. 11B). La clase D tiene una mayor fracción de residuos propensos a formar regiones hélice (0.32% ± 0.03%; n = 2759) respecto a la clase A (0.27% ± 0.03%; n = 13,177) y C (0.30 % ± 0.01%; n = 6,564). La clase A tiene una menor fracción de residuos aromáticos (0.06% ± 0.01%) respecto a las clases C (0.10% ± 0.01%) y D (0.10% ± 0.02%). De forma similar, la clase A tiene una menor entropía a nivel de secuencia (3.98% ± 0.05 bits) respecto a la clase C (4.09 ± 0.05 bits) y la clase D (4.09 ± 0.06 bits). Curiosamente, en la clase C, la familia enzimática EC asociada a *Escherichia* (Fig. 18C y 20) muestra altos valores de entropía. Lo anterior sugiere que pese a estar limitadas a un solo género, estas betalactamasas cuentan con una alta diversidad en composición de secuencia.

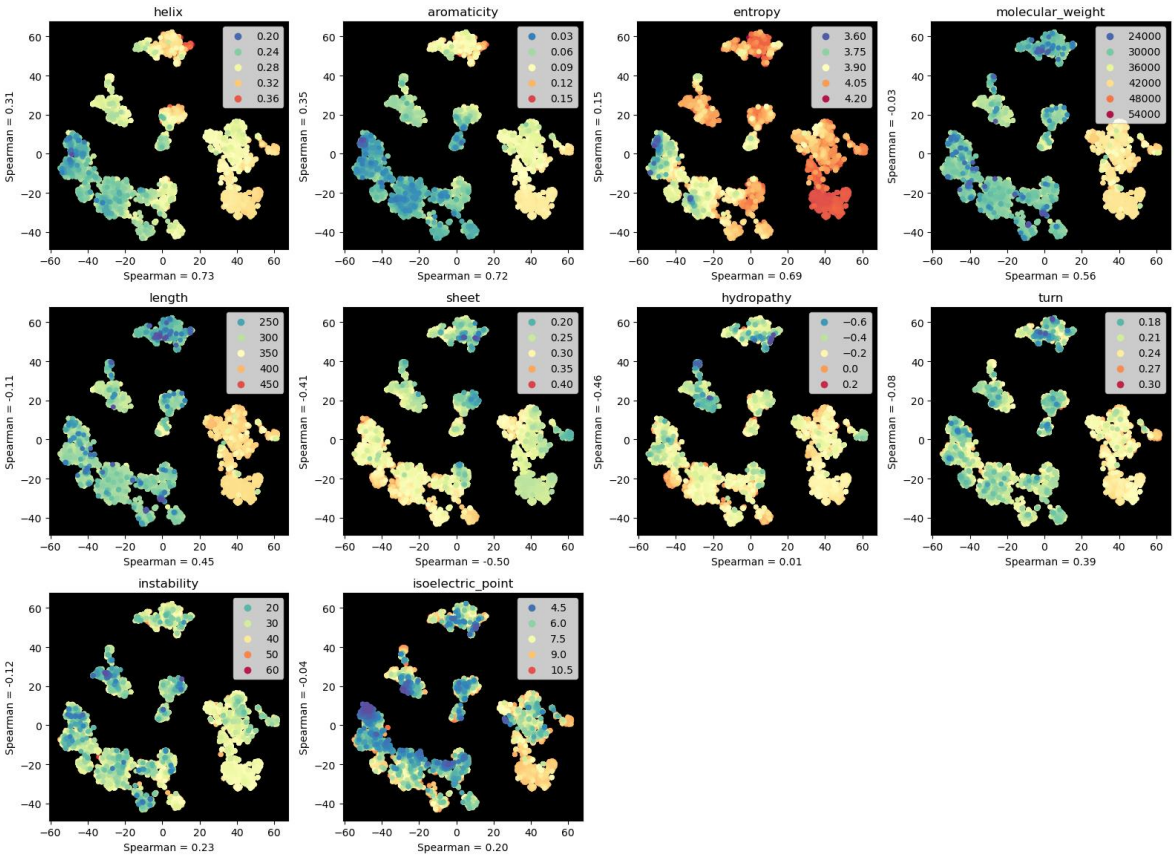


Figura 24. Propiedades cuantitativas de las serinbetalactamasas. Se indica en el encabezado de cada panel la propiedad correspondiente a cada una de las subclases. Se omite el nombre de las dos dimensiones de tSNE y en su lugar se muestran los valores de correlación de Spearman que tiene dicho eje respecto a los valores de la propiedad en cuestión. Para una mejor visualización, se removieron los 179 coordenadas correspondientes a secuencias cuya longitud no cae dentro de un rango de ±30% del valor de la mediana de cada clase, esto ayuda a remover valores extremos que sesgan el rango de visualización. Para más detalles sobre las propiedades consultar la Tabla 2 y las Figuras suplementarias 10 y 11.

El resto de las propiedades muestra una correlación media, baja o nula. Sin embargo, dentro de cada clase de serinbetalactamasas existe una estructura interna más fina. Por ejemplo, las secuencias de Sphingomonadaceae dentro del grupo A2P se caracterizan por tener un alto punto isoeléctrico con relación al resto de la clase A. Tomando los resultados en conjunto, sugiero que las clases de betalactamasas son la principal propiedad que separa a las serinbetalactamasas. Posteriormente se genera una estructura interna basada en variaciones estructurales y de similitud de secuencia dentro de cada clase. Y solo algunas pocas propiedades fisicoquímicas/estadísticas tiene buena correlación con las dimensiones de representación de tSNE.

Para saber si las representaciones de tSNE de ESM-1b se ven alteradas por el número de secuencias con alta similitud (>90% identidad), computé una nueva representación de tSNE usando sólo secuencias representativas. Observé que las clases A, C y D se separan, sugiriendo que la capacidad de ESM-1b para identificar a las tres clases no se ve muy alterada (Fig. 25A). Sin embargo, en la clase A si existen variaciones, pues el grupo asociado a Firmicutes no se distanció tanto respecto al grupo de Proteobacteria y Actinobacteriota (Fig. 25B); esto puede deberse a que al ser un menor número de secuencias de Firmicutes la fuerza de repulsión de tSNE no es tal para distanciarlo del grupo principal. Por otro lado, el grupo de secuencias del filo Bacteroidota (subclase A2) se sigue separando considerablemente del grupo principal. Sin embargo, el grupo de Proteobacteria similar a la subclase A2 no se separa tanto como lo hizo al usar el conjunto de datos completos. Lo anterior, sugiere que la organización local de la clase A (y posiblemente también el de la clase C) se ve parcialmente influenciada por la cantidad de secuencias con alta similitud de secuencia, lo cual probablemente se debe a las fuerzas atractivas y repulsivas de tSNE.

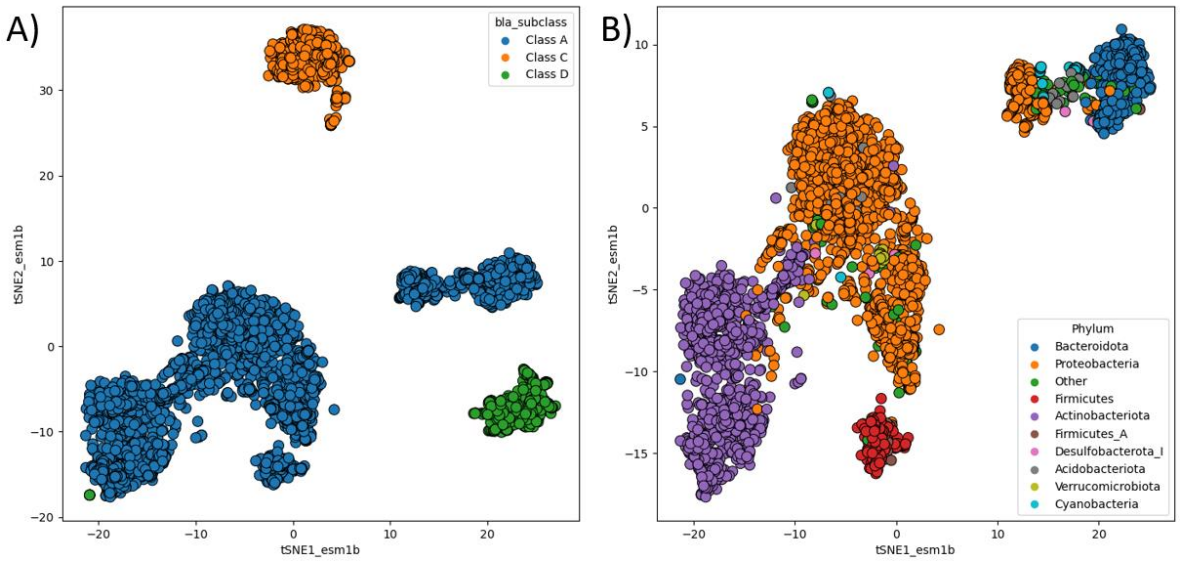


Figura 25. Representaciones de tSNE de ESM-1b al considerar las secuencias representativas de serinbetalactamasas. (A) Organización de las serinbetalactamasas. Después de los filtros de longitud e identidad de secuencia a se obtuvieron un total de 4,062 secuencias representativas de la clase A, 780 de la clase C y 608 de la clase D. Parámetros: divergencia de Kullback-Leibler = 0.36, perplexity = 400 y 1500 iteraciones. (B) Mapeo taxonómico de la clase A. Se muestran solo los 10 filos más abundantes y el resto fue etiquetado como “Otros”.

Al estimar las métricas de distancia entre las 100 secuencias muestreadas al azar por cada clase, observé que los *embeddings* de ESM-1b cuentan con la información suficiente para distinguir a las tres clases (Fig. 26). Por el contrario, el modelo Bepler no logra distinguir dichos grupos, lo cual también se refleja en el análisis con PCA de este modelo (Fig. 15).

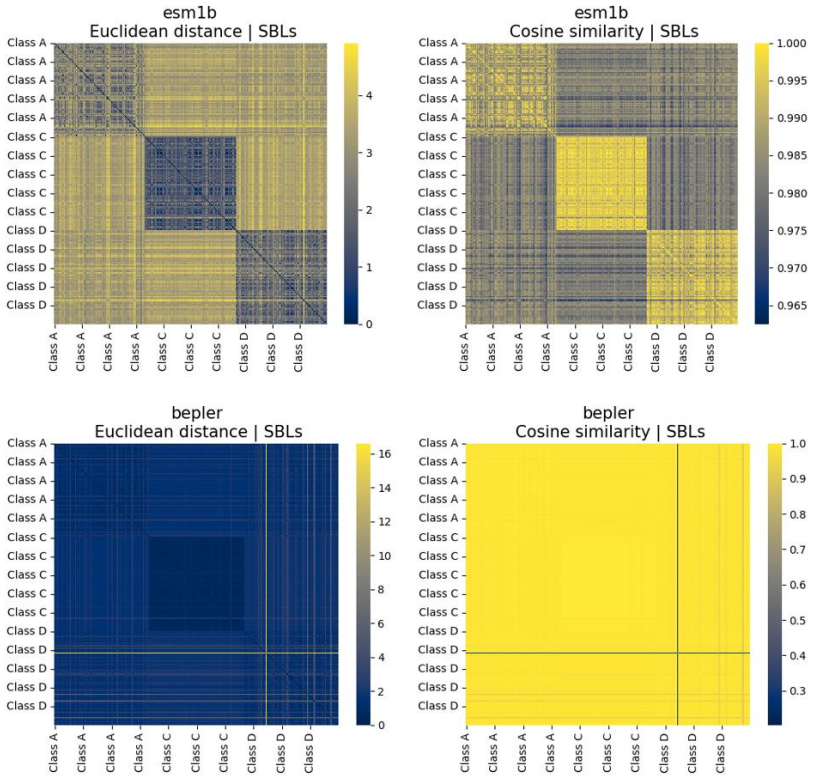


Figura 26. Organización de las clases de serinbetalactamasas usando métricas de distancia. Se indica en cada panel la métrica de distancia y el modelo de lenguaje de proteínas.

Posteriormente, usé las mismas 300 secuencias con *kmeans* para evaluar la capacidad de detección de las clases A, C y D (Fig. 27). Cuando el valor de K = 2, los *embeddings* de ESM-1b distinguen a la clase C como un grupo distinto a las clases A y D (ARI = 1), lo cual parece ser consistente con la estrecha relación entre estas dos clases[13,16]. Por el contrario, con el modelo Bepler no se distingue a ninguna de las clases (ARI = -0.003). Cuando el valor de K = 3, ESM-1b identifica bien a las tres clases de serinbetalactamasas (ARI = 0.86), sin embargo, unas pocas secuencias de Bacteroidota fueron erróneamente etiquetadas como clase D. Por el contrario, con el modelo Bepler, *kmeans* solo distingue a la clase C de las clases A y D (ARI = 0.54). Este comportamiento es similar al observado con las metalobetalactamasas, sugiriendo nuevamente que ESM-1b en combinación con *kmeans* puede detectar correctamente grupos de proteínas.

Dada la buena calidad de la información contenida en los *embeddings* de ESM-1b, sugiero entrenar algoritmos de clasificación supervisada para etiquetar serinbetalactamasas a múltiples niveles (clase, subclase, grupo, familia y subfamilia). Un sistema como este, podría ser valioso para automatizar la anotación de nuevas serinbetalactamasas tal y como Philippon *et al*.[15] sugirieron previamente. Y de hecho, es un proyecto en el que nuestro grupo de investigación está trabajando actualmente.

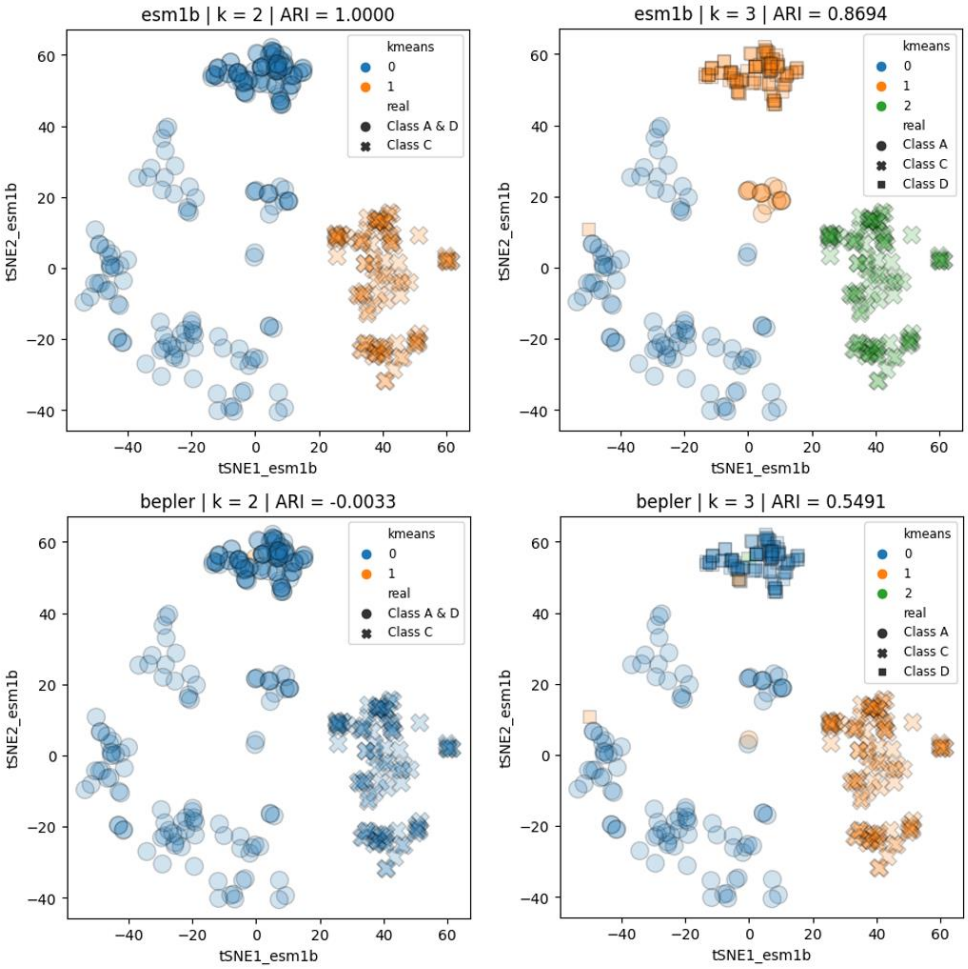


Figura 27. Organización de las serinbetalactamasas usando *kmeans*. Se indica en el encabezado de cada panel el modelo de lenguaje de proteínas, el valor de k de kmeans y el valor de adjusted rand index (ARI) obtenido al comparar los grupos de kmeans contra las clases reales. Se indica en la leyenda los grupos inferidos por *kmeans* (kmeans) y con distintos marcadores a los grupos de serinbetalactamasas (real).

PARTE 2. PREDICCIÓN DE LA FUNCIÓN CATALÍTICA

PROCESAMIENTO DE LAS BASES DE DATOS

Esta parte del proyecto consiste en crear una referencia de la actividad catalítica de betalactamasas representativas contra distintos antibióticos. Posteriormente, usar los *embeddings* de distintos PLM para entrenar regresores y predecir la actividad catalítica de betalactamasas que no han sido funcionalmente caracterizadas. Para ello, accedí, descargué (05 de Enero del 2022), ordené y etiqueté todos los datos disponibles en una segunda base de datos curada de betalactamasas (BLDB2)[102]. La BLDB2 integra datos para 215 betalactamasas representativas, clasificadas en cuatro clases y 41 familias (Fig. sup. 20). Estas enzimas representan un total de 2,383 datos de concentraciones mínimas inhibitorias (MICs[M]) para 21 conjuntos de antibióticos, cinco de ellos en combinación con inhibidores (Fig. 28). Dado que la BLDB2 no incluye las secuencias, mapeé los nombres de las betalactamasas contra la base de datos que curé en la primer parte de esta tesis (más detalles del proceso de curación en el Jupyter notebook "Create_functional_datasets").

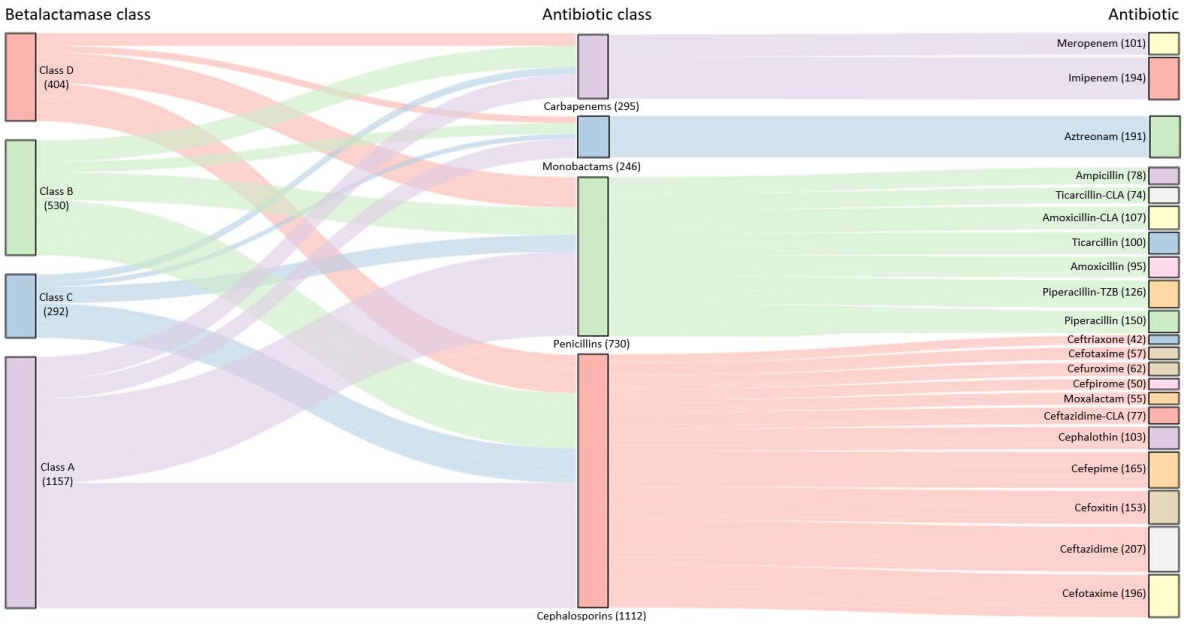


Figura 28. Distribución de los datos de concentraciones mínimas inhibitorias por clase de betalactamasas y de antibióticos. Se indica entre paréntesis el número de datos disponibles en cada categoría. Abreviaciones: CLA, ácido clavulánico; TZB, tazobactam.

[M] El término "MIC" es la abreviación de *Minimum inhibitory concentration* (concentración mínima inhibitoria).

En cada caso, existen datos de MICs que permitió crecer a las bacterias con y sin la betalactamasa. La BLDB2 usa estos dos datos para estimar la tasa de cambio (fold) generada por la betalactamasa. La interpretación de los valores de fold es la siguiente:

fold > 1: la betalactamasa provocó resistencia.

fold = 1: la betalactamasa no provocó resistencia.

fold < 1: la betalactamasa tuvo efectos contraproducentes en el crecimiento celular.

La condición "fold < 1" se consideró como "sin sentido" y se cambiaron los 54 casos de este tipo a una condición de "fold = 1". Dado que las mediciones de MICs se realizan en cantidades exponenciales (64, 128, 256 µg/mL, etc.), los valores de fold siguen la misma razón. Este tipo de datos no son fáciles de procesar por algoritmos de regresión, pues hace que los valores muy pequeños se vean iguales entre sí respecto a valores muy grandes[50,172]. Por ello, normalicé los valores de fold usando el logaritmo base dos, lo cual es una práctica común en este tipo de trabajos[173,174]. A estos valores normalizados les llamaré "incremento de resistencia" de aquí en adelante (Fig. 29).

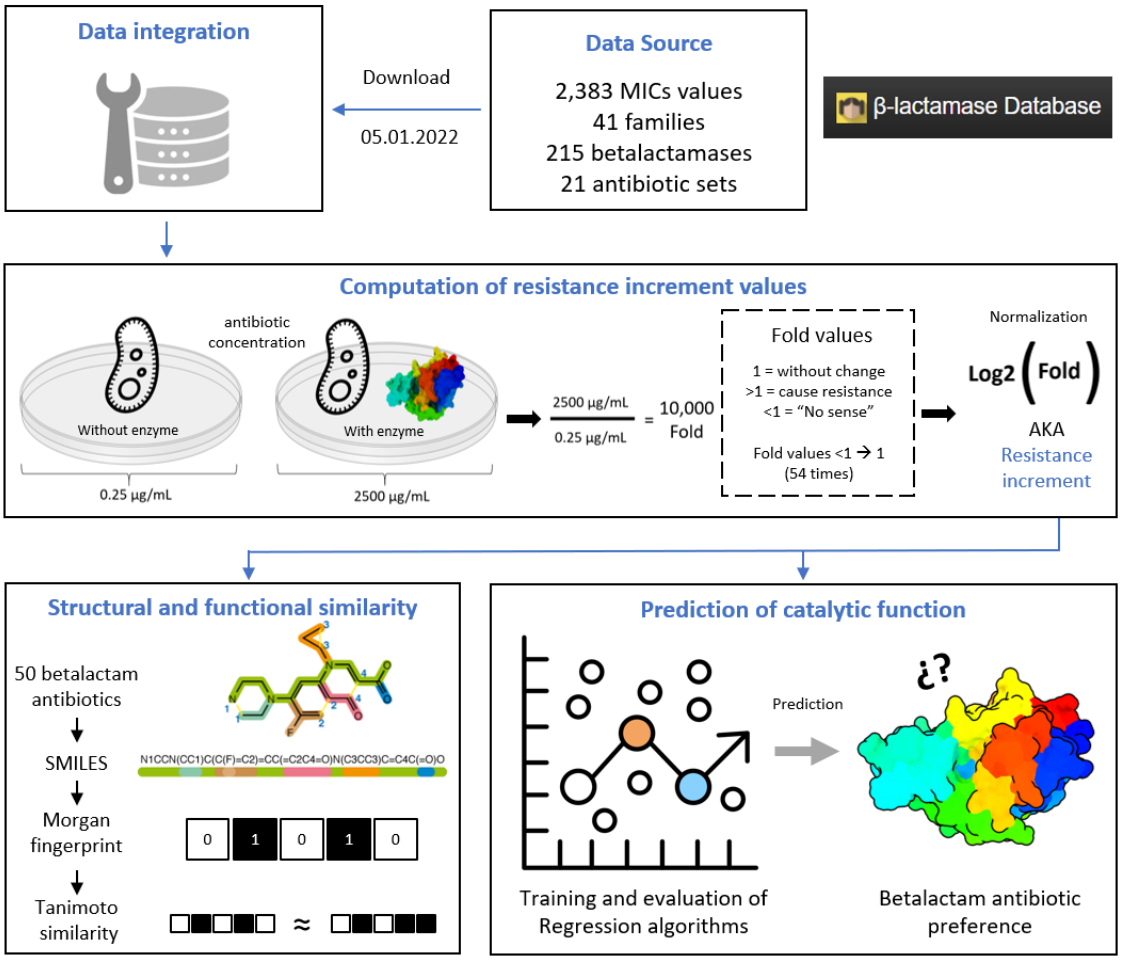


Figura 29. Estrategia de análisis y predicción de la actividad catalítica de las betalactamasas. Se muestra un diagrama con el procesamiento de los valores de fold e incremento de resistencia. Con estos últimos, se realizaron análisis de la relación estructural y funcional entre antibióticos betalactámicos, así como el entrenamiento de modelos de regresión.

EVALUACIÓN DE LOS PERFILES DE RESISTENCIA

La distribución de los datos de MICs reflejan el sesgo de investigación a betalactamasas de interés clínico. El 77.76% de los datos corresponden a variantes representativas de familias reconocidas de la clasificación funcional (Fig. sup. 20). También se observa un sesgo a favor de la clase A, cuyos datos equivalen al 48.55% del total (Fig. 28). Igualmente, hay más datos de cefalosporinas y penicilinas, los cuales representan el 46.66% y 30.63% del total, respectivamente. Lo anterior es consistente con el hecho de que estas dos clases de antibióticos representan más de la mitad de las ventas globales de antibióticos[2,3], así como que la clase A es la más estudiada[9,48].

Para evaluar qué clase de betalactamasa genera alta resistencia contra las clases de antibióticos betalactámicos, removí tres valores extremos de fold (Fig. sup. 21) y posteriormente, grafiqué la distribución de los valores de incremento de resistencia (Fig. 30A). Observé que la clase A genera alta resistencia contra penicilinas y monobactamas, moderadamente contra cefalosporinas y poca contra carbapenemas. La clase B genera alta resistencia contra penicilinas, moderadamente contra cefalosporinas y carbapenemas y poca contra monobactamas. Además, la clase B es la que mayor resistencia genera contra carbapenemas respecto al resto de clases, lo cual es consistente con la literatura[6,175]. La clase C genera alta resistencia contra cefalosporinas, penicilinas y monobactamas, y poca contra carbapenemas. La clase D genera alta resistencia contra penicilinas y poca contra cefalosporinas, carbapenemas y monobactamas. Es notable que todas las clases generan una relativamente alta resistencia contra penicilinas, lo cual sugiere que podrían ser un sustrato “común” entre las betalactamasas.

Para evaluar el efecto de los inhibidores sobre el incremento de resistencia, seleccioné los conjuntos con inhibidores y grafiqué su distribución (Fig. 30B). Observé que el conjunto Piperacillin-Tazobactam parece no tener diferencias significativas en sus efectos inhibitorios contra las clases A, B, C y D. De hecho, el conjunto Piperacilina-Tazobactam es comúnmente conocido por sus capacidades de inhibición contra betalactamasas de la clase A[176], sin embargo, también se ha registrado efectos inhibitorios contra las betalactamasas clase C[177], algunas de la clase D[178] y en menor medida de la clase B[179]. Por otro lado, los conjuntos con ácido clavulánico son comúnmente conocidos por su efecto inhibidor contra las betalactamasas clase A y en menor medida contra la clase D[180] (Fig. 2), dicho efecto se observa en los conjuntos de ácido clavulánico en combinación con Ceftazidima y Ticarcilina, pues no hay diferencias significativas entre las betalactamasas de las clases A y D. El efecto de inhibición contra la clase A también se aprecia en el conjunto de ácido clavulánico con Cefotaxima, donde la clase A presenta un menor incremento de resistencia respecto a la clase C. Sin embargo, el conjunto de ácido clavulánico con amoxicilina no se observan diferencias considerables de inhibición entre las clases de betalactamasas. Esta repuesta heterogénea ante la presencia de los inhibidores comúnmente asociados a una clase de betalactamasas sugiere que es necesario una reconsideración de su especificidad, pues las todas las clases de betalactamasas pueden generar resistencia a las combinaciones con inhibidores[43,45]. Tomando los resultados en conjunto, sugiero que los datos de MICs de BLDB2 son consistentes con los perfiles funcionales de las clases de betalactamasas reportados en la literatura.

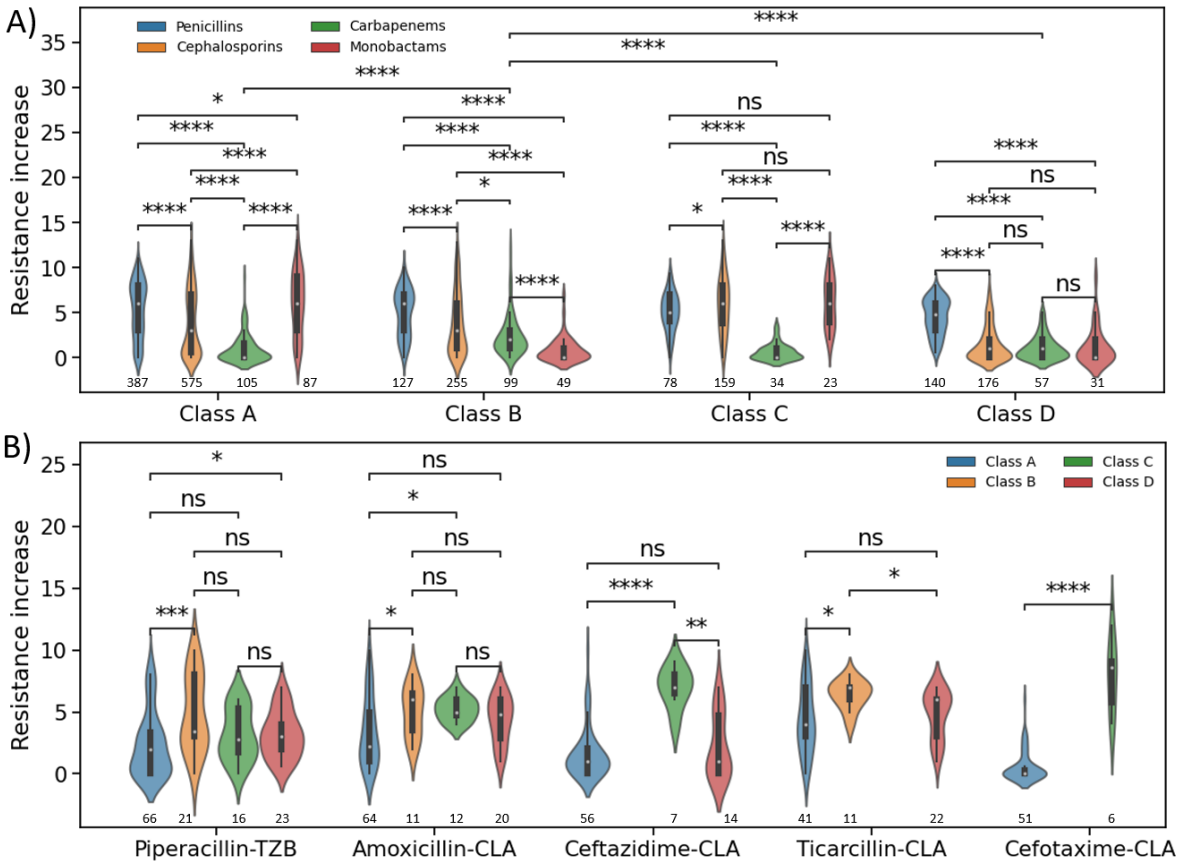


Figura 30. Distribución de los valores de incremento de resistencia. (A) Distribución por clase de betalactamasas. (B) distribución por conjuntos antibiótico-inhibidor. Se indica con un "*" si existen diferencias significativas entre los grupos al realizar un prueba de Mann-Whitney implementada en la librería de Python statannotations, y con un "ns" si las diferencias no son estadísticamente significativas. Se indica debajo de cada violín el número de datos correspondientes al grupo.

ANÁLISIS DE SIMILITUD ESTRUCTURAL Y FUNCIONAL ENTRE ANTIBIÓTICOS

Para este análisis consideré un los 50 antibióticos betalactámicos de la base de datos CBMAR[103] (Fig. sup. 22 y 23). Para conocer la similitud estructural entre estos antibióticos tomé sus respectivas SMILES de base de datos PubChem[181], generé sus respectivos *fingerprints*[N] de Morgan (bits = 2048, radius = 2), calculé la similitud de Tanimoto por pares usando RDKit[182] y visualicé esta matriz con un agrupamiento jerárquico. Las SMILES son una forma de representar compuestos mediante cadenas de caracteres[183]. Los *fingerprints* son vectores de presencia o ausencia de subestructuras presentes en un conjunto de compuestos[184]. La similitud de Tanimoto es una métrica que evalúa la similitud entre *fingerprints*, y toma valores entre 0 y 1, siendo valores ≈1 indicativos de una alta similitud[185,186].

Observé que las penicilinas se agruparon en un solo grupo y, además, son más similares entre sí respecto al resto de antibióticos (Fig. 31). Por ejemplo, piperacilina, azlocillina y mezlocilina muestran una alta similitud entre sí (≈0.8). Piperacilina tiene un grupo hidroxilo y etilo más respecto a azlocillina, mientras que mezlocilina difiere de estas dos por la presencia de un grupo sulfonilo (Fig. sup. 23). Algunas penicilinas tienen una considerable similitud (≈0.5) con cefalosporinas de primer y segunda generación, por ejemplo ampicilina-cefalexina o amoxicilina-cefadroxilo, lo cual podría relacionarse con las respuestas alergénicas debidas a reacciones cruzadas entre estas clases de antibióticos[187,188]. De hecho, por esta razón los grupos radicales de las recientes generaciones de cefalosporinas han sido diseñados para diferenciarse de los grupos radicales de las pencialinas[189]. Los carbapenemas se agruparon en un solo grupo. Por el contrario, las cefalosporinas se dividen en dos grupos principales; el primer grupo compuesto por cefalosporinas de primer y segunda generación junto con dos cefalosporinas de tercera generación (cefoperazona y moxalactama), y el segundo grupo compuesto por cefalosporinas de tercera, cuarta y quinta generación junto con las monobactamas que podrian haberse agrupado dado la similitud de sus grandes grupos radicales (por ejemplo, cefixima vs carumonam; Fig. sup. 23). Tomando los resultados en conjunto, sugiero que la similitud de Tanimoto puede ser una buena métrica para detectar similitudes estructurales entre antibióticos.

Para evaluar la similitud funcional entre antibióticos betalactámicos, tomé los valores de incremento de resistencia de los 21 conjuntos registrados en la BLDB2 y computé una matriz de correlación de Spearman. Observé cuatro grupos con redundancia funcional (Fig. 32), es decir, cuando una betalactamasa es buena degradando al antibiótico A, también es relativamente buena degradando al antibiótico B. El primero incluye a las combinaciones de penicilinas con inhibidores junto con los dos carbapenemas. El segundo incluye a tres penicilinas (ampicilina, amoxicilina y ticarcilina) que muestran una alta correlación entre sí ($\rho \geq 0.86$). El tercero incluye los conjuntos de cefalosporinas con ácido clavulánico junto con moxalactam y cefoxitina. El cuarto incluye al resto de las cefalosporinas junto con piperacilina y aztreonam, mostrando entre sí correlaciones moderadas ($\rho \geq 0.50$). En este grupo el par cefotaxima-ceftriaxona muestra la correlación más alta de todos ($\rho = 0.93$),

[N] El término "*fingerprints* de Morgan" es un anglicismo que se usa para referirse a una representación vectorial de una sustancia química. Esta representación codifica información sobre la conectividad de los átomos de una molécula, lo cual permite realizar análisis de similitud a partir de la presencia y ausencia de subestructuras químicas.

consistente con reportes que indican que estas cefalosporinas generan una resistencia similar pese a su diferencia farmacocinética[190,191]. Hay casos similares entre penicilinas; por ejemplo, cuando se prescriben amino-penicilinas (*i.e.* ampicilina y amoxicilina) es común encontrar reacciones cruzadas[188] y co-selección entre genes de resistencia ante estos antibióticos[192]. Esto también ocurre entre amino-penicilinas y carboxi-penicilinas (*i.e.* ticarcilina), registrando niveles de resistencia similares[193]. Curiosamente, el par moxalactam-ampicillin tiene la correlación más negativa de todos, consistente con un reporte que sugiere un efecto parcialmente sinérgico entre estos antibióticos[194].

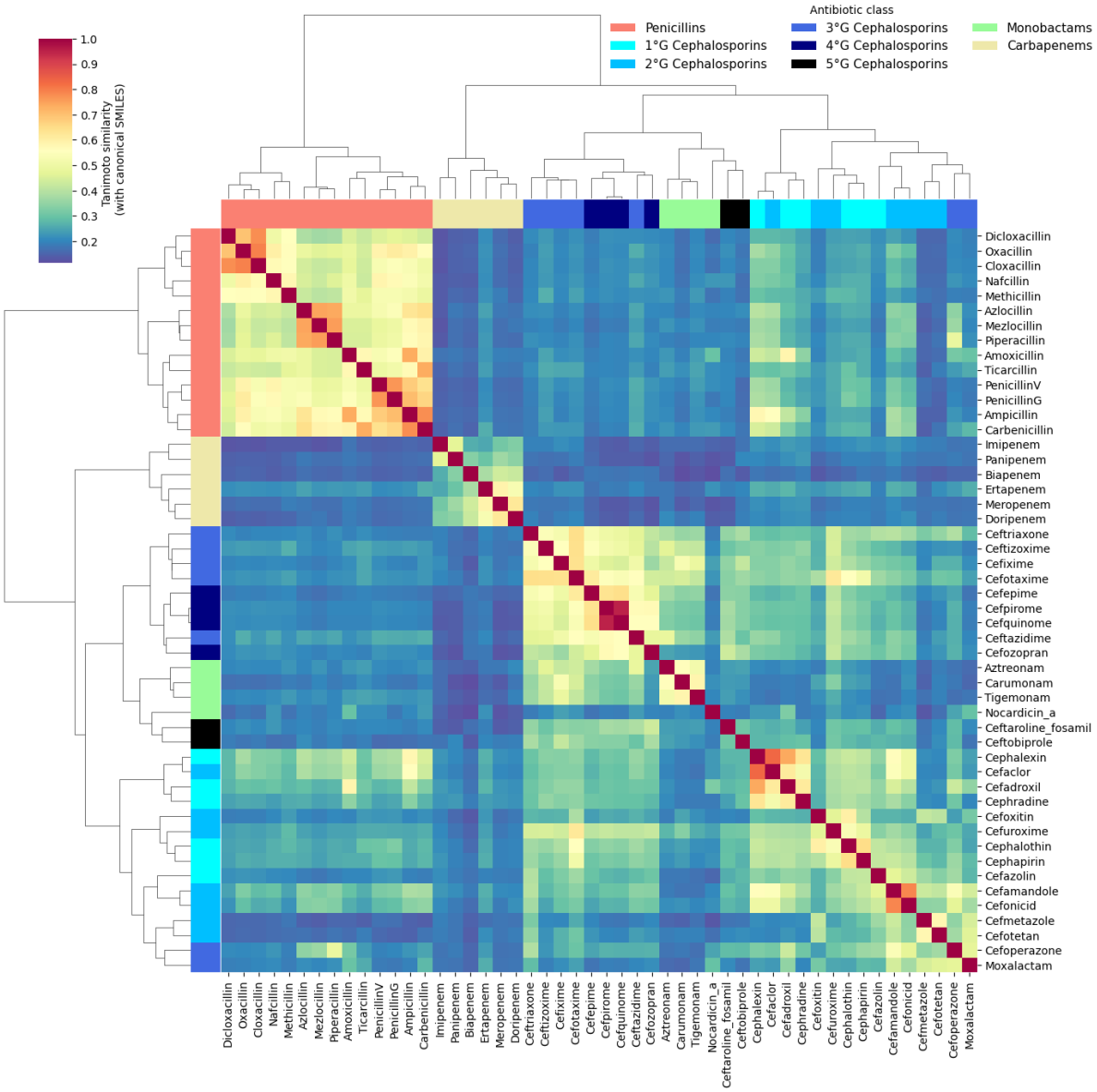


Figura 31. Similitud de Tanimoto entre 50 antibióticos betalactámicos. Se muestra una barra de colores con los valores de similitud, así como una leyenda con las cuatro clases de antibióticos, dividiendo a las cefalosporinas por generaciones. Los valores de similitud se obtuvieron con las SMILES canónicas, sin embargo, la PubChem también ofrece SMILES isoméricas que incluyen información estereoquímica. Analicé ambos tipos de SMILES y dieron los mismos resultados. Por convención, presento los resultados de las SMILES canónicas (más detalles el Jupyter notebook "SMILES_analysis_Tanimoto"). Los datos se visualizan con clustermap (method = ward, metric = euclidean) de la librería Seaborn.

Tomando los resultados en conjunto, sugiero que existe una redundancia funcional entre determinados pares de antibióticos, lo cual se reflejaría en la capacidad de una betalactamasa para degradar ambos sustratos de forma similar. Sin embargo, desde una perspectiva funcional, la división entre clases de antibióticos no es tan clara como lo fue a nivel estructural. Este análisis también me permitió definir una estrategia de regresión y decidí entrenar un modelo de regresión por cada uno de los 21 conjuntos de antibióticos betalactámicos.

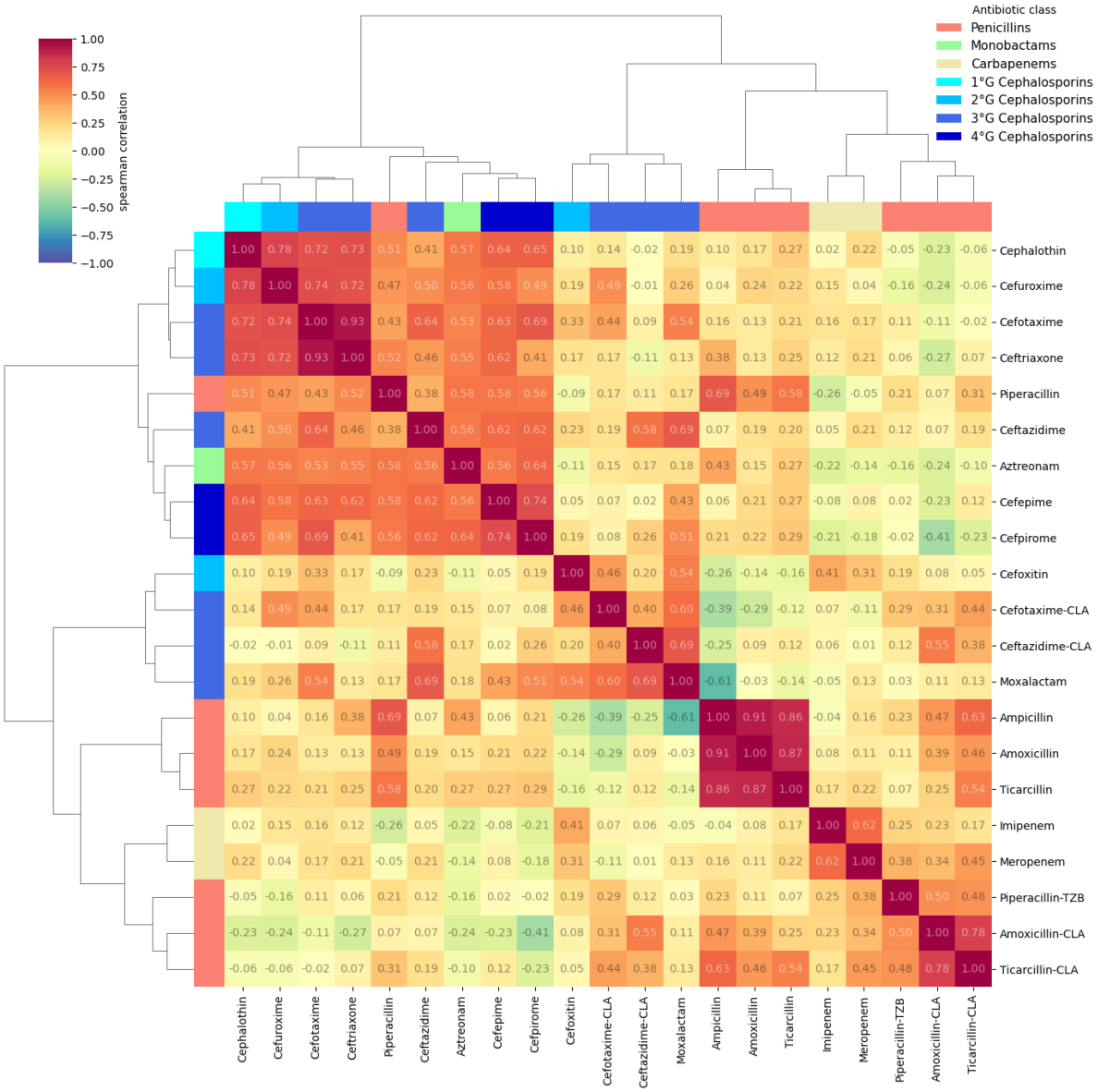


Figura 32. Matriz de correlación de Spearman de los valores de incremento de resistencia entre 21 conjuntos de antibióticos betalactámicos. Se muestra una barra de colores con los valores de correlación y su valor en cada recuadro, así como una leyenda de las cuatro clases de antibióticos, dividiendo a las cefalosporinas por generaciones. Se usaron los valores de incremento de resistencia como columnas y las 215 betalactamasas como filas para crear una tabla pivote a partir de la cual se computó la correlación por pares con la librería Pandas. Los datos se visualizan con clustermap (method = ward, metric = euclidean) de la librería Seaborn.

Tres observaciones derivadas de estos análisis resultan interesantes. La primera es que todas las clases de betalactamasas parecen buenas al degradar penicilinas respecto al resto de clases de antibióticos (Fig. 30A). La segunda es que la similitud de Tanimoto entre penicilinas es considerable tal que todas se agruparon juntas (Fig. 31). Y la tercera es que dos de los cuatro grupos de redundancia funcional incluyen penicilinas con buenas correlaciones (*i.e.* $\rho \geq 0.7$; Fig. 32). Visto lo anterior, podría ser posible que las penicilinas sean sustratos más comunes de degradar por parte de las betalactamasas respecto al resto de las clases de antibióticos.

Para cuantificar la relación entre la similitud estructural y funcional de los antibióticos betalactámicos, analicé en conjunto la similitud de Tanimoto e incremento de resistencia de 16 conjuntos de antibióticos sin inhibidores. Para este conjunto, existen 120 pares no redundantes con los que generé una tabla ordenada por incremento de resistencia y calculé la correlación de Spearman entre los valores de similitud de Tanimoto y correlación de Spearman de los valores de incremento de resistencia. En términos globales, observé una correlación moderada ($\rho = 0.66$) entre estos valores, y dicha correlación tiende a aumentar conforme conservo los valores con mayor correlación de incremento de resistencia (Fig. sup. 24). Lo anterior sugiere que a mayores valores de correlación de incremento de resistencia existe una mayor similitud de Tanimoto. Para evaluar este comportamiento en las clases de antibióticos, segmenté los pares de acuerdo con su clase (Fig. 33). Observé una alta correlación ($\rho = 0.9856$) entre pares de penicilinas, lo cual parece consistente con su considerable similitud estructural (Fig. 31) y funcional (Fig. 32), así como con sus altos valores de incremento de resistencia en todas las clases de betalactamasas (Fig. 30A). Notablemente, los pares de cefalosporinas tienen una correlación moderada ($\rho = 0.5017$). El resto de pares mostraron bajas correlaciones o bajos niveles de confianza.

Tomando los resultados en conjunto, sugiero que las penicilinas pueden ser un sustrato común de degradar por todas las clases de betalactamasas, lo cual posiblemente se debe a su considerable similitud estructural. Esto también podría aplicar a las cefalosporinas en menor medida debido a la heterogeneidad estructural de las cinco generaciones.

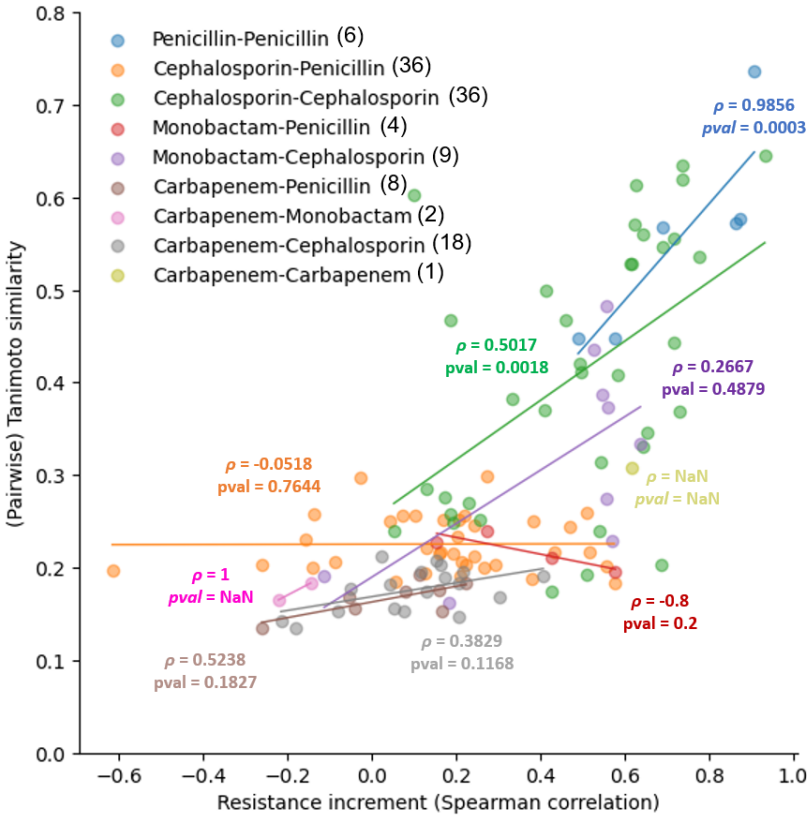


Figura 33. Correlación de Spearman entre la similitud de Tanimoto y correlación de incremento de resistencia por clase de antibióticos. Cada punto representa un par de antibióticos para los cuales tengo sus valores de similitud de Tanimoto (Fig. 30) y sus respectivos valores de correlación de Spearman de incremento de resistencia (Fig. 31). Se muestra con letras de colores el respectivo valor de correlación de Spearman y sus *p-values* asociados. En la leyenda se indica la clase de los pares de antibióticos y entre paréntesis la cantidad de datos contenidos del total de 120 pares no redundantes.

ENTRENAMIENTO DE REGRESORES Y PREDICCIÓN DE LA FUNCIÓN CATALÍTICA

Para predecir los valores de incremento de resistencia de betalactamasas sin anotación funcional es necesario determinar el mejor PLM y algoritmo de regresión. Para ello, consideré los siete PLM (Tab. 1) y añadí dos representaciones de PCA de ESM-1b y ProtT5-XL-U50 que modelan el 99% de la varianza (Fig. sup. 25). Solo usé 1,850 datos de serinbetalactamasas (Fig. sup. 26) debido a que la catálisis en las metalobetalactamasas es altamente dependiente de las concentraciones de Zinc[27], y además, no hay homología entre estas superfamilias[9].

Para identificar el mejor algoritmo de regresión usé la librería LazyPredict[195], la cual entrena automáticamente 42 regresores distintos (*i.e.* sin optimizar sus hiperparámetros). Por cada una de las nueve codificaciones[O] y los 21 conjuntos de antibióticos, dividí al azar los datos en 80% para el conjunto de entrenamiento y 20% para el conjunto de prueba. Después del entrenamiento, tomé los diez mejores regresores basado en su error cuadrático medio y grafiqué su frecuencia considerando solo al mejor regresor, los mejores tres y la totalidad de ellos (Fig. sup. 27). El regresor más frecuente entre los mejores fue NuSVR, el cual además se ubicó en el tercer lugar al considerar los tres mejores y en el sexto lugar al considerar todos los regresores.

Para identificar la mejor codificación, tomé el mejor regresor entrenado por LazyPredict (*i.e.* el de menor error cuadrático medio) y calculé la correlación de Spearman con los datos del conjunto de prueba. Por cada codificación, calculé el promedio de la correlación entre los 21 antibióticos, así como el número de veces que se obtiene un $\rho > 0.5$ (Fig. sup. 28A). Contrario a lo que esperaba, ESM y Prot-T5-BFD tienen los mejores valores en ambos filtros, lo cual contrasta con los buenos resultados de ESM-1b (y Prot-T5XL-U50) en la primera parte del proyecto. Al observar los valores de correlación por antibiótico (Fig. sup. 29B) encontré que no hay una relación entre la cantidad de datos disponibles y un mejor desempeño del regresor (*i.e.* mayor correlación de Spearman). Por ejemplo, Moxalactam y Ceftazidime cuentan con 34 y 156 datos, respectivamente, y en ambos casos la mayoría de las codificaciones obtienen un $\rho \approx 0.7$. También observé que la mayoría de las codificaciones en Piperacilin-TZB y Cefepime obtienen un $\rho < 0.5$, lo cual sugiere que son dos conjuntos difíciles de modelar. Tomando los resultados en conjunto, sugiero que la combinación entre NuSVR y Prot-T5-BFD es una buena estrategia para el entrenamiento.

El conjunto de datos de serinbetalactamasas no representan un escenario ideal para obtener buenos modelos de regresión debido a la distribución desbalanceada de las clases de betalactamasas. De hecho, la clase A representa el 62.38% de todos los datos (Fig. sup. 26). Además, existen muy pocos datos por cada antibiótico, siendo Ceftazidime (156 datos) y Ceftriaxone (33 datos) los antibióticos que más y menos datos tienen, respectivamente. Para reducir el mayor sesgo posible y asegurar la capacidad de generalizar a datos fuera del conjunto de entrenamiento, usé una estrategia de validación cruzada estratificada (Fig. sup. 29). Por cada antibiótico dividí al azar los datos en 70% para el conjunto de entrenamiento y 30% para el conjunto de prueba, asegurando que la cantidad de

[O] El término "codificaciones" se usa para referir tanto a los *embeddings* de los distintos modelos de lenguaje de proteinas así como a las representaciones de baja dimensión inferidas con PCA para los modelos ESM-1b y ProtT5-XL-U50.

datos de las tres clases de serinbetalactamasas fuera proporcional entre ambos conjuntos. Posteriormente, determiné los mejores valores de los hiperparámetros del regresor NuSVR mediante una validación cruzada aleatoria de cinco pliegues en los datos de entrenamiento usando la función GridSearchCV de SciKit-learn[196]. Finalmente, entrené un nuevo regresor usando los mejores hiperparámetros y los datos del conjunto de entrenamiento. Con este regresor final, realicé las predicciones sobre el conjunto de datos de prueba y calculé cinco métricas de desempeño de regresión entre los valores reales y predichos. Considerando estas métricas, clasifiqué los 21 modelos en cuatro categorías de calidad: alta, media, baja y basura (Fig. 34). Aunque esta estrategia ayuda a entrenar buenos modelos, siempre existirá un sesgo en los datos que puede asociarse a múltiples categorías. Por ejemplo, por la clase de serinbetalactamasas, familia enzimática e incluso, por la categoría taxonomía asociada a las secuencias como previamente se ha visto[197]. Del total de 21 antibióticos, solo conseguí un modelo de buena alta para Cefoxitina. Además, obtuve cinco modelos de calidad media, cinco modelos de calidad baja y 10 modelos basura.

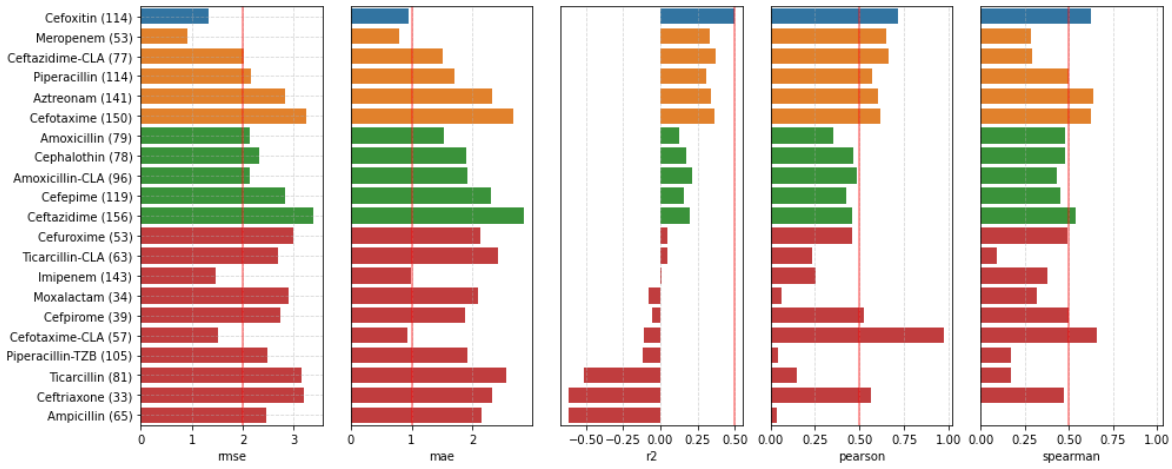


Figura 34. Calidad de los modelos de regresión usando NuSVR y los *embeddings* de ProtT5-BFD. Se muestran los valores de la raíz del error cuadrático medio (rmse), error absoluto medio (mae), coeficiente de determinación (r2) y de correlación de Pearson y Spearman entre los valores reales y las predicciones de los datos del conjunto de prueba. Se indica entre paréntesis la cantidad de datos disponibles de cada antibiótico y con una línea roja, el valor límite considerado para determinar la calidad del modelo: rojo = basura, verde = baja calidad, naranja = calidad media, azul = buena calidad. Esta clasificación es subjetiva y con especial énfasis en el coeficiente de determinación.

Aunque el regresor de Cefoxitina es de alta calidad, sus coeficientes de correlación son moderados ($\rho \approx 0.6$, $r \approx 0.6$), por lo cual se espera que el detalle de sus predicciones no sea tan fino y en su lugar, puede ser considerado de grano grueso. Para evaluar la habilidad predictiva de este regresor, comparé los valores de incremento de resistencia reales contra los predichos. El conjunto de datos usados durante el entrenamiento se compone de 114 datos distribuidos en 24 familias enzimáticas reconocidas (Fig. 35A). Del total de 114 datos, las clases A, C y D representan el 62.28%, 22.81% y 14.91%, respectivamente (Fig. sup. 26 y 29). Al observar las predicciones de todo el conjunto de serinbetalactamasas noté varias familias con altos valores predichos que no se encuentran en las 24

familias usadas durante el entrenamiento (Fig. 35B). La mayoría de estas familias son clase C, lo cual contrasta con los pocos datos de esta clase que fueron usados en el entrenamiento. Curiosamente, se ha reportado que algunas variantes dentro de estas familias son capaces de generar alta resistencia contra Cefoxitina, como es el caso de Ear-1[198], MOX-4[199] o CMH-2[200]. Además de estos casos, existen otras familias reconocidas que cuentan con menos de 10 variantes, así como otras familias no reconocidas por la BLDB, cuyos valores predichos son altos. Tomando los resultados en conjunto, sugiero que el regresor de Cefoxitina es capaz de generalizar a datos fuera del conjunto de entrenamiento. Además, puede ser útil en la identificación de betalactamasas que generen alta resistencia contra Cefoxitina, con las cuales se puede realizar análisis detallados a nivel de secuencia y estructura para dar una explicación a su alta actividad contra Cefoxitina.

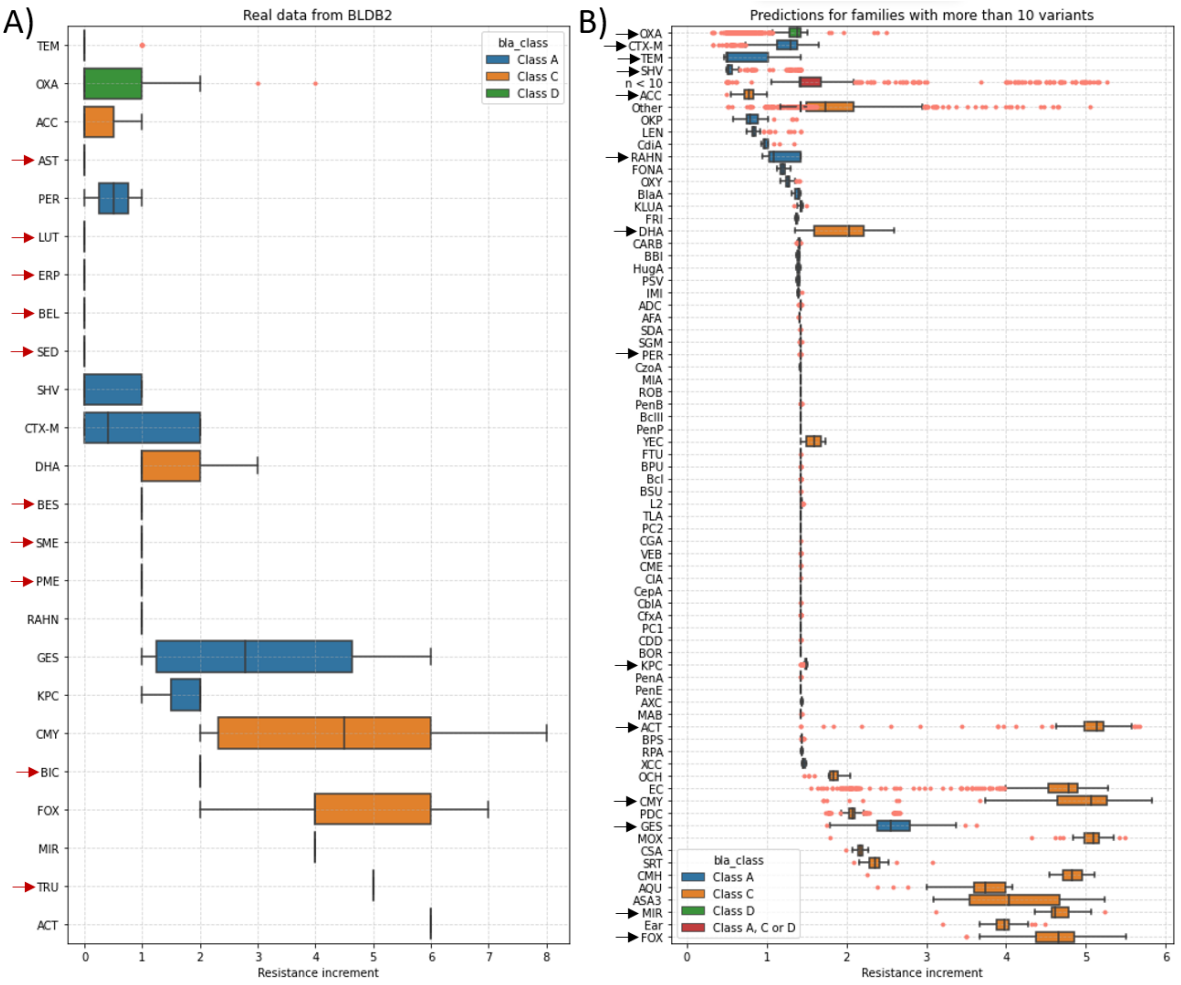


Figura 35. Distribución de los valores de incremento de resistencia reales y predichos para Cefoxitina. (A) Distribución de los datos reales. Se señala con flechas rojas a las familias que tienen menos de 10 variantes y que no aparecen en el panel B. (B) Distribución de los datos predichos. Para una mejor visualización, creé una nueva categoría llamada "n < 10" la cual contiene todas las familias compuestas por menos de 10 variantes dentro de la clase "Clase A, C o D". 73 familias cuentan con >10 variantes, las cuales se indican con su respectivo nombre. Se señala con flechas negras a las familias con más de 10 variantes que están presentes en los datos de entrenamiento del panel A.

No fue posible obtener buenos regresores para la gran mayoría de antibióticos, lo cual sugiere que es necesario contar con más datos para poder predecir la actividad catalítica de las betalactamasas, o bien, explorar otras estrategias como el *fine-tunning*[P] de los modelos de lenguaje de proteínas. Con más datos también podría ser posible reevaluar la clasificación funcional de las betalactamasas, de tal forma que no solo se considere a las variantes de alto interés clínico. Para evaluar lo anterior, consideré el regresor de Cefoxitina y los cinco regresores de mediana calidad para realizar una agrupación jerárquica basada en el valor de las predicciones (Fig. sup. 30). Observé que las secuencias de la clase C se separan claramente de las clases A y D. Estas dos últimas clases también muestran una relativa buena agrupación, sin embargo, no es tan limpia como la de la clase C. Lo anterior sugiere que las clases de serinbetalactamasas se distinguen claramente por sus capacidades catalíticas contra distintos antibióticos betalactámicos. Además, sugiere que es posible crear un esquema de clasificación funcional basado en predicciones el cual puede mejorar con la inclusión de más datos.

[P] El término "*fine-tunning*" es un anglicismo que se usa para referir a una técnica de optimización usada en el área de procesamiento de lenguaje natural. Específicamente para el caso del modelado de proteínas, el *fine-tunning* consiste en entrenar el modelo de lenguaje pre-entrenado usando solo secuencias homologas de interés, permitiendo aprender al modelo una distribución de probabilidad especializada a dicho conjunto de homólogos, lo cual se traduce en una mejora de las capacidades predictivas como se demuestra en al referencia [85].

Discusión

RECAPITULACIÓN Y DESARROLLO

PARTE 1

La primera parte de mi tesis consistió en crear una estrategia de modelado basada en PLM, evaluar qué tipo de información se detecta y comparar los resultados contra la clasificación molecular y estimaciones fisicoquímicas de las secuencias. Para ello, usé los *embeddings* de siete PLM con distintas arquitecturas junto con tres estrategias: reducción de la dimensionalidad, métricas de distancia y *kmeans* (Fig. 5).

Respecto al desempeño de los PLM observé que ESM-1b y Prot-T5XL-U50 son los modelos que mejor distinguen a las clases de serin y metalobetalactamasas (Fig. 6-8 y 15-17). Por el contrario, el modelo Bepler fue el que peor identifica a dichos grupos, posiblemente, debido a que es un modelo con pocas capas de redes neuronales (Tab. 1). Recientemente el modelo CARP, ha sido propuesto como una alternativa competitiva a los *transformers*[77]. Sin embargo, mis resultados sugieren que CARP no es competitivo respecto a moldeos como ESM-1b en la identificación de grupos de betalactamasas, pues sus *embeddings* no contienen la suficiente información para separar a las clases A y D (Fig. 16 y 17). Lo anterior es interesante pues pese a que la forma en que se entrenó a CARP es muy similar a la de ESM-1b, no logra los mismos resultados. Esto podría deberse a que el sesgo inductivo que aportan los *transformers* es más adecuado para modelar información de secuencias, mientras que las redes neuronales convolucionales que usa CARP, aportan un sesgo inductivo orientado a la detección de patrones locales de distintas estructuras de datos.

La razón por la que el resto de los PLM (*i.e*. XLNet, ESM y Prot-T5-BFD) no cuentan con el mismo desempeño que ESM-1b y Prot-T5XL-U50 posiblemente recae en las variaciones de sus arquitecturas, esquemas de entrenamiento, cantidad y calidad de datos. XLNet cuenta con 409 millones de parámetros, mientras que ESM-1b y Prot-T5XL-U50 cuentan con 650 millones y 3,000 millones de parámetros, respectivamente (Tab. 1). Aunque en general se ha visto que aumentar el número de parámetros de un modelo se traduce en mejores resultados[89,93], más parámetros no siempre significa mejores resultados[76,92,95]. Por ejemplo, ESM-1b deriva de la optimización de ESM, al que aplicaron una serie de cambios en la arquitectura que resultaron en un modelo que aún con 20 millones de parámetros menos (*i.e.*, ESM-1b), supera a ESM[86]. Prot-T5-BFD tiene más parámetros respecto a Prot-T5XL-U50, sin embargo, Prot-T5-BFD fue entrenado con más secuencias redundantes, lo cual ocasionó un rendimiento ligeramente menor respecto a Prot-T5XL-U50 que fue entrenado con los mismos datos, y además, fue reajustado sobre un conjunto de datos de mayor calidad (*i.e*.

UniRef50)[98]. Lo anterior ilustra cómo la arquitectura, escala del modelo, cantidad y calidad de los datos guardan un balance en relación con el desempeño del modelo.

Respecto a la estrategia de reducción de la dimensionalidad, observé que los algoritmos que usan transformaciones no lineales (*i.e*. tSNE y UMAP) representan mejor la información codificada en los *embeddings*, pues identifican mejor a los grupos de betalactamasas. Por el contrario, los resultados con PCA no identifican claramente a dichos grupos (Fig. 6-8 y 15-17). Esto es esperado, pues tSNE y UMAP representan datos de una alta dimensión intentando preservar la estructura del vecindario en la baja dimensión[146,201,202]. Además, otros trabajos[56] también han recomendado usar métodos no lineales basados en espacios hiperbólicos para analizar *embeddings* de PLM. Sin embargo, una desventaja de tSNE y UMAP, además de que necesariamente distorsionan el espacio de la alta dimensión, es que sus resultados dependen mucho de valores de sus respectivos parámetros. Ante esto, es recomendable comparar representaciones derivadas de varios valores de parámetros que permitan identificar una configuración estable[140], siempre y cuando la cantidad de datos y capacidad de cómputo lo permita. En este sentido, comparé las representaciones obtenidas a partir de distintos parámetros y logré identificar grupos estables (Fig. sup. 2-4). Observé que, al considerar un mayor número de secuencias vecinas, *i.e*. un mayor número de perplexity o n_neighbors en tSNE o UMAP respectivamente, se obtienen mejores representaciones. Esto es interesante, pues estos parámetros suelen ser configurados con valores bajos, contrario a múltiples recomendaciones que sugieren que, para conjuntos de más de 10,000 datos es mejor considerar un mayor número de vecinos (Fig. sup. 2). Además, esta observación soporta los resultados Johnson *et al*.[130], quienes sugieren que es mejor considerar un mayor número de vecinos respecto a los valores comúnmente usados en otros conjuntos de datos fuera del dominio biológico (por ejemplo, MNIST).

Debido a que las representaciones de tSNE dependen de menos parámetros respecto a UMAP y, además, que la divergencia de Kullback-Leibler me permite saber su fidelidad, decidí analizar en detalle las representaciones con tSNE. Sin embargo, si hubiera trabajado con >100,000 secuencias, sería recomendable usar implementaciones de tSNE con heurísticas[53,203] o UMAP debido a que tSNE tiene alto coste de cómputo. También es importante considerar que las representaciones de tSNE son no euclidianas, a diferencia de PCA. Esto hace que la interpretación de las representaciones de tSNE sea cualitativa, es decir, si dos grupos se agrupan de forma discreta pero cercana entre sí, es porque tienen características similares en los datos de alta dimensionalidad, pero no necesariamente idénticas. Lo anterior se ejemplifica con el agrupamiento de los ancestros clase A y secuencias AFAM pese que muestran una identidad de secuencia ≥40% (Fig. 19B y Fig. sup. 15).

Para realizar análisis más detallados solo consideré a ESM-1b debido a sus buenos resultados, pues mapear los detalles de los seis modelos restantes habría extendido considerablemente el análisis. Tanto para serin y metalobetalactamasas, encontré que ESM-1b identifica varias propiedades biológicas (Tab. 2), sin embargo, la única que parece ser consistente son las clases de betalactamasas, distinguiendo claramente plegamientos que no son betalactamasas (Fig. sup. 9A y 9B). Por ejemplo, ESM-1b logra distinguir bien a las tres clases de serinbetalactamasas (Fig. 18A). Sin embargo, la clase A se dividió en tres grupos, contrario a los dos grupos que podríamos esperar correspondieran a sus

dos subclases[23], o a los seis grupos que podríamos esperar si fuera consistente con los grupos filogenéticos de Philippon *et al*.[15] (Fig. 21).

Para comprender el motivo por el que la clase A se dividió en tres grupos, comparé la similitud estructural y de secuencia entre un subconjunto de betalactamasas muestreadas al azar (Fig. 22). Observé que la cantidad de betalactamasas con alta (≥60%) o baja (<30%) similitud de secuencia favorece que las fuerzas atractivas y repulsivas de tSNE[143] agrupen o separen a las betalactamasas (Fig. 21 y 25B). Es decir, tSNE genera un efecto de "agrupación contingente". Este factor pudo haber ocasionado que las betalactamasas del grupo A1F (Firmicutes) se separaran del grupo principal (A01) pese a no tener diferencias marcadas a nivel estructural (Fig. 23). Una explicación similar podría aplicar a las betalactamasas del grupo A02, las cuales si tienen diferencias marcadas a nivel estructural y de secuencia. Dentro de este grupo, las betalactamasas del grupo A2P asociadas a la familia Sphingomonadaceae podrían representar una nueva subclase A3 debido a que las predicciones estructurales sugieren diferencias estructurales en su *loop* omega (Fig. sup. 16-19). Algo similar ocurre en la región cercana al sitio catalítico de una posible subclase C2 (Fig. sup. 12).

Mi análisis permitió identificar tres errores en la BLDB: 1) BLEG-1 qué es una serin proteasa (Fig. sup. 9A); 2) BCI-532 que es un canal iónico (Fig. sup. 9B) y; 3) siete secuencias OXA-D# que son proteínas de unión a penicilinas (Fig. 18A). En la representación de tSNE estas secuencias son puntos extremos en sus respectivas clases. Sin embargo, ESM-1b tiene conflictos al agrupar casos más difíciles como las fusiones entre distintas clases de betalactamasas (Fig. sup. 9C y 9D; Fig. sup. 14), o las fusiones entre betalactamasas y distintos plegamientos (Fig. sup. 9E). Además de haber sido extraño encontrar estos casos en una base de datos manualmente curada como la BLDB, estas fusiones son difíciles de resolver para ESM-1b. En la representación de tSNE, estas fusiones se agruparon dentro de una clase de betalactamasas, lo cual se debe a que el *embedding* si contiene información que deriva de una betalactamasa. Lo ideal sería que al contener información de dos plegamientos distintos en un mismo *embedding*, ESM-1b los hubiera detectado como puntos extremos tal como fue el caso para los plegamientos que no son betalactamasas. Posiblemente, estos casos de fusiones estarían ilustrando un caso de "ataques adversarios" donde se está engañando al modelo.

A nivel taxonómico observé una mala agrupación en las categorías taxonómicas salvo a nivel de género o especie. Sin embargo, ESM-1b podría no estar captando señales filogenéticas, sino que serían los límites filogenéticos los que estarían promoviendo una similitud de secuencia particular. Los límites filogenéticos influyen en la capacidad de recombinación y/o transferencia horizontal de los genes de betalactamasas, lo cual se estaría reflejando en la identificación de grupos claramente distinguibles en función de la categoría taxonómica (Fig. 11 y 20).

A nivel bioquímico, encontré que ciertas características son más fáciles de detectar respecto a otras. Por ejemplo, tanto para serin y metalobetalactamasas, la longitud de secuencia aporta una alta señal, mientras que el punto isoeléctrico aporta una baja señal (Fig. 12 y 24). Esto es interesante, la primer versión de la clasificación funcional considera al punto isoeléctrico en su clasificación[24,45]. Sin embargo, con la identificación de más variantes, el punto isoeléctrico y otras definiciones podrían ya no ser tan adecuadas[43,44,204]. Ademas, hay que considerar que no todas las similitudes entre proteínas

se deben a sus secuencias. Otras propiedades como factores fisicoquímicos, modificaciones postraduccionales o interacciones con proteínas y/o ligandos, también pueden afectar la similitud entre proteínas.

Al usar estrategias no basadas en reducción de la dimensionalidad observé que la calidad de la información de los *embeddings* varía según el PLM. Con ESM-1b observé un buen desempeño, sin embargo, al usar *kmeans* unas pocas secuencias no fueron correctamente clasificadas (Fig. 14 y 27). Esto puede deberse a que las distancias entre puntos en espacios de alta dimensión tienden a ser muy similares[79,143]. Además de que la preservación de las distancias en la alta dimensión no siempre se traduce en la preservación de la estructura del vecindario[143,146]. Por ello, tSNE y UMAP son el estándar para visualizar datos de alta dimensión, pues se enfocan en preservar la estructura del vecindario. El modelo Bepler es un ejemplo de lo anterior, pues con PCA no se logra separar a las clases de betalactamasas (Fig. 6 y 15), mientras que UMAP y tSNE logran una mejor separación respecto a PCA (Fig. 7, 8, 16 y 17).

En general, los resultados me permiten proponer como estrategia de modelado de proteínas el uso de ESM-1b (o Prot-T5XL-U50) junto con algoritmos no lineales de reducción de la dimensionalidad (como tSNE y UMAP). Y si no se cuenta con etiquetas asociadas a las secuencias, es posible usar kmeans y posiblemente otros métodos de agrupación no supervisada como el agrupamiento jerárquico o HDBSCAN. Se espera que los grupos identificados guarden similitudes estructurales que ayuden en la definición de subgrupos, como lo son las subclases, familias y subfamilias de betalactamasas.

Al evaluar la clasificación molecular de las betalactamasas con una nueva perspectiva (*i.e.* usando ESM-1b y predicciones estructurales) pude identificar al grupo A2P que podría representar una nueva subclase A3, así como secuencias representativas de una posible subclase C2. Considero que vale la pena realizar un esfuerzo para conseguir una estructura cristalográfica de este tipo de betalactamasas para corroborar su plegamiento, así como realizar análisis más detallados para determinar si se trata de nuevas subclases o no (Fig. 23; Fig. sup. 12).

Además, sugiero que la identificación de grupos de betalactamasas puede ser considerada como una prueba más en la evaluación de las capacidades de nuevos PLM[92,99]. Recomiendo usar estos nuevos algoritmos junto con herramientas bioinformáticas clásicas para enriquecer los análisis.

PARTE 2

La segunda parte de mi tesis consistió en evaluar los perfiles funcionales de las betalactamasas para tener una referencia, y posteriormente, entrenar regresores para predecir la actividad catalítica de betalactamasas cuyos perfiles funcionales no han sido caracterizados. Como esperaba, encontré una mayor cantidad de datos asociados a familias de alto interés clínico de la clase A (Fig. sup. 20), así como para cefalosporinas y penicilinas (Fig. 28). Estos datos son consistentes con los perfiles funcionales de las clases de betalactamasas reportados en la literatura (Fig. 30) y fueron mi referencia.

Al analizar la relación funcional y estructural de los antibióticos betalactámicos (Fig. 31 y 32), noté que las penicilinas tienen una considerable similitud estructural entre sí, la cual podría promover que sean un sustrato más común de ser degradado por betalactamasas respecto a las otras clases de antibióticos (Fig. 33). Además, observé algunas cefalosporinas de primer y segunda generación que muestran una ligera similitud con las penicilinas (Fig. 31), lo cual parece ser consistente con los casos reportados de reacciones alergénicas cruzadas entre estas clases de antibióticos[187,188].

La estrategia que seguí para el entrenamiento de los regresores puede ser considerado como un enfoque "proteín-céntrico" considerablemente reduccionista[205]. Sin embargo, existen otros esquemas que integran información genómica que podrían ser más apropiados, pues la resistencia contra antibióticos es un fenómeno emergente que involucra la actividad enzimática especifica, la fisiología del hospedero y su medio ambiente[206]. Sería ideal contar con un conjunto de datos del orden de varias centenas para cada uno de los antibióticos y que además estuvieran bien distribuidos en las clases de betalactamasas. Sin embargo, debido al sesgo de investigación a favor de variantes de interés clínico, este escenario parece complicado. Además de estos problemas, se suma la variabilidad de las condiciones experimentales con las que se colectaron los datos de MICs, por ejemplo, la cepa bacteriana específica, condiciones de cultivo (expresión génica), instrumentos de medición, etc. Estas condiciones no fueron consideradas en mi análisis debido a lo complejo que sería curar dichos datos para las 215 variantes de betalactamasas con las que trabajé (Fig. 29).

Un escenario ideal, sería que un mismo laboratorio midiera los valores de MICs usando la misma cepa bacteriana, bajo las mismas condiciones de cultivo para múltiples familias enzimáticas. Particularmente para la clase B, será importante medir las concentraciones de iones de Zinc. Sin embargo, dicho escenario ideal parece bastante costoso y logísticamente difícil de lograr. Aunque las betalactamasas son muy estudiadas, resulta sorprendente que apenas en 2022[32] se fijaron directrices de cómo y cuáles antibióticos hay que cuantificar. Con este nuevo régimen de colecta de datos podemos esperar un mejor escenario para la creación de modelos de aprendizaje automático.

Para lidiar con el problema de la escasez de datos, usé una estrategia de entrenamiento de aprendizaje por transferencia[56,79,82], la cual ha demostrado que es posible entrenar buenos regresores con unas cuantas decenas de datos etiquetados[83–85]. Para el entrenamiento solo consideré los datos de serinbetalactamasas, aunque en realidad las tres clases tienen variaciones en sus microambientes catalíticos[17–20]. Además, usé una estrategia de validación cruzada aleatoria y

estratificada la cual permite aminorar los sesgos durante el entrenamiento cuando los datos son muy escasos. Pese a todas estas consideraciones dentro de mi análisis, sólo logré entrenar un regresor de alta calidad del total de los 21 antibióticos considerados, lo cual sugiere que se necesitan más datos o que es necesario optimizar los modelos de lenguaje mediante *fine-tunning*.

Otra crítica que también se extiende a la primera parte de mi proyecto es la forma en que representé mis distintos objetos de estudio. Por ejemplo, para evaluar la similitud estructural de los antibióticos usé SMILES (Fig. 31). Esta forma de representación es considerada una convención en quimioinformática, sin embargo, existen otras representaciones como las SELFIES e incluso representaciones que derivan de moldeos de lenguaje que podrían ofrecer un mejor rendimiento[183]. En este sentido, la similitud de Tanimoto solo considera la información derivada de las SMILES, y además, se basa en la premisa de que dos antibióticos son similares si comparten entre sí muchos fragmentos en común. Sin embargo, la definición de similitud molecular es complicada, pues se puede optar por definiciones basadas en principios geométricos, farmacocinéticos, etcétera.

La convención para crear los *embeddings* es extraerlos de la última capa del PLM pues es donde más información codificada se encuentra[76,77,86,98]. Sin embargo, no necesariamente estos *embeddings* podrían ser los mejores en tareas predictivas. Igualmente, la creación de los *embeddings* por proteína suele hacerse mediante "average pooling" (Fig. 3). Sin embargo, existen otros esquemas capaces de resaltar residuos catalíticos que podrían tener un mejor desempeño en tareas predictivas[80,81]. Además, el reciente trabajo de Zeming *et al*.[91] sugiere que los PLM como ESM-2 tienen mejores capacidades predictivas con ciertos grupos de proteínas en función en función del número de homólogos que se hayan incluido durante el entrenamiento (ver Fig. 1B y 1C en la referencia [91]). Por lo tanto, no todos los plegamientos son igualmente modelados por los PLM. Por esta razón, podemos esperar que plegamientos abundantes (como las betalactamasas) presenten buenos resultados respecto a los plegamientos con pocos representantes.

Contrario a lo que esperaba, Prot-T5-BFD fue el PLM que mejor desempeño mostró en la tarea de regresión (Fig. sup. 29A), mientras que ESM-1b tuvo un mal desempeño con relación a Prot-T5-BFD. Lo anterior ilustra cómo los PLM pueden tener distintos desempeños en distintas tareas, pues en la primera parte de mi tesis Prot-T5-BFD no fue el PLM con mejores resultados, sino que fue ESM-1b. Esto sugiere que en función del objetivo de estudio y en medida de lo posible, es buena idea comparar distintos PLM y tomar el de mejor desempeño.

Considero que la clasificación funcional de betalactamasas necesita ser actualizada tomando en cuenta una mayor diversidad de secuencias y no solo aquellas de interés clínico. Una actualización a este esquema podría provenir de un análisis como el que realicé considerando los regresores de calidad media y alta (Fig. sup. 30), con lo cual observé que las clases de betalactamasas se pueden distinguir claramente de acuerdo con las predicciones de los valores de incremento de resistencia contra distintos antibióticos. Una mayor cantidad y calidad de los datos nos permitiría predecir el perfil funcional de las betalactamasas de mejor manera gracias al uso de PLM, el cual debe complementarse con otras técnicas bioinformáticas y experimentales para validar las predicciones.

PERSPECTIVAS

1. Realizar un *fine-tuning*[84,85] de ESM-1b usando secuencias no redundantes de las superfamilias *PBP-like* y metalobetalactamasas y repetir los análisis realizados para evaluar una posible mejora en el modelado de las betalactamasas.
2. Realizar análisis detallados con las secuencias y estructuras predichas de las betalactamasas de las posibles nuevas subclases A3 y C2 para evaluar los detalles que las harían ser nuevas subclases, los cuales deben ser validados mediante cristalografía de rayos X.
3. Entrenar un modelo de clasificación supervisada múltiple que permita predecir las etiquetas a nivel de clase, subclase, familia y subfamilia usando los *embeddings* de ESM-1b en conjunto con otras etiquetas como su clasificación taxonómica, propiedades fisicoquímicas, etc. Recomiendo que dicho modelo de clasificación sea comparado con otro no supervisado, por ejemplo, basado en modelos *autoencoders*, y comparar sus resultados contra otros modelos similares no basados en redes neuronales profundas[111,207].
4. Predecir las estructuras de todas las betalactamasas usando ESMFold[91] para poder realizar análisis de agrupamiento de estructuras con FoldSeek[208] y evaluar las posibles diferencias estructurales entre grupos y subgrupos de betalactamasas a nivel estructrual.
5. Entrenar modelos de regresión usando la base de datos que curé con datos de parámetros de cinética enzimática (kcat, km y kcat/km) y comparar los resultados con las predicciones realizadas usando los datos de MICs. Dichos modelos de regresión también pueden ser adaptados usando esquemas basados en teoría de grafos[209,210] en la representación de las estructuras de las betalactamasas.
6. Aplicar la estrategia de modelado de secuencias que propongo pero usando betalactamasas derivadas de la GTDB, las cuales se pueden identificar usando modelos ocultos de Márkov disponibles[26].
7. Explorar las secuencias con altos valores predichos usando el regresor de Cefoxitina y evaluar sus secuencias y estructuras en busca de firmas que puedan explicar su actividad catalítica.

Conclusiones

1. Los PLM han aprendido múltiples propiedades biológicas que les permitieron detectar los grupos de betalactamasas, lo cual sugiere que son herramientas útiles en la detección de grupos de proteínas que no han sido tan extensamente caracterizados.
2. La escasez de datos limitó la capacidad de entrenar buenos regresores, obteniéndose un solo regresor capaz de realizar predicciones consistentes con las literatura y subrayando la importancia de conjuntos de datos más abundantes para este tipo de tareas.
3. Hasta donde conozco, este es el primer trabajo en usar PLMs para modelar la clasificacion molecular y funcional de betalactamasas, el cual puede servir como referencia para investigaciones futuras.
4. No existe un mejor PLM, por lo que es recomendable comparar varios modelos pues su desempeño es variable en distintas tareas predictivas o de modelado.
5. Se sugiere que las penicilinas pueden ser un sustrato más fácil de ser degradado por betalactamasas debido a su considerable similitud estructural.

Referencias

1. Stokes, J. M., Lopatkin, A. J., Lobritz, M. A. & Collins, J. J. Bacterial Metabolism and Antibiotic Efficacy. *Cell Metab.* **30**, 251–259 (2019).

2. Van Boeckel, T. P. *et al.* Global antibiotic consumption 2000 to 2010: An analysis of national pharmaceutical sales data. *Lancet Infect. Dis.* **14**, 742–750 (2014).

3. Klein, E. Y. *et al.* Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3463–E3470 (2018).

4. Lima, L. M., Nascimento, B., Barbosa, G. & Barreiro, E. J. Betalactam antibiotics : An overview from a medicinal chemistry perspective. *Eur. J. Med. Chem.* **208**, 112829 (2020).

5. Abraham, E. P. & Chain, E. An enzyme from Bacteria able to destroy penicillin. *Nature* 3713 (1940).

6. Bush, K. & Patricia A. Bradfordb. Epidemiology of BetaLactamase-Producing Pathogens. *Clin. Microbiol. Rev.* **33**, 1–37 (2020).

7. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160**, 882–892 (2015).

8. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. *JACS* (2013).

9. Fröhlich, C., Chen, J. Z., Gholipour, S., Erdogan, A. N. & Tokuriki, N. Evolution of β-lactamases and enzyme promiscuity. *Protein Eng. Des. Sel.* (2021).

10. Ambler, R. P. The structure of beta-lactamases. *Philos. Trans. R. Soc. London* (1980) doi:10.1016/0009-2614(94)00453-6.

11. Bush, K. & Jacoby, G. A. Updated functional classification of β-lactamases. *Antimicrob. Agents Chemother.* **54**, 969–976 (2010).

12. Ronni, M. J., Al-Mahmeed, A., Fazal, K. D. & Mohammad, S. *Trends in Beta-Lactamase Classification. In: Beta-Lactam Resistance in Gram-Negative Bacteria*. (Springer, 2022).

13. Hall, B. G. & Barlow, M. Structure-based phylogenies of the serine β-lactamases. *J. Mol. Evol.* **57**, 255–260 (2003).

14. Hall, B. G., Salipante, S. J. & Barlow, M. The metallo-β-lactamases fall into two distinct phylogenetic groups. *J. Mol. Evol.* **57**, 249–254 (2003).

15. Philippon, A., Jacquier, H., Ruppé, E. & Labia, R. Structure-based classification of class A beta-lactamases, an update. *Curr. Res. Transl. Med.* **67**, 115–122 (2019).

16. Hall, B. G. & Barlow, M. Evolution of the serine betalactamases : past , present and future. *Drug Resist. Updat.* **7**, 111–123 (2004).

17. Mora-Ochomogo, M. & Lohans, C. T. β-Lactam antibiotic targets and resistance mechanisms: From covalent inhibitors to substrates. *RSC Medicinal Chemistry* vol. 12 1623–1639 at https://doi.org/10.1039/d1md00200g (2021).

18. Pratt, R. F. β-Lactamases: Why and How. *J. Med. Chem.* (2016).

19. Knox, J. R., Moews, P. C. & Frere, J. M. Molecular evolution of bacterial β-lactam resistance. *Chem. Biol.* **3**, 937–947 (1996).

20. Golemi, D., Maveyraud, L., Vakulenko, S., Samama, J. P. & Mobashery, S. Critical involvement of a carbamylated lysine in catalytic function of class D β-lactamases. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14280–14285 (2001).

21. Massova, I. & Mobashery, S. Kinship and diversification of bacterial penicillin-binding proteins and β-lactamases. *Antimicrob. Agents Chemother.* **42**, 1–17 (1998).

22. Meroueh, S. O., Minasov, G., Lee, W., Shoichet, B. K. & Mobashery, S. Structural aspects for evolution β-lactamases from penicillin-binding proteins. *J. Am. Chem. Soc.* **125**, 9612–9618 (2003).

23. Philippon, A., Slama, P., Dény, P. & Labia, R. A structure-based classification of class A β-Lactamases, a broadly diverse family of enzymes. *Clin. Microbiol. Rev.* **29**, 29–57 (2016).

24. Yoon, E.-J. & Jeong, S. H. Class D β-lactamases. *J.*

Antimicrob. Chemother. (2021).

25. Philippon, A., Arlet, G. & Labia, R. Class C b-Lactamases : Molecular Characteristics. *Clin. Microbiol. Rev.* (2022).

26. Silveira, M. C. *et al.* Systematic Identification and Classification of β-Lactamases Based on Sequence Similarity Criteria: β-Lactamase Annotation. *Evol. Bioinforma.* **14**, (2018).

27. Lopez, C. et al. Deciphering the evolution of metallo-β-lactamases A journey from the test tube to the bacterial periplasm. *JBC Rev.* (2022).

28. Dideberg, O. Is it necessary to change the classification of b -lactamases ? *J. Antimicrob. Chemother.* 1051–1053 (2005) doi:10.1093/jac/dki155.

29. Hall, B. G. & Barlow, M. Revised Ambler classification of β-lactamases. *J. Antimicrob. Chemother.* **55**, 1050–1051 (2005).

30. Hall, B. G., Salipante, S. J. & Barlow, M. Independent Origins of Subgroup Bl + B2 and Subgroup B3 Metallo-b -Lactamases. **3**, 133–141 (2004).

31. Roanna, G. A. & Baker, D. One origin for metallo-β-lactamase activity, or two An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees. *J. Mol. Evol.* (2014).

32. Bradford, P. A. *et al.* Consensus on betalactamase nomenclature. *Antimicrob. Agents Chemother.* (2022).

33. Tooke, C. L. *et al.* β-Lactamases and β-Lactamase Inhibitors in the 21st Century. *J. Mol. Biol.* **431**, 3472–3500 (2019).

34. Christensen, H., Martin, M. T. & Waley, S. G. β-Lactamases as fully efficient enzymes. Determination of all the rate constants in the acyl-enzyme mechanism. *Biochem. J.* **266**, 853–861 (1990).

35. Bulychev, A. & Mobashery, S. Class C β-lactamases operate at the diffusion limit for turnover of their preferred cephalosporin substrates. *Antimicrob. Agents Chemother.* **43**, 1743–1746 (1999).

36. Davidi, D., Longo, L. M., Jabłońska, J., Milo, R. & Tawfik, D. S. A Bird's-Eye View of Enzyme Evolution: Chemical, Physicochemical, and Physiological Considerations. *Chem. Rev.* **118**, 8786–8797 (2018).

37. Garau, G. *et al.* Update of the standard numbering scheme for class B β-lactamases. *Antimicrob. Agents Chemother.* **48**, 2347–2349 (2004).

38. Mack, A. R. *et al.* A standard numbering scheme for class C β-lactamases. *Antimicrobial Agents and Chemotherapy* vol. 64 at https://doi.org/10.1128/AAC.01841-19 (2020).

39. Matagne, A. *et al.* The diversity of the catalytic properties of class A β-lactamases. *Biochem. J.* **265**, 131–146 (1990).

40. Keshri, V. & et. al. The functional convergence of antibiotic resistance in β- lactamases is not conferred by a simple convergent substitution of amino acid. *Evol. Appl.* (2019).

41. Giske, C. G. *et al.* Redefining extended-spectrum β-lactamases: Balancing science and clinical need. *J. Antimicrob. Chemother.* **63**, 1–4 (2009).

42. Bush, K. *et al.* Comment on: Redefining extended-spectrum β-lactamases: Balancing science and clinical need. *J. Antimicrob. Chemother.* **64**, 212–213 (2009).

43. Bush, K. The ABCD's of β-lactamase nomenclature. *J. Infect. Chemother.* **19**, 549–559 (2013).

44. Livermore, D. M. Defining an extended-spectrum β-lactamase. *Clin. Microbiol. Infect.* **14**, 3–10 (2008).

45. Bush, K. Past and Present: Perspectives on betalactamases. *Antimicrobial* 1–20 (2018).

46. Brandt, C. & Al, E. In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. *Sci. Rep.* (2017).

47. Maatouk, M. *et al.* New Beta-lactamases in Candidate Phyla Radiation : Owning Pleiotropic Enzymes Is a Smart Paradigm for Microorganisms with a Reduced Genome. *Int. J. Mol. Sci.* (2022).

48. Ambler, R. P. & Al., E. A standard numbering scheme for the Class A betalactamases. *Biochem. J.* **276**, 1990–1991 (1991).

49. Naas, T. *et al.* Beta-lactamase database (BLDB) – structure and function. *J. Enzyme Inhib. Med. Chem.* **0**, 917–919 (2017).

50. Kandathil, S. M. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **0123456789**,.

51. Eckmann, J. P. & Tlusty, T. Dimensional reduction in

complex living systems: Where, why, and how. *BioEssays* **43**, 1–10 (2021).

52. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* **15**, 399–400 (2018).

53. Nguyen, L. H. & Holmes, S. Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* **15**, 1–19 (2019).

54. Wang, S., Sontag, E. D. & Lauffenburger, D. A. What cannot be seen correctly in 2D visualizations of single-cell 'omics data? *Cell Syst.* **14**, 723–731 (2023).

55. Parasa, N. A., Namgiri, J. V., Mohanty, S. N. & Dash, J. K. Introduction to Unsupervised Learning in Bioinformatics. *Data Anal. Bioinforma. A Mach. Learn. Perspect.* 35–49 (2021) doi:10.1002/9781119785620.ch2.

56. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nat. Commun.* 1–12 (2022) doi:10.1038/s41467-022-29443-w.

57. Littmann, M. *et al.* Clustering FunFams using sequence embeddings improves EC purity. *Bioinformatics* **37**, 3449–3455 (2021).

58. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

59. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).

60. AlQuraishi, M. & Sorger, P. K. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. Methods* **18**, 1169–1180 (2021).

61. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**-**Decem**, 5999–6009 (2017).

62. Amatriain, X. Transformer models: an introduction and catalog. *arXiv* (2023).

63. Vu, M. H. *et al.* Linguistically inspired roadmap for building biologically reliable protein language models. *Nat. Mach. Intell.* 1–26 (2023).

64. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654-669.e3 (2021).

65. Ofer, D., Brandes, N. & Linial, M. The Language of Proteins : NLP , Machine Learning & Protein Sequences. *Comput. Struct. Biotechnol. J.* (2021) doi:10.1016/j.csbj.2021.03.022.

66. Iuchi, H. *et al.* Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **19**, 3198–3208 (2021).

67. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

68. Dallago, C., Goldman, S., Bhattacharya, N., Madani, A. & Yang, K. K. FLIP : Benchmark tasks in fitness landscape inference for proteins. *NeurIPS 2021 Datasets Benchmarks* (2021).

69. Peter Mørch Groth, Richard Michael, Pengfei Tian, Jesper Salomon, W. B. FLOP Tasks for Fitness Landscapes Of Protein families using sequence- and structure-based representations. *ICLR 2023 Conf.* (2022).

70. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *arXiv* 1–20 (2019).

71. Notin, P. *et al.* Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *Proc. Mach. Learn. Res.* (2022).

72. Ferruz, N. & Höcker, B. Towards Controllable Protein design with Conditional Transformers. *arx* 1–17 (2022).

73. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, (2022).

74. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-022-01618-2.

75. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. *arXiv* (2020) doi:10.1101/2020.06.26.174417.

76. Brandes, N. & et al. ProteinBERT : a universal deep-learning model of protein sequence and function. *Bioinformatics* 1–9 (2022).

77. Yang, K. K., Lu, A. X. & Fusi, N. Convolutions are competitive with transformers for protein sequence pretraining. doi:10.1101/2022.05.19.492714.

78. Teufel, F. *et al.* SignalP 6.0 predicts all five types of

signal peptides using protein language models. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-021-01156-3.

79. Fenoy, E., Edera, A. A. & Stegmayer, G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Brief. Bioinform.* **23**, 1–19 (2022).

80. Goldman, S., Das, R., Yang, K. K. & Coley, C. W. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS Comput. Biol.* **18**, 1–20 (2022).

81. Yamaguchi, H. & Saito, Y. Evotuning protocols for Transformer-based variant effect prediction on multi-domain proteins. *Brief. Bioinform.* **22**, 1–9 (2021).

82. Ferguson, A. L. & Ranganathan, R. Data-Driven Protein Design. *ACS Macro Lett.* (2021) doi:10.1021/acsmacrolett.0c00885.

83. Biswas, S. *et al.* Toward machine-guided design of proteins. *bioRxiv* 1–10 (2018) doi:10.1101/337154.

84. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).

85. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).

86. Rao, R. & et al. Transformer protein language models are unsupervised structure learners. *bioRxiv* 1–24 (2020).

87. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021).

88. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

89. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. (2021) doi:10.1101/2022.07.20.500902.

90. Hesslow, D. & et. al. RITA : a Study on Scaling Up Generative Protein Sequence Models. *arXiv* (2022).

91. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *Science (80-.).* **1130**, 2022.07.20.500902 (2023).

92. Elnaggar, A. *et al.* Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv* 1–29 (2023).

93. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. *arXiv* (2020).

94. Chowdhery, A. *et al.* PaLM : Scaling Language Modeling with Pathways. *arXiv* 1–83 (2022).

95. Hoffmann, J. *et al.* Training Compute-Optimal Large Language Models. (2022).

96. Birhane, A., Prabhu, V. U. & Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv* (2021).

97. Chen, B. *et al.* xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. *bioRxiv* 2023.07.05.547496 (2023).

98. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**, 1–17 (2021).

99. Buchfink, B., Ashkenazy, H., Reuter, K., Kennedy, J. A. & Drost, H.-G. Sensitive clustering of protein sequences at tree-of-life scale using DIAMOND DeepClust. *bioRxiv* 1–35 (2023).

100. Bepler, T. & Berger, B. Learning protein sequence embeddings using information from structure. *7th Int. Conf. Learn. Represent. ICLR 2019* 1–17 (2019).

101. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–46 (2021).

102. Keshri, V. & et al. An Integrative Database of Betalactamase Enzymes. *Antimicrob. Agents Chemother.* **63**, 1–8 (2019).

103. Srivastava, A., Singhal, N., Goel, M., Virdi, J. S. & Kumar, M. CBMAR: A comprehensive b-lactamase molecular annotation resource. *Database* **2014**, 1–8 (2014).

104. Wang, C. Y. *et al.* ProtaBank : A repository for protein design and engineering data. *Protein Sci.* **27**, 1113–1124 (2018).

105. Schug, A., Tenaillon, O., Weigt, M. & Figliuzzi, M. Coevolutionary Landscape Inference and the Context-

Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2015).

106. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116-124.e3 (2018).

107. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

108. Verkuil, R. *et al.* Language models generalize beyond natural proteins. *bioRxiv* 2022.12.21.521521 (2022).

109. Lundstrøm, J., Korhonen, E., Lisacek, F. & Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin–Glycan Binding Prediction. *Adv. Sci.* **9**, 1–16 (2022).

110. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with predictions of protein-protein interactions. *Cell Syst.* 1–14 (2021) doi:10.1016/j.cels.2021.08.010.

111. Pandey, D., Singhal, N. & Kumar, M. β-LacFamPred: An online tool for prediction and classification of β-lactamase class, subclass, and family. *Frontiers in Microbiology* vol. 13 at https://doi.org/10.3389/fmicb.2022.1039687 (2023).

112. Cohen, R. D. & Pielak, G. J. A cell is more than the sum of its (dilute) parts: A brief history of quinary structure. *Protein Sci.* **26**, 403–413 (2017).

113. Risso, V. A., Gavira, J. A., Gaucher, E. A. & Sanchez-Ruiz, J. M. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins Struct. Funct. Bioinforma.* **82**, 887–896 (2014).

114. Risso, V. A. *et al.* De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.* **8**, 1–13 (2017).

115. Modi, T. & et al. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. *Nat. Commun.* (2021).

116. Dallago, C. *et al.* Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Curr. Protoc.* **1**, 1–26 (2021).

117. Sternke, M., Tripp, K. W. & Barrick, D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U. S. A.* **166**, 11275–11284 (2019).

118. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

119. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

120. Bond, S. R., Keat, K. E., Barreira, S. N. & Baxevanis, A. D. BuddySuite: Command-line toolkits for manipulating sequences, alignments, and phylogenetic trees. *Mol. Biol. Evol.* **34**, 1543–1546 (2017).

121. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

122. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, 1–10 (2016).

123. Dorfer, T. protlearn: A Python package for extracting protein sequence features. at https://github.com/tadorfer/protlearn (2020).

124. Kleikamp, H. B. C. *et al.* Comparative metaproteomics demonstrates different views on the complex granular sludge microbiome. *bioRxiv* 2022.03.07.483319 (2022).

125. Parks, D. H. *et al.* GTDB : an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent , rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **202**, 1–10 (2021).

126. Buchfink, B., Reuter, K. & Drost, H. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, (2021).

127. Lindeløv, J. K. Common statistical tests are linear models. *Https://Lindeloev.Github.Io/Tests-As-Linear/* 2019 at https://lindeloev.github.io/tests-as-linear/ (2019).

128. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648 (2018).

129. Chari, T. & Pachter, L. The specious art of single-cell

genomics. *PLoS Comput. Biol.* **19**, 1–20 (2023).

130. Johnson, E. M., Kath, W. & Mani, M. EMBEDR: Distinguishing signal from noise in single-cell omics data. *Patterns* **3**, 100443 (2022).

131. Yeung, W. *et al.* Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Brief. Bioinform.* **24**, 1–10 (2023).

132. Yang, Y. *et al.* Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep.* **36**, (2021).

133. Huang, H., Wang, Y., Rudin, C. & Browne, E. P. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun. Biol.* **5**, (2022).

134. Schütze, K., Heinzinger, M., Steinegger, M. & Rost, B. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Front. Bioinforma.* **2**, 1–12 (2022).

135. Michael, R. *et al.* Assessing the performance of protein regression models. *bioRxiv* 2023.06.18.545472 (2023).

136. Hie, B. L. *et al.* Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* 1–12 (2022) doi:10.1016/j.cels.2022.01.003.

137. Bhattacharya, N. Single Layers of Attention Suffice to Predict Protein Contacts. *ICLR 2021 Conf.* 1–25 (2020).

138. Schreiber, A. Transformers, proteins, and persistent homology. 1–38 (2023).

139. Powel, V. Principal Component Analysis: Explained Visually. at https://doi.org/https://setosa.io/ev/principal-component-analysis/ (2015).

140. Wattenberg, M., Viégas, F. & Johnson, I. How to Use t-SNE Effectively. at https://doi.org/http://doi.org/10.23915/distill.00002 (2016).

141. Coenen, A. & Pearce, A. Understanding UMAP. at https://doi.org/https://pair-code.github.io/understanding-umap/.

142. Heiser, C. N. & Lau, K. S. A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Rep.* **31**, 107576 (2020).

143. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding how dimension reduction tools work: An empirical approach to deciphering T-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J. Mach. Learn. Res.* **22**, 1–73 (2021).

144. Oskolkov, N. tSNE vs. UMAP: Global structure. at https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17 (2019).

145. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).

146. Damrich, S., Böhm, J. N., Hamprecht, F. A. & Kobak, D. Contrastive learning unifies t-SNE and UMAP. *arXiv* (2022).

147. Oskolkov, N. How to tune hyperparameters of tSNE. 1–19 at https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868 (2019).

148. Belkina, A. C. *et al.* Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.* **10**, 1–12 (2019).

149. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, (2019).

150. Zhou, H., Wang, F. & Tao, P. T-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *J. Chem. Theory Comput.* **14**, 5499–5510 (2018).

151. Ata, R. E. D., Grabski, I. N., Street, K. & Irizarry, R. A. Significance analysis for clustering with single-cell RNA-sequencing data. *Nat. Methods* 1–13 (2023).

152. Porebski, B. T. & Buckle, A. M. Consensus protein design. *Protein Eng. Des. Sel.* **29**, 245–251 (2016).

153. Sternke, M., Tripp, K. W. & Barrick, D. *The use of consensus sequence information to engineer stability and activity in proteins*. *Methods in Enzymology* vol. 643 (Elsevier Inc., 2020).

154. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).

155. Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T. & Poon, A. F. Y. Ancestral Reconstruction. *PLoS Comput. Biol.* **12**, 1–20 (2016).

156. Grootendorst, M. 9 Distance Measures in Data Science. *Towards Data Science* at https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa (2021).

157. Lin & et. al. The Vibrio cholerae var regulon encodes a metallo-β-lactamase and an antibiotic efflux pump, which are regulated by VarR, a LysRtype transcription factor. *PLoS One* (2017).

158. Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A. & Handelsman, J. Functional metagenomics reveals diverse B-lactamases in a remote Alaskan soil. *ISME J.* **3**, 243–251 (2009).

159. Unsal, S. *et al.* Learning functional properties of proteins with language models. *Nat. Mach. Intell.* **4**, 227–245 (2022).

160. Hu, M. *et al.* Exploring evolution-aware & -free protein language models as protein function predictors. *arXiv* (2022).

161. Au, S. X. *et al.* Dual activity bleg-1 from bacillus lehensis g1 revealed structural resemblance to b3 metallo-β-lactamase and glyoxalase ii: An insight into its enzyme promiscuity and evolutionary divergence. *Int. J. Mol. Sci.* **22**, 1–21 (2021).

162. Mirdita, M., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* 2021.08.15.456425 (2022).

163. Lu, S. *et al.* An active site loop toggles between conformations to control antibiotic hydrolysis and inhibition potency for CTX-M β-lactamase drug-resistance enzymes. *Nature Communications* vol. 13 at https://doi.org/10.1038/s41467-022-34564-3 (2022).

164. Pares, S. & et. al. X-ray structure of Streptococcus pneumoniae PBP2x, a primary penicillin target enzyme. *Nat. Struct. Mol. Biol.* **2**, 534–539 (1996).

165. Goddard, T. D. *et al.* UCSF ChimeraX : Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

166. Silveira, M. C., Catanho, M. & de Miranda, A. B. Genomic analysis of bifunctional class C-class D β-lactamases in environmental bacteria. *Mem. Inst. Oswaldo Cruz* **113**, (2018).

167. Salverda, M. L. M., de Visser, J. A. G. M. & Barlow, M. Natural evolution of TEM-1 β-lactamase: Experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev.* **34**, 1015–1036 (2010).

168. Risso, V. A. & Sanchez-Ruiz, J. M. Resurrected Ancestral Proteins as Scaffolds for Protein Engineering. in *Directed Enzyme Evolution: Advances and Applications* 1–284 (2017). doi:10.1007/978-3-319-50413-1.

169. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

170. Yi, H. *et al.* Twelve Positions in a b -Lactamase That Can Expand Its Substrate Spectrum with a Single Amino Acid Substitution. *PLoS One* **7**, (2012).

171. Tranier, S. *et al.* The high resolution crystal structure for class A β-lactamase PER-1 reveals the bases for its increase in breadth of activity. *J. Biol. Chem.* **275**, 28075–28082 (2000).

172. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **0123456789**, (2022).

173. Li, Y. *et al.* Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting b-Lactam Resistance Levels in Streptococcus pneumoniae. *MBio* (2016) doi:10.1128/mBio.00756-16.Editor.

174. Valizadehaslani, T., Zhao, Z., Sokhansanj, B. A. & Rosen, G. L. Amino Acid k-mer Feature Extraction for Quantitative Antimicrobial Resistance (AMR) Prediction by Machine Learning and Model Interpretation for Biological Insights. *Biol.* (2020) doi:10.3390/biology9110365.

175. Meletis, G. Carbapenem resistance: overview of the problem and future perspectives. *Ther. Adv. Infect. Dis.* **3**, 15–21 (2016).

176. Gin, A. *et al.* Piperacillin-tazobactam: A β-lactam/β-lactamase inhibitor combination. *Expert Rev. Anti. Infect. Ther.* **5**, 365–383 (2007).

177. Bush, K., Macalintal, C., Rasmussen, B. A., Lee, V. J. & Yang, Y. Kinetic interactions of tazobactam with β-lactamases from all major structural classes. *Antimicrob. Agents Chemother.* **37**, 851–858 (1993).

178. Antunes, N. T. & Fisher, J. F. Acquired class D β-Lactamases. *Antibiotics* **3**, 398–434 (2014).

179. Jamal, W. Y., Albert, M. J. & Rotimi, V. O. High prevalence of New Delhi metallo-β-lactamase-1 (NDM-1) producers among carbapenem-resistant Enterobacteriaceae in Kuwait. *PLoS One* **11**, 1–12 (2016).

180. Drawz, S. M. & Bonomo, R. A. Three decades of β-lactamase inhibitors. *Clin. Microbiol. Rev.* **23**, 160–201 (2010).

181. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).

182. RDKit Open-source cheminformatics. at httpwww.rdkit.org.

183. Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1–19 (2022) doi:10.1002/wcms.1603.

184. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).

185. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).

186. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Syst.* **11**, 461-476.e10 (2020).

187. Committee, U. P. T. *Allergic Cross-reactivity of Select Antimicrobials*. *UCDH Pharmacy Therapeutics Committee* vol. 27 https://health.ucdavis.edu/antibiotic-stewardship/pdfs/abx_cross_reactivity.pdf (2017).

188. Caruso, C., Valluzzi, R. L., Colantuono, S., Gaeta, F. & Romano, A. β-Lactam Allergy and Cross-Reactivity: A Clinician's Guide to Selecting an Alternative Antibiotic. *Journal of Asthma and Allergy* vol. 14 31–46 at https://doi.org/10.2147/JAA.S242061 (2021).

189. Chaudhry, S. B. & Veve, M. P. Cephalosporins : A Focus on Side Chains and. *Pharmacy* vol. 7 1–16 at (2019).

190. Pilmis, B. *et al.* No significant difference between ceftriaxone and cefotaxime in the emergence of antibiotic resistance in the gut microbiota of hospitalized patients: A pilot study. *Int. J. Infect. Dis.* **104**, 617–623 (2021).

191. Smith, C. R. *et al.* Ceftriaxone compared with cefotaxime for serious bacterial infections. *J. Infect. Dis.* **160**, 442–447 (1989).

192. Pouwels, K. B. *et al.* Association between use of different antibiotics and trimethoprim resistance: Going beyond the obvious crude association. *J. Antimicrob. Chemother.* **73**, 1700–1707 (2018).

193. Laroche, E., Pawlak, B., Berthe, T., Skurnik, D. & Petit, F. Occurrence of antibiotic resistance and class 1, 2 and 3 integrons in Escherichia coli isolated from a densely populated estuary (Seine, France). *FEMS Microbiol. Ecol.* **68**, 118–130 (2009).

194. Landesman, S. H., Corrado, M. L., Cherubin, C. E. & Sierra, M. F. Activity of moxalactam and cefotaxime alone and in combination with ampicillin or penicillin against group B streptococci. *Antimicrob. Agents Chemother.* **19**, 794–797 (1981).

195. Pandala, S. R. LazyPredict. https://github.com/shankarpandala/lazypredict (2019).

196. Pedregosa., F. & Al, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2011).

197. Armenteros, J. J. A., Johansen, A. R., Winther, O. & Nielsen, H. Language modelling for biological sequences – curated datasets and baselines. *bioRxiv* 1–8 (2020).

198. Preston, K. E., Radomski, C. C. A. & Venezia, R. A. Nucleotide sequence of the chromosomal ampC gene of Enterobacter aerogenes. *Antimicrob. Agents Chemother.* **44**, 3158–3162 (2000).

199. Ye, Y., Xu, X. H. & Li, J. Bin. Emergence of CTX-M-3, TEM-1 and a new plasmid-mediated MOX-4 AmpC in a multiresistant Aeromonas caviae isolate from a patient with pneumonia. *J. Med. Microbiol.* **59**, 843–847 (2010).

200. Ingti, B. *et al.* Molecular and in silico analysis of a new plasmid-mediated AmpC β-lactamase (CMH-2) in clinical isolates of Klebsiella pneumoniae. *Infect. Genet. Evol.* **48**, 34–39 (2017).

201. Venna, J., Kaski, S., Aidos, H., Nybo, K. & Peltonen, J. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.* **11**, 451–490 (2010).

202. Corin Wagen. Dimensionality Reduction in Cheminformatics. https://corinwagen.github.io/public/blog/20230417_

dimensionality_reduction.html (2023).

203. Poličar, P. G., Stražar, M. & Zupan, B. OpenTSNE: A modular Python library for t-SNE dimensionality reduction and embedding. *bioRxiv* 1–2 (2019) doi:10.1101/731877.

204. Verma, D., Jacobs, D. J. & Livesay, D. R. Variations within Class-A β-Lactamase Physiochemical Properties Reflect Evolutionary and Environmental Patterns, but not Antibiotic Specificity. *PLoS Comput. Biol.* **9**, (2013).

205. Lau, H. J., Lim, C. H., Foo, S. C. & Tan, H. S. The role of artificial intelligence in the battle against antimicrobial-resistant bacteria. *Curr. Genet.* **67**, 421–429 (2021).

206. Socha, R. D., Chen, J. & Tokuriki, N. The Molecular Mechanisms Underlying Hidden Phenotypic Variation among Metallo-β-Lactamases. *Journal of Molecular Biology* vol. 431 1172–1185 at https://doi.org/10.1016/j.jmb.2019.01.041 (2019).

207. White, C., Ismail, H. D., Saigo, H. & KC, D. B. CNN-BLPred: A Convolutional neural network based predictor for β-Lactamases (BL) and their classes. *BMC Bioinformatics* **18**, (2017).

208. Kempen, M. van *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022).

209. Shroff, R. *et al.* Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* (2020) doi:10.1021/acssynbio.0c00345.

210. Viñas, R. *et al.* Graphein - a Python Library for Geometric Deep Learning and Network Analysis on Protein Structures and Interaction Networks. *bioRxiv* 1–15 (2021).

211. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

Material Suplementario

SOFTWARE Y HARDWARE EMPLEADO

Tabla suplementaria 1. Software empleado para el procesamiento, análisis y visualización de los datos.

Nombre	Descripción de uso	Ref.
SeqKit	Manipulación y procesamiento de secuencias	122
Bio_embeddings	Uso de lenguajes de proteínas, t-SNE y UMAP	116
ChimeraX	Visualización de proteínas	165
ColabFold	Predicción *de novo* de proteínas usando AlphaFold2 y ESMFold	67,89,162
BioPandas	Manipulación de archivos PDB	
Biopython y ProtLearn	Cómputo de características fisicoquímicas derivadas de secuencia	121,123
Mafft	Creación de alineamientos múltiples de secuencia. Tanto con el servidor web, así como con la plataforma MPI Bioinformatics Toolkit.	119
Cd-hit	Cómputo de secuencias representativas a un porcentaje de identidad	118
Diamond2	Alineamiento de secuencias eficiente	126
EVCouplings	Estimación de la Identidad de secuencia a partir de un alineamiento	211
TMalign	Alineamiento estructural de proteínas	169
GTDB2DIAMOND	Conjunto de scripts para la asignación de taxonomía basado en alineamientos y último ancestro en común	124
Python (Librerías)	Manipulación, procesamiento y visualización de datos: Pytorch, Numpy, Scipy, SciKit-Learn, statannotations, Pandas, Matplotlib, Seaborn, Os, Re, Time, Joypy, mpl_sankey, 3Dblocks, Plotly, HoloViews, itertools, Conda, Jupyter y mlxtend	
Lazypredict	Comparación de algoritmos de regresión	195
BuddySuite	Construcción de secuencias consenso	120
RDKit	Cómputo de los *fingerprints* de Morgan y similitud de Tanimoto	182

Tabla suplementaria 2. Hardware empleado para el procesamiento, análisis y visualización de los datos

Componente	FOS – IBt	TIM – IBt	GAMA – Local
Procesador	(2x) Intel Xeon E5-2680 v4	Intel i7 9ª generación	Ryzen 7 Pro 4750G
RAM	500GB	32GB	32GB
Almacenamiento	12TB	2TB	3.25TB
GPU	N/A	RTX 2060 Super	RTX 3060

REPOSITORIOS EN GITHUB

RECURSOS DIDÁCTICOS

Como parte de mi tesis de maestría me di a la tarea de facilitar el acceso a personas interesadas en aprender ciencia de proteínas basada en inteligencia artificial. Para ello, creé el siguiente repositorio el cual contiene distintos recursos para introducirse al tema.

Repositorio: https://github.com/miangoar/ciencia-de-proteinas-basada-en-IA

Contenido:

1. Seminarios: serie de videos serie de videos en YouTube que preparé para introducirse al tema de
    a) evolución de proteínas
    b) herramientas útiles para el procesamiento de secuencias y estructuras proteínas con énfasis en aprendizaje automático
    c) trabajos contemporáneos destacados de ciencia de proteínas basada en aprendizaje automático
    d) hardware y software
    e) aprendizaje automático y aprendizaje profundo
    f) modelos de lenguaje
    g) sesgos que tiene la ciencia de proteínas basada en inteligencia artificial
2. Videos recomendados: serie de videos en YouTube para introducirse al tema de la evolución molecular y de proteínas, así como de ciencia de proteínas basada en inteligencia artificial.
3. Tutoriales: Jupyter notebooks escritos por mi u otras personas para manejar herramientas relacionadas a la ciencia de proteínas basada en inteligencia artificial.
4. Herramientas recomendadas: serie de herramientas bioinformáticas que personalmente recomiendo para la realización de ciencia de proteínas basada en inteligencia artificial
5. Repositorios útiles: serie de repositorios donde acceder a más herramientas de ciencia de proteínas basada en inteligencia artificial
6. Webservers: serie de páginas recomendadas que ofrecen análisis bioinformáticos de ciencia de proteínas basada en inteligencia artificial sin necesidad de escribir código.
7. Recursos de aprendizaje: serie de páginas que recomiendo con cursos completos para aprender ciencia de proteínas, ciencia de datos y aprendizaje automático.
8. Literatura recomendada: serie de artículos curados que personalmente recomiendo para introducirse al tema de evolución molecular y de proteínas, así como de ciencia de proteínas basada en inteligencia artificial

CÓDIGO DE LA PRESENTE TESIS

En el siguiente repositorio se encuentra todo el código que escribí en formato *Jupyter notebook* para el procesamiento, análisis y visualización de los datos de betalactamasas, así como las bases de datos que curé y construí.

Repositorio:

https://github.com/miangoar/protein_language_models_for_betalactamases_analysis

FIGURAS SUPLEMENTARIAS

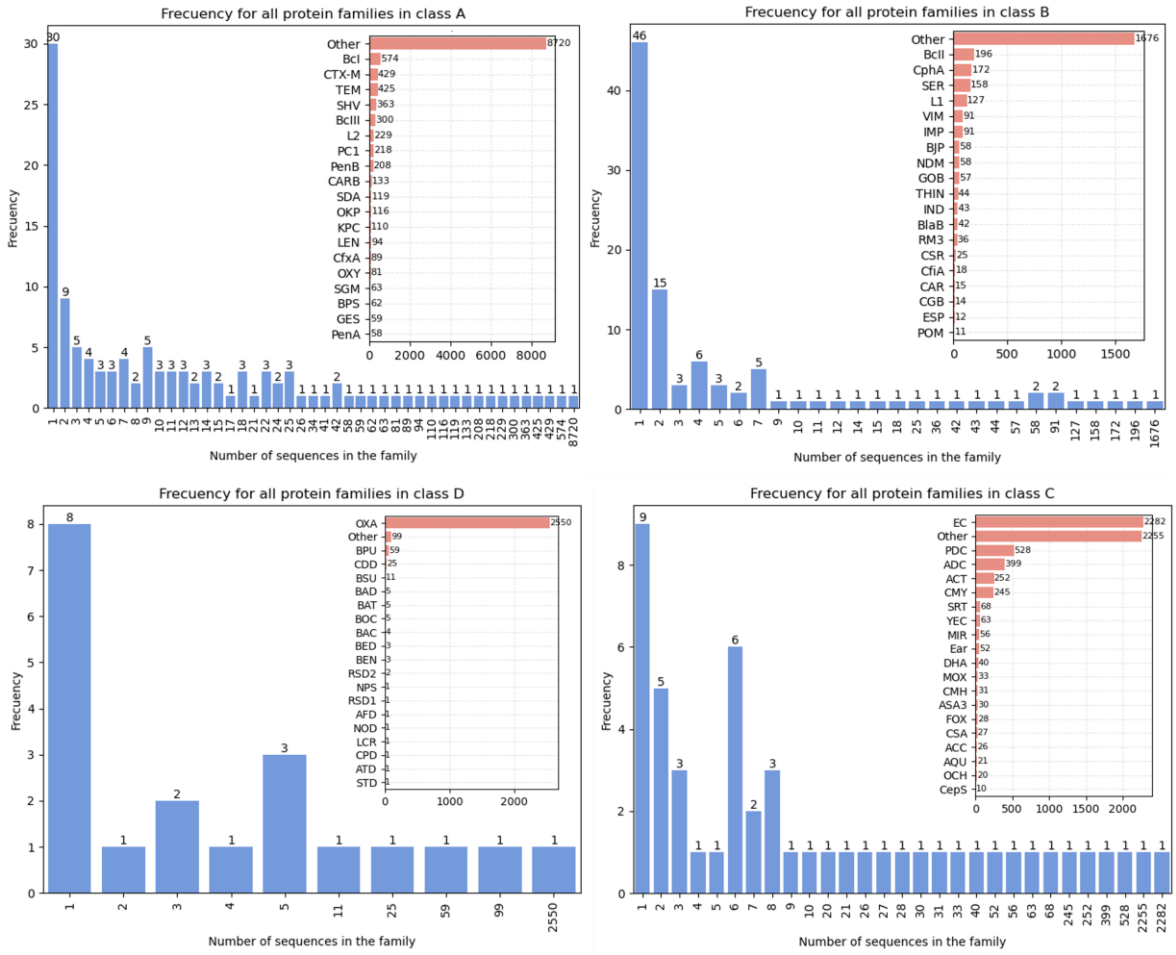


Figura suplementaria 1. Distribución del número de miembros en las familias enzimáticas por clase de betalactamasas. Se muestra en el eje X el número de miembros contenido en la familia enzimática y en el eje Y el número de familias con dicha cantidad de secuencias. Dentro de cada panel, se muestra un gráfico de barras con el número de secuencias contenido en las 19 familias más abundantes junto con las secuencias etiquetadas como "Otras" al no corresponder a alguna familia reconocida. Por ejemplo, para el total de 2,779 secuencias de la clase D, la BLDB reconoce 19 familias, siendo OXA la familia más abundante con 2,550 secuencias mientras que solo 99 secuencias fueron etiquetadas como "otras". En general, se observa que solo unas pocas familias cuentan con un gran número de secuencias.

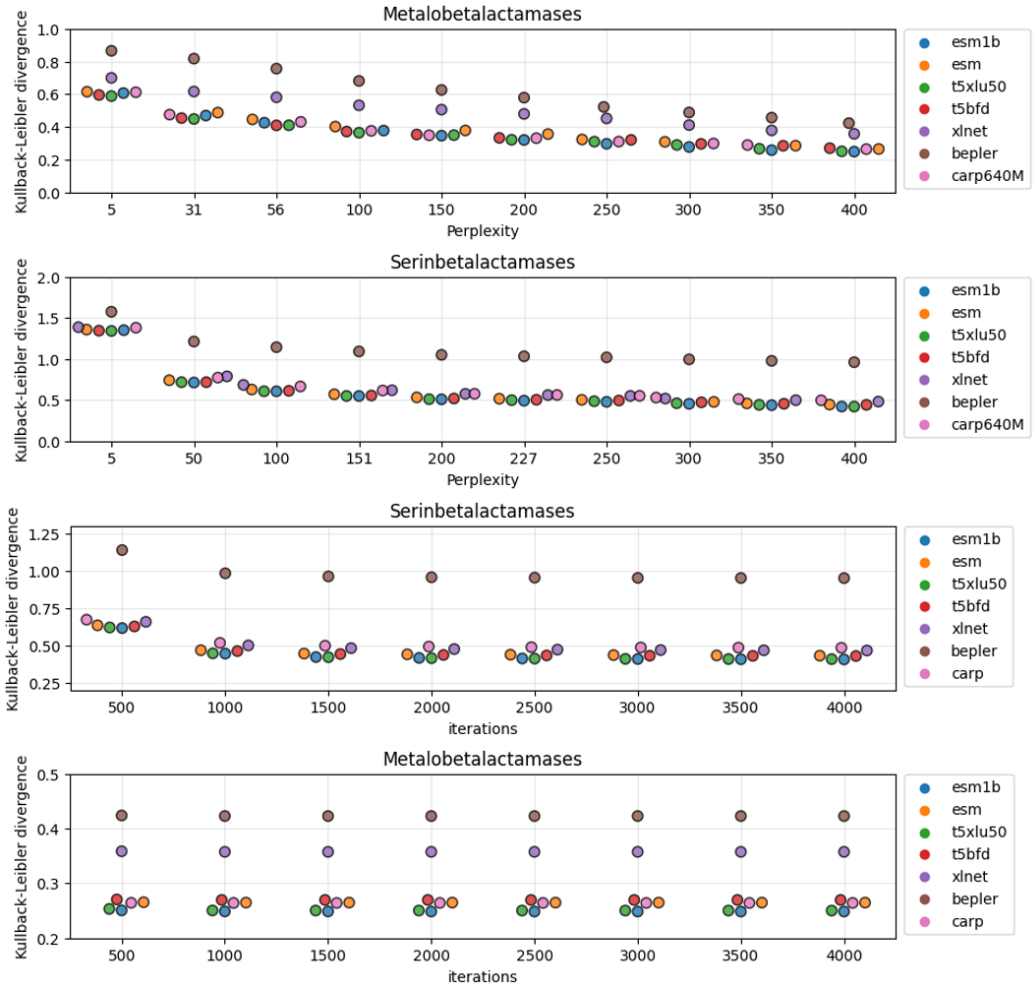


Figura suplementaria 2. Divergencia de Kullback-Leibler en función de valores de *perplexity* y número de iteraciones. Se indica en el encabezado de cada panel el tipo de betalactamasas y el parámetro a evaluar (*perplexity* y número de iteraciones). *Perplexity* es un parámetro que indica el número de datos vecinos considerados para computar la distribución de probabilidades que se usan para representar los datos; dichas distribuciones son específicas para cada modelo así como de la estructura del conjunto de datos[148]. *Perplexity* debe tomar un valor menor al número total de datos analizados, o menor al total de datos de una categoría si se conocen previamente, y usualmente toma valores entre 30 y 50. Sin embargo, se ha sugerido que para conjuntos de datos grandes (>10,000 datos), considerar una *perplexity* de 30 causara un pobre desempeño en la preservación de la estructura global[203], por lo que se recomienda altos valores de *perplexity* (*e.g*. 500) para grandes conjuntos de datos así como probar varios valores para identificar configuraciones estables[140]. Johnson et al.[130] sugieren que, dependiendo de la cantidad del conjunto de datos, un valor de *perplexity* ≥ 1,000 puede ser bueno, especialmente cuando los datos tienen varias escalas de organización como en el caso de las betalactamasas (*i.e.* clase, subclase, familias, subfamilias, variantes). Oskolkov[144,147] sugiere determinar el valor de *perplexity* al usar una ley de potencias de la forma *perplexity = n^(1/2)*, siendo *n* el número de datos a analizar. Kobak & Berens[149] sugieren determinar el valor de *perplexity* asignando un valor equivalente al 1% del total de los datos analizados. Heiser & Lau[142] sugieren determinar el valor de *perplexity* asignando un valor equivalente entre el 3% y 10% del total de los datos. Incluso Zhou *et al*.[150] sugieren usar valores tan altos de *perplexity* como *n/3* (*n* = número de datos) para asegurar que la probabilidad conjunta de todos los puntos de datos con respecto a cada otro sea calculada. Considerando estas sugerencias, determiné un conjunto de 10 valores de *perplexity* para cada tipo de betalactamasas. Tanto para las serin y metalobetalactamasas en todos los modelos de lenguaje, se observa que un valor alto de *perplexity* se traduce en menores valores de divergencia de Kullback-Leibler, razón por la cual conservé la representación obtenida con *perplexity* = 400. Al evaluar el efecto de las iteraciones usando una *perplexity* = 400 se observa que no hay cambios en los valores de divergencia finales después de 1500 iteraciones, razón por la que elegí este valor para los experimentos posteriores. Para más detalles de la organización de las serin y metalobetalactamasas revisar el *Jupyter notebook* perplexity_optimization y tsne_iterations en el repositorio de esta tesis. Los siguientes parámetros fueron usados en todos los casos: *random_state* = 420, metric = “cosine”.

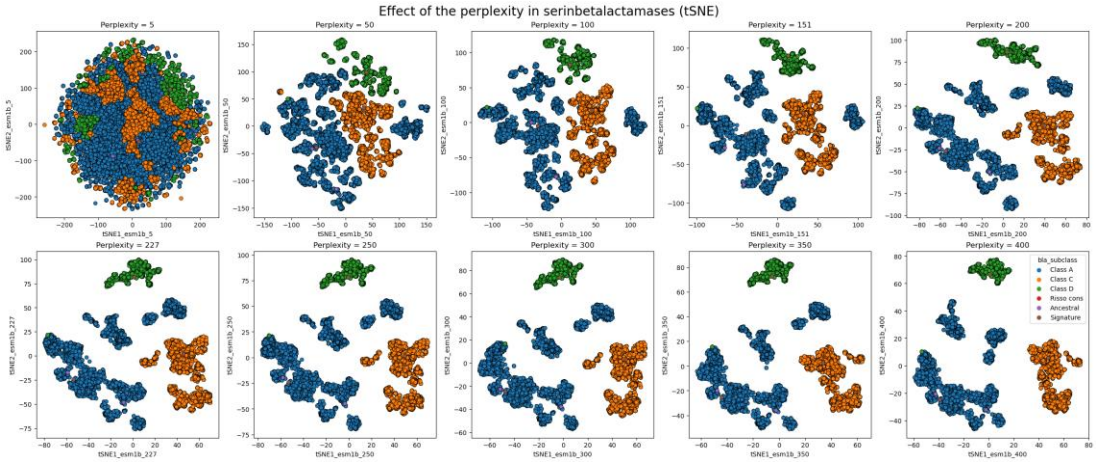


Figura suplementaria 3. Organización de las serinbetalactamasas en función de valores perplexity usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usaron 1500 iteraciones.

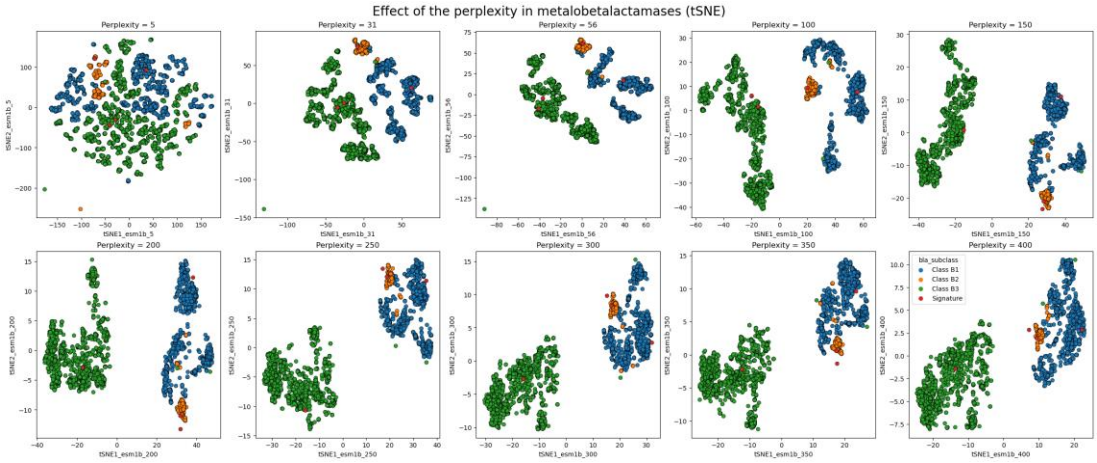


Figura suplementaria 4. Organización de las metalobetalactamasas en función de valores perplexity usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usaron 1500 iteraciones.

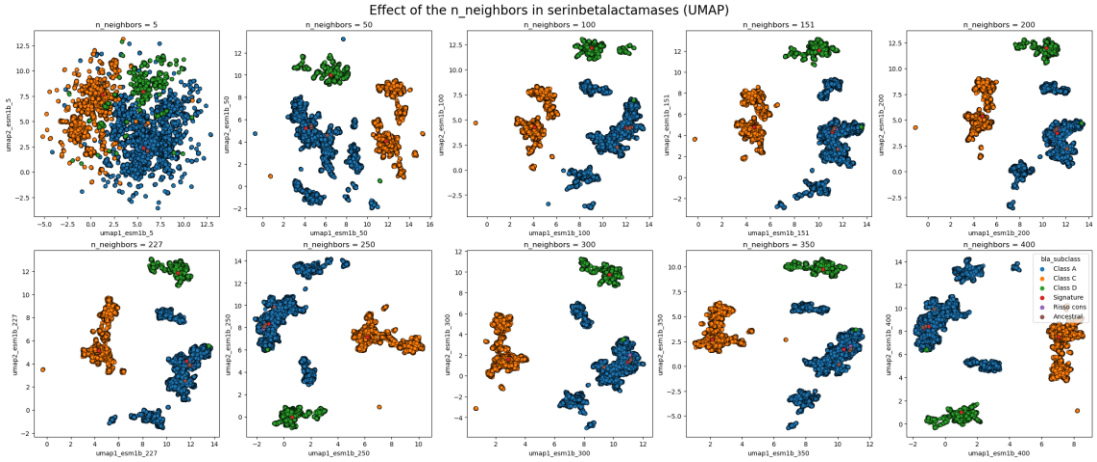


Figura suplementaria 5. Organización de las serinbetalactamasas en función de valores n_neighbors usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usaron valores de min_dist y spread = 0.2.

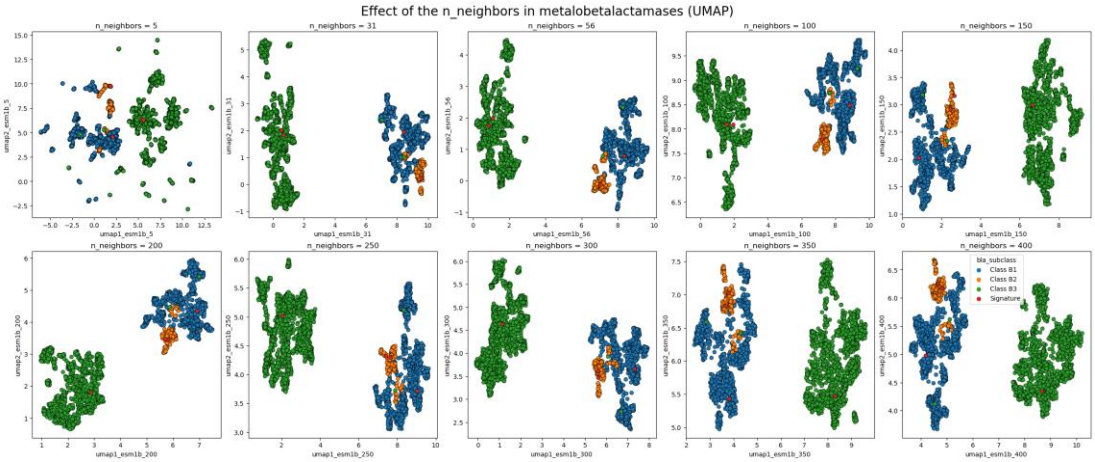


Figura suplementaria 6. Organización de las metalobetalactamasas en función de valores n_neighbors usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usaron valores de min_dist y spread = 0.2.

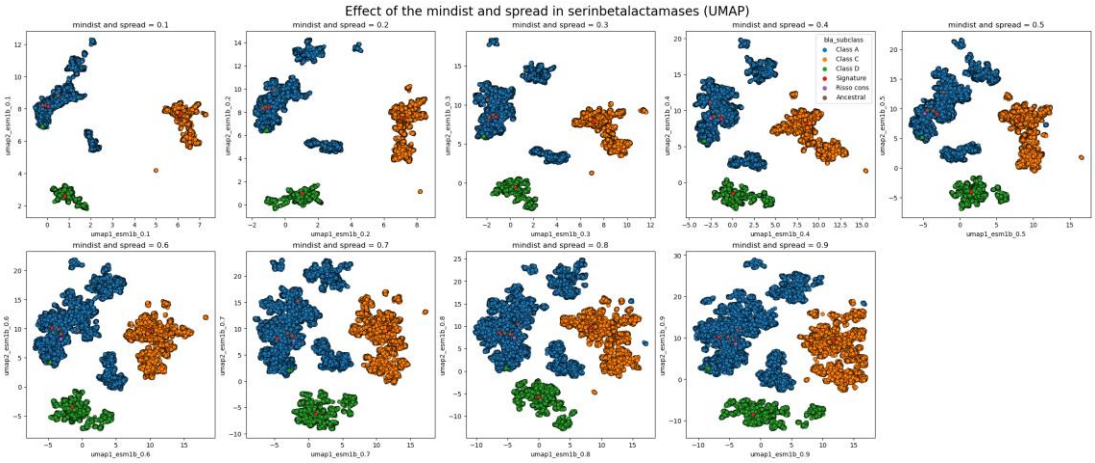


Figura suplementaria 7. Organización de las serinbetalactamasas en función de valores mindist y spread usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usó un valor de n_neighbors = 400.

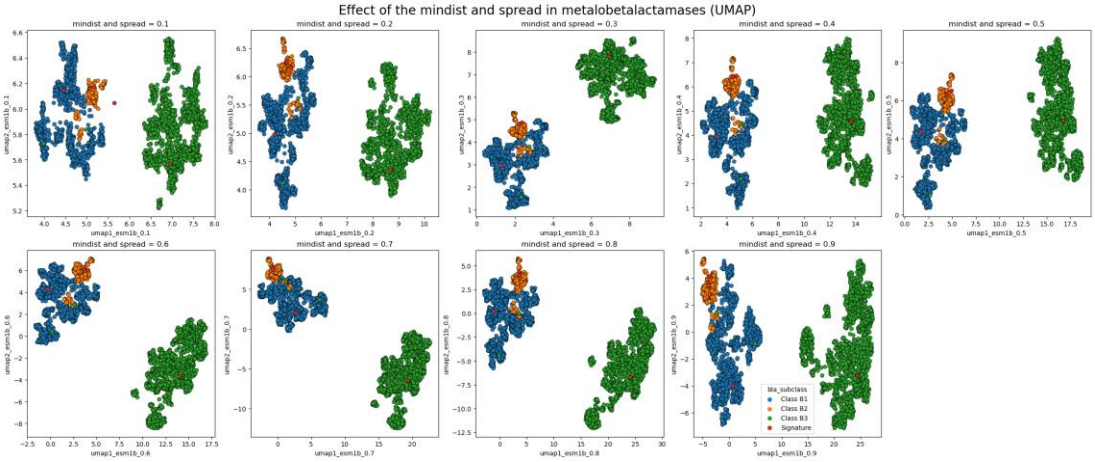


Figura suplementaria 8. Organización de las metalobetalactamasas en función de valores mindist y spread usando ESM-1b. Para una mejor visualización solo se muestra la leyenda en un panel. Se usó un valor de n_neighbors = 400.

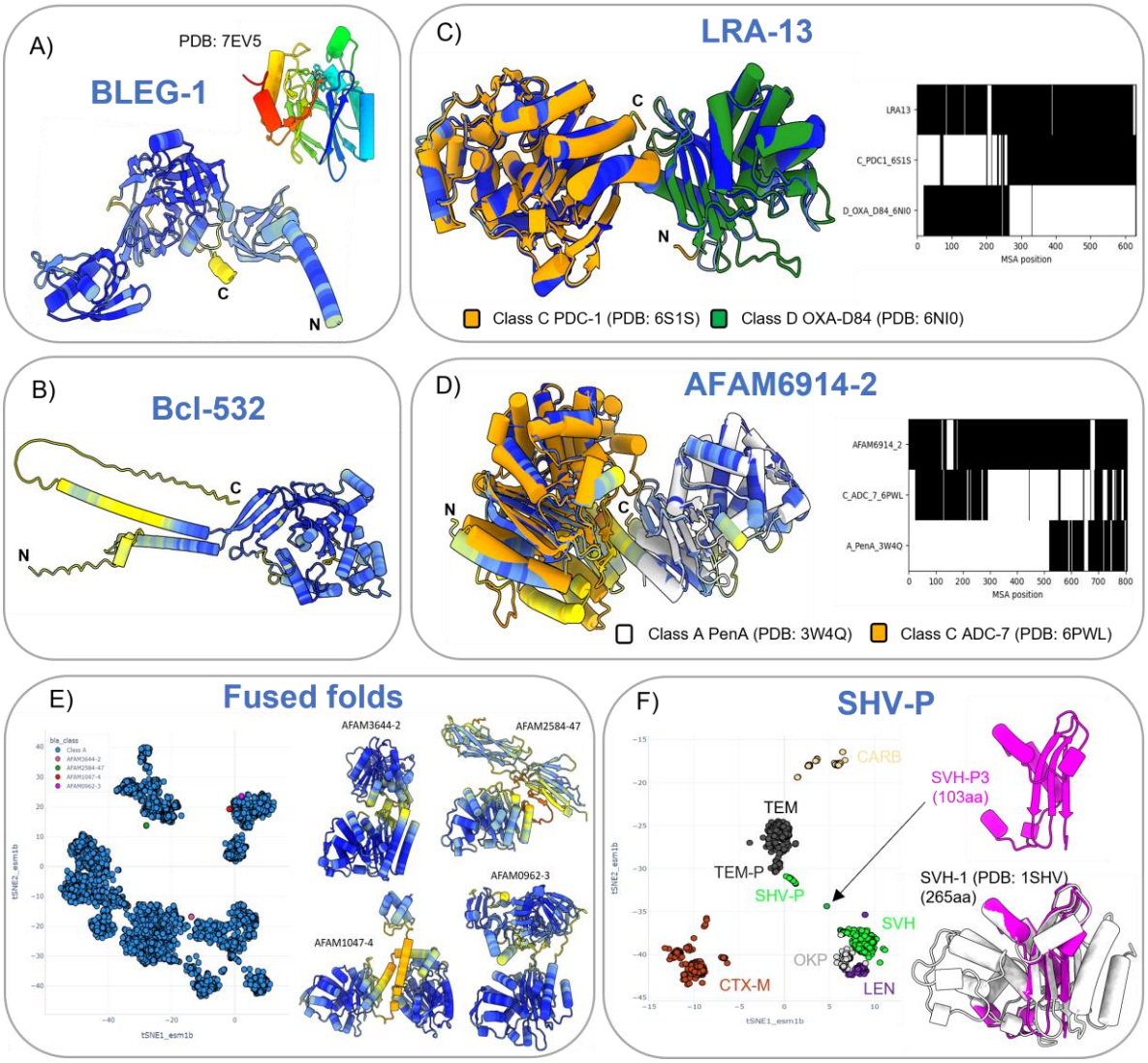


Figura suplementaria 9. Estructuras de las secuencias anómalas. Las estructuras se predijeron usando la versión de ESMFold[89] implementada en ColabFold[162] con parámetros predeterminados. Los modelos se colorean por sus valores de pLDDT y se señala con una N y una C las partes N-terminal y C-terminal. (A) BLEG-1 (ID: AIC95013.1); se muestra como comparación la una estructura depositada en la PDB para la proteína BLEG-1 coloreada de por número de residuos de N a C terminal (PDB: 7EV5), mostrando claramente que son plegamientos distintos. (B) BcI-532 (ID: AAB49182.1); un canal iónico de *Homo sapiens*. (C) LRA-13 (ID: WP_063839877.1); se indican en la parte inferior la clase de betalactamasa, nombre de la enzima y los PDB ID de las estructuras cristalográficas de referencia y un gráfico de la ocupación de las secuencias en un alineamiento, mostrando de color negro y blanco la presencia y ausencia de aminoácidos, respectivamente. (D) AFAM6914-2 (ID: WP_105000841.1); una putativa betalactamasa dimérica entre la clase A y C. Notablemente, una región intermediaria de seis hélices de esta betalactamasa cuenta con baja calidad de predicción. (E) Dentro de todas las clases, algunas pocas secuencias cuentan con una longitud atípica para su respectivo grupo. Solo en la clase A se presentaron este tipo de anomalías bastante marcadas y dichas secuencias demostraron ser plegamientos de betalactamasas fusionadas con otros plegamientos que no son betalactamasas, y pese a ello, fueron incluidas en la BLDB que supuestamente es manualmente curada por dos laboratorios. Se muestra como ejemplo cuatro de estos casos, así como su ubicación en la representación de tSNE de ESM-1b de la clase A. (F) Secuencias parciales de la familia SHV. Unas pocas secuencias de las familias SHV y TEM incluyen en sus encabezados la indicación de que son secuencias parciales, es decir, secuencias incompletas. Se muestra la predicción estructural de una de ellas, SHV-P3 la cual tiene apenas 103 residuos, mientras que la SHV-1 tiene 265. Pese a ser parciales, se encuentran relativamente bien agrupadas.

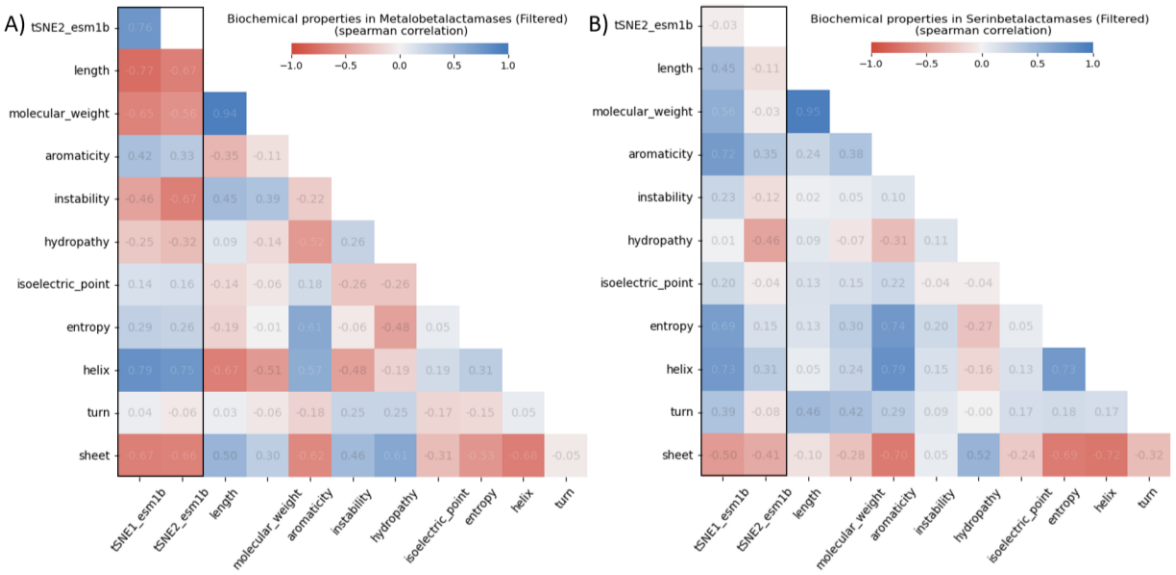


Figura suplementaria 10. Matriz de correlación entre las dimensiones de tSNE con ESM-1b y las propiedades cuantitativas. Las correlaciones se estimaron usando las secuencias obtenidas de la BLDB y filtradas con un rango de ±30% del valor de la mediana de la longitud de secuencia de cada grupo. Se resalta en un recuadro negro las dos dimensiones de tSNE.

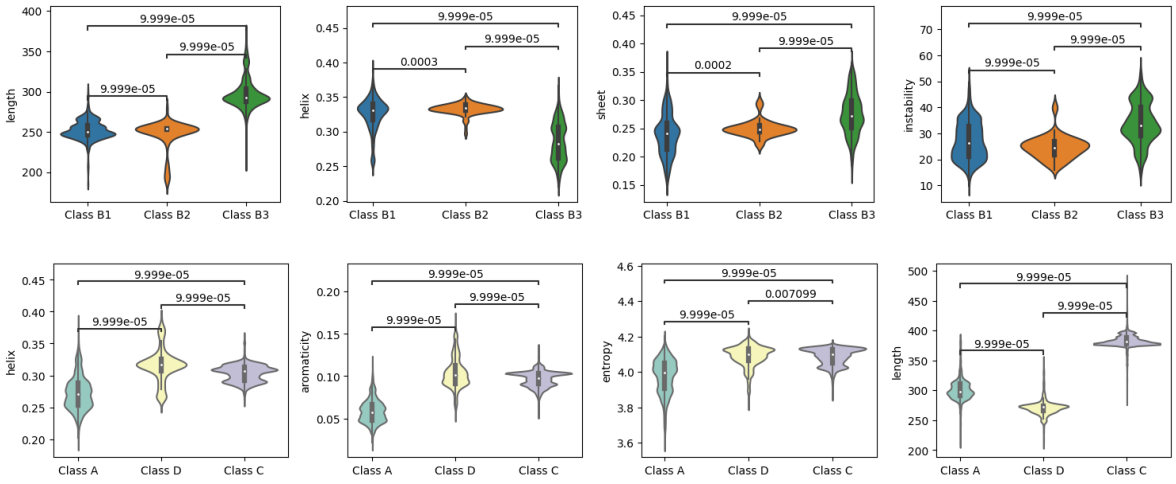


Figura suplementaria 11. Distribución las propiedades cuantitativas de las betalactamasas con buena correlación de Spearman (ρ > 0.50 con alguno de sus ejes). Las distribuciones se estimaron usando las secuencias obtenidas de la BLDB y filtradas con un rango de ±30% del valor de la mediana de la longitud de secuencia de cada grupo. Se indican los valores de *p-value* obtenidos mediante una prueba de permutaciónes no paramétrica (permutation_test) implementada en la librería de Python mlxtend (parametros: func='x_mean != y_mean', method='approximate', num_rounds=10000, seed=0, paired=False). Número de secuencias por grupo: B1 = 1,205; B2 = 201; B3 = 1,693, A = 13,177; C = 6,564; D = 2,759.

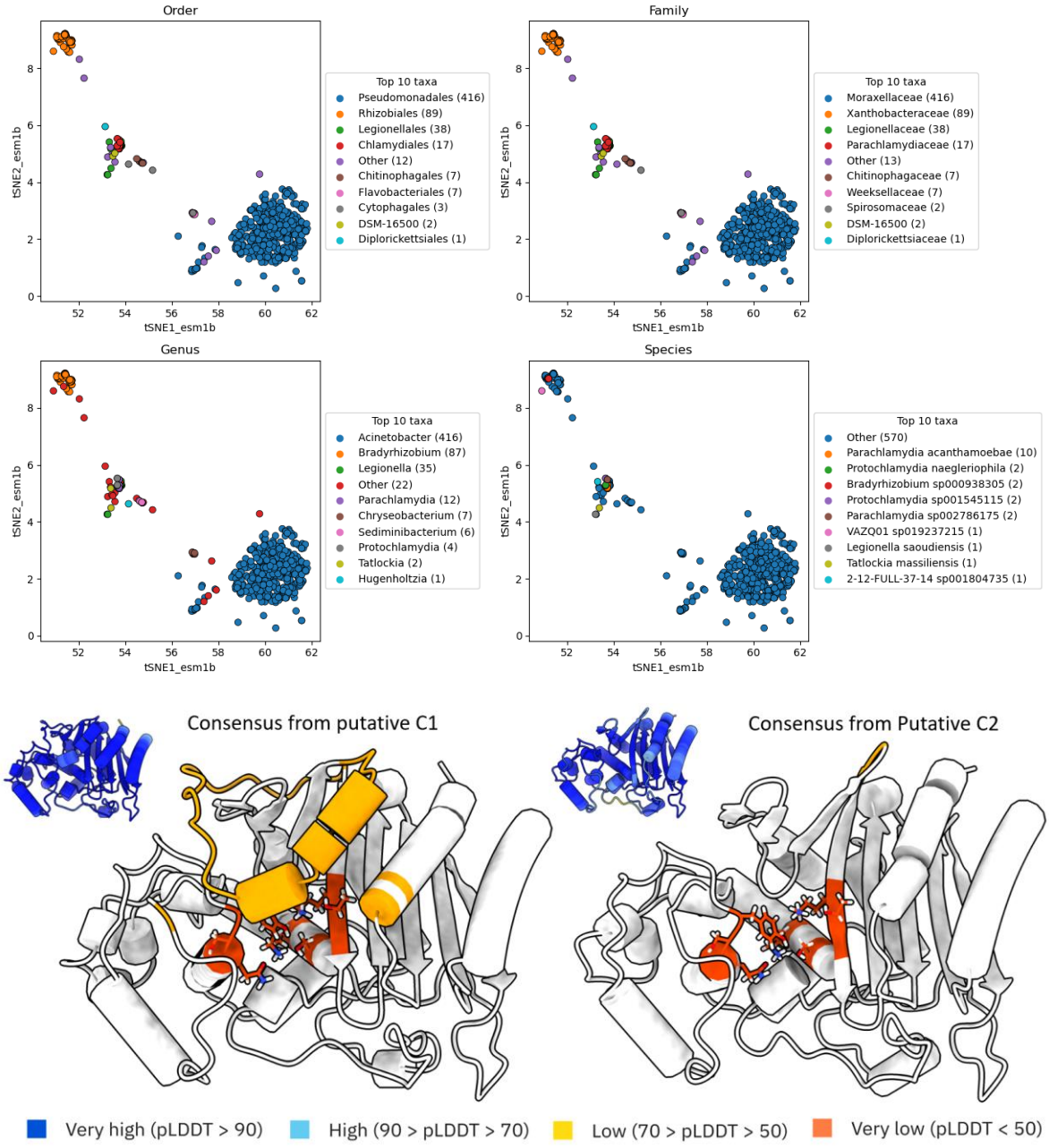


Figura suplementaria 12. Organización taxonómica de la putativa subclase C2. (Superior) clasificación taxonómica de la putativa subclase C2. Los paneles corresponden a la sección indicada con un recuadro rojo de la clase C de la Figura 17A. Se indica en la leyenda de cada panel su categoría taxonómica. Para una mejor visualización se enlista en orden decreciente los 10 taxa más abundantes y el resto se indica como "Otros". (Inferior) Diferencias estructurales de betalactamasas clase C. Para conocer las posibles diferencias estructurales entre las putativas subclases C1 y C2 tomé las secuencias representativas de cada grupo las filtré por número de residuos (C2 = 320 a 430; C1 = 370 a 410), resultando en 92 secuencias para C2 y 648 secuencias para C1. Posteriormente, creé un alineamiento usando Mafft y generé una secuencia consenso con un enfoque por pesos con Alignbuddy. Finalmente predije la estructura de dichas secuencias consenso con AlphaFold2. Se muestra en naranja las diferencias estructurales detectadas con TM-align. En rojo se muestran las tres firmas catalíticas (SxxK, YxN, KTG). Se muestra la correspondiente estructura en pequeño y coloreada por valores de pLDDT. Actualmente la BLDB registra 18 familias en la clase C que han sido cristalizadas y depositadas en la PDB, y todas portan la estructura señalada en rojo compuesta por dos alfa hélices y una región de giro (PDB IDs: ACC-1=6K8X, ACT-1=2ZC7, ADC-1=4NET, BUR-1=5E2G, CHV-1=5EVL, CMH-T1=6LC7, CMY-2=8DI7, EC-1=1FSY, FOX-4=5CGW, HaBLA=3WRT, MOX-1=3W8K, MYC-1=5E2H, PDC-1=2WZX, PFL-P1=2QZ6, PSY-1=5EVI, RHO-1=7MQN, SUC-1=6NJK, TRU-1=6FM6).

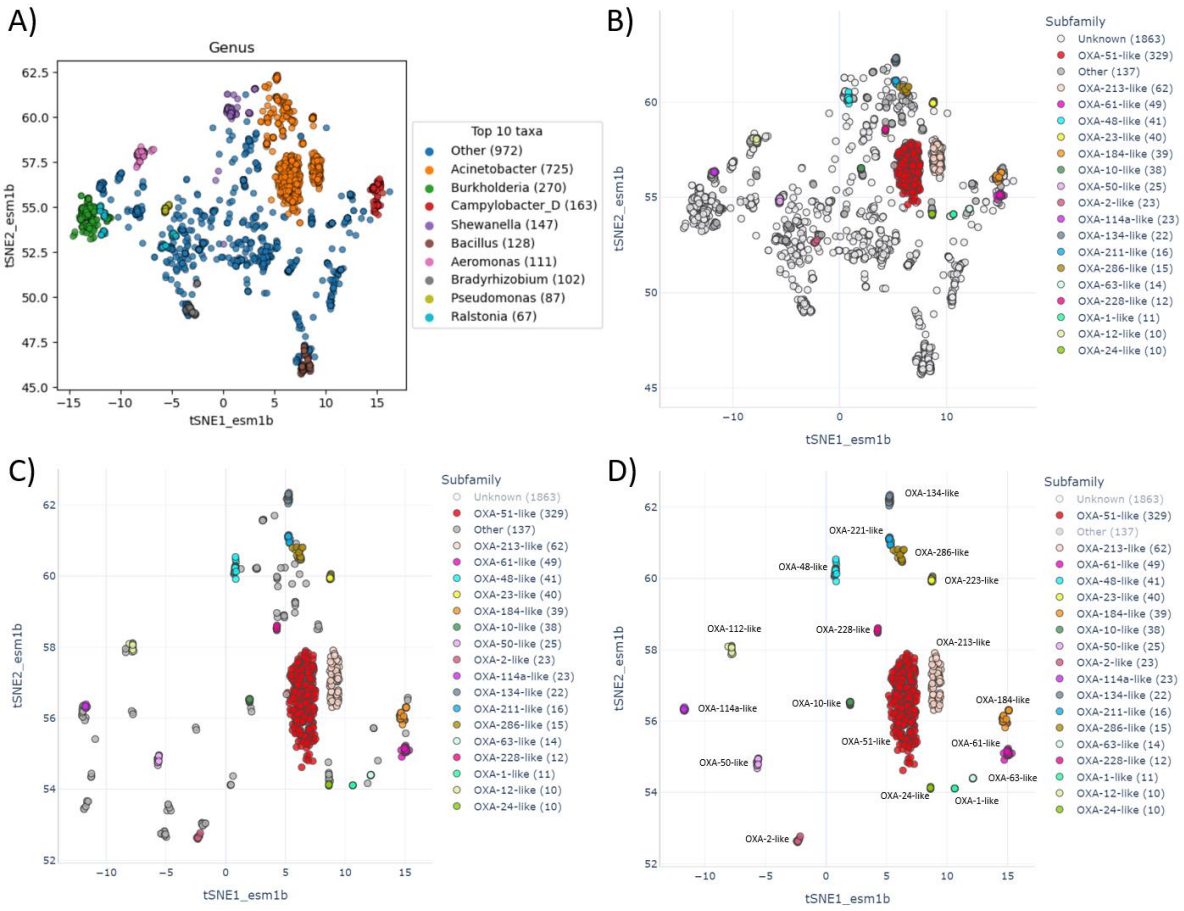


Figura suplementaria 13. Organización por subfamilias de la Clase D. (A) Organización taxonómica a nivel de género de las betalactamasas clase D. Se indican los nueve géneros más abundantes (B) Organización taxonómica a nivel de subfamilia. Se indican las 18 subfamilias más abundantes y se etiqueta como "Otra" a las subfamilias que no están dentro de este rango, mientras que todas las secuencias que no tienen asociada una subfamilia se señalan como "Desconocidas". (C) Mismas coordenadas que en el panel B, pero sin las secuencias clasificadas como "Desconocidas". (D) Mismas coordenadas que en el panel C, pero sin las secuencias clasificadas como "Otras". Se indican entre paréntesis los conteos de cada subfamilia. Para una mejor visualización, se omitieron las siete secuencias OXA-D# de las coordenadas totales.

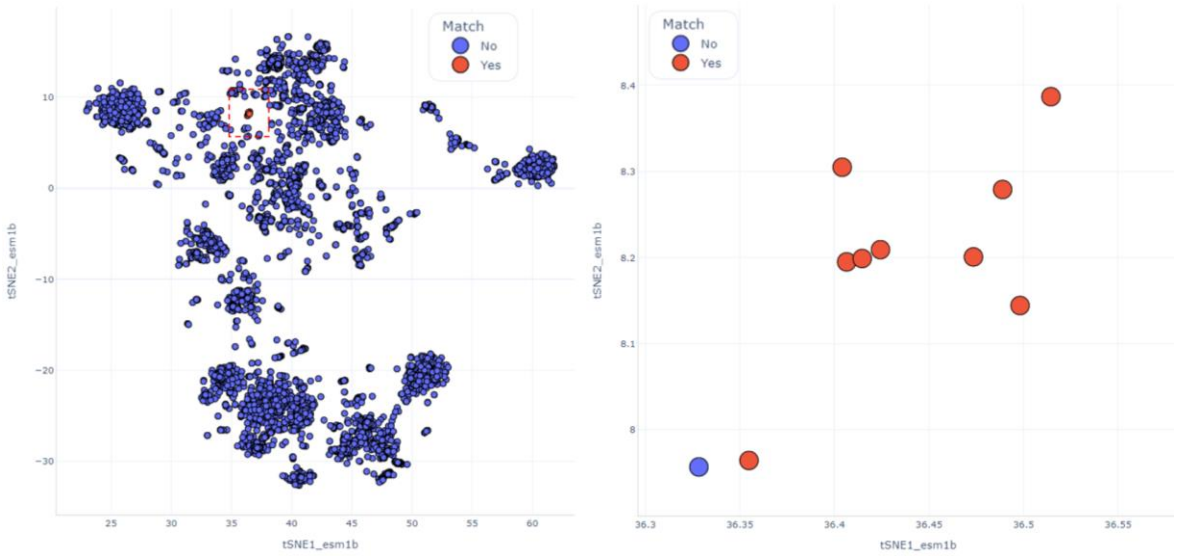


Figura suplementaria 14. Posibles betalactamasas diméricas. Se mapeó los identificadores de las nueve secuencias reportadas por Chaves-Silveira *et al*.[166] a la organización de serinbetalactamasas, teniendo registro en la región de la clase C. Se señala en el panel de la izquierda con un recuadro rojo punteado a la región que se hace un acercamiento en el panel de la derecha. Las nueve secuencias señaladas en rojo se encuentran en un espacio en común con otra secuencia (ID: ELX10257.1) de 619 residuos la cual no fue registrada por Chaves Silveira *et al*., pero que potencialmente representa otra secuencia dimérica.

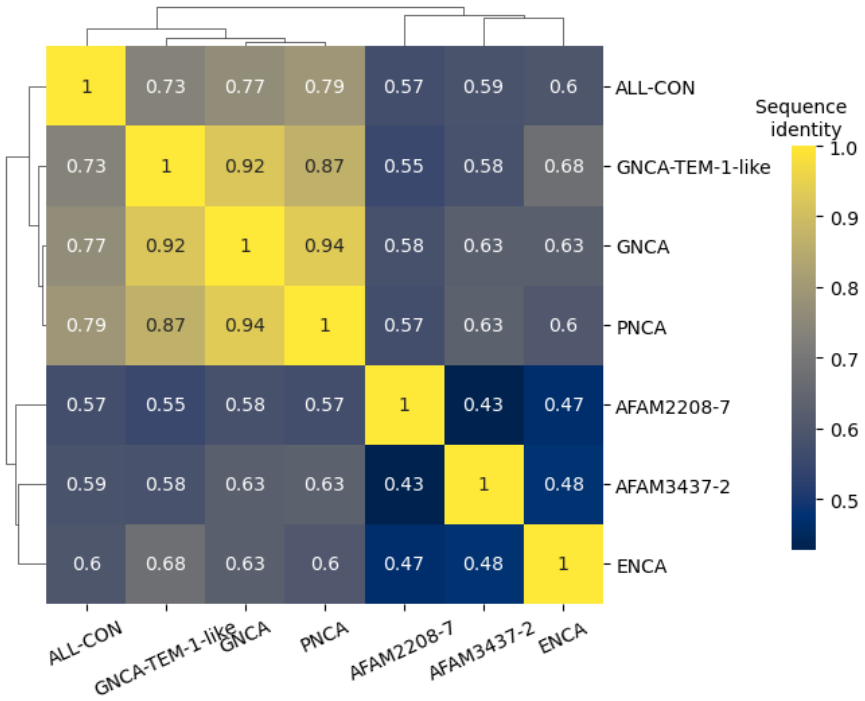


Figura suplementaria 15. Porcentaje de identidad estimado de los ancestros y el consenso computados por Risso *et al*. Para estimar la identidad de secuencia generé un alineamiento con las respectivas secuencias usando Mafftv7[119] y limpié los gaps con el comando "trimal gappyout" de AlignBuddy, posteriormente, computé una matriz de identidad usando EvCouplings[211] y la visualicé con clustermap (method = ward, metric = euclidean) de la librería Seaborn.

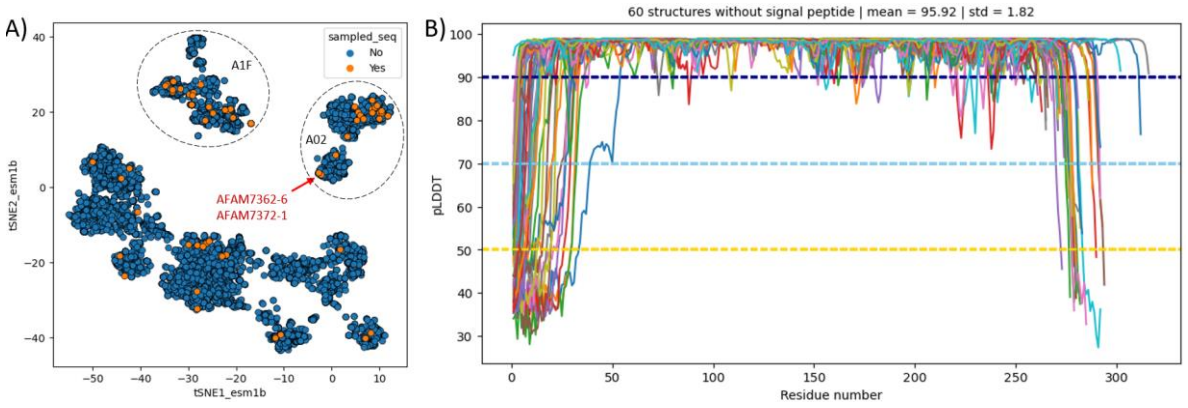


Figura suplementaria 16. Características de los 60 modelos de la clase A predichos con AlphaFold2 (A) Ubicación de las estructuras muestreadas al azar. Se señalan con círculos punteados a las secuencias que fueron consideradas dentro de la subclase A2 y la putativa subclase A3, y las secuencias fuera de estos círculos fueron consideradas como A1. Una vez se delimitaron los tres grupos, se tomaron al azar 20 secuencias por grupo, se removieron los péptidos señales detectados por SignalPv6[78] y se predijeron sus estructuras con la implementación de ColabFold[162] del algoritmo de AlphaFold2[67]. Se indica con flechas rojas los nombres de varias secuencias cuyas estructuras se presentan en la Figura suplementaria 20. (B) Calidad de las predicciones estructurales. Se indica en el encabezado el promedio de pLDDT entre las 60 estructuras y su desviación estándar. Se muestran los valores de pLDDT de cada uno de los residuos, así como tres líneas punteadas de referencia indicando la calidad de las predicciones de acuerdo con el código de AlphaFold2: Azul obscuro = Alta confianza; Azul claro = Buena confianza; Amarillo = baja confianza.

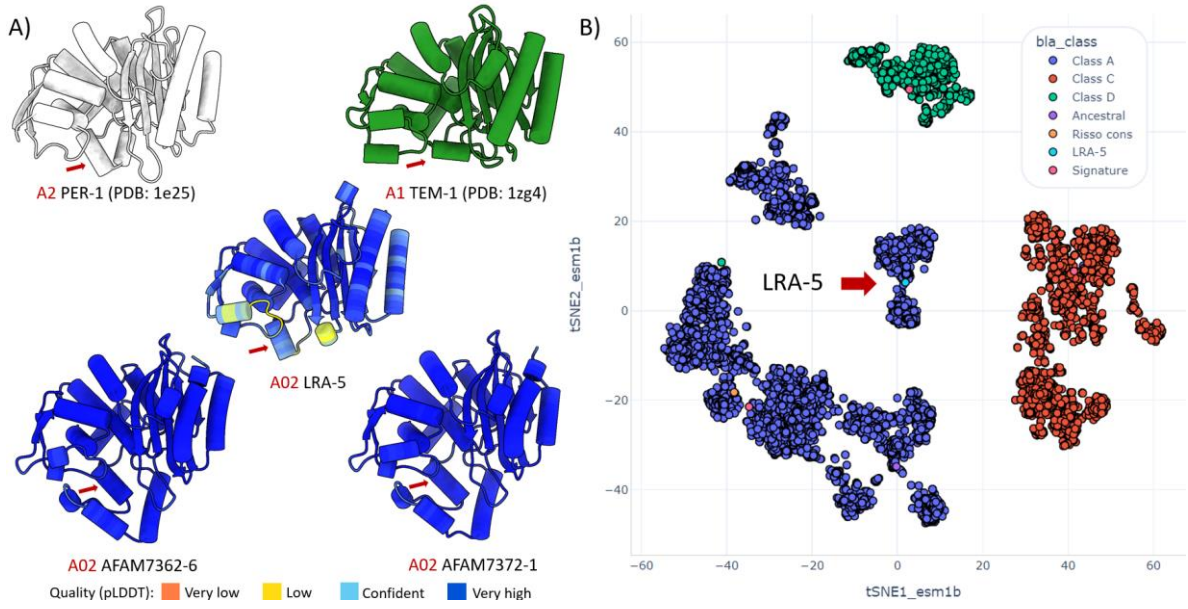


Figura suplementaria 17. Serinbetalactamasas clase A y LRA-5. (A) Diversidad estructural. Se muestra en blanco la estructura de PER-1 de la subclase A2 y en verde la estructura de TEM-1 de la subclase A1. Se muestran las estructuras predichas con AlphaFold2 de las secuencias LRA-5, AFAM7362-6 y AFAM7372-1 coloreadas por valores de pLDDT. Se señala con flechas la ubicación del omega *loop*. (B) Ubicación de LRA-5 en la representación de tSNE de ESM-1b. Se señala con una flecha roja la ubicación de la secuencia LRA-5 (ID: WP_063842268.1). Parámetros: Divergencia de Kullback-Leibler = 0.42; perplexity = 400; iteraciones = 1500.

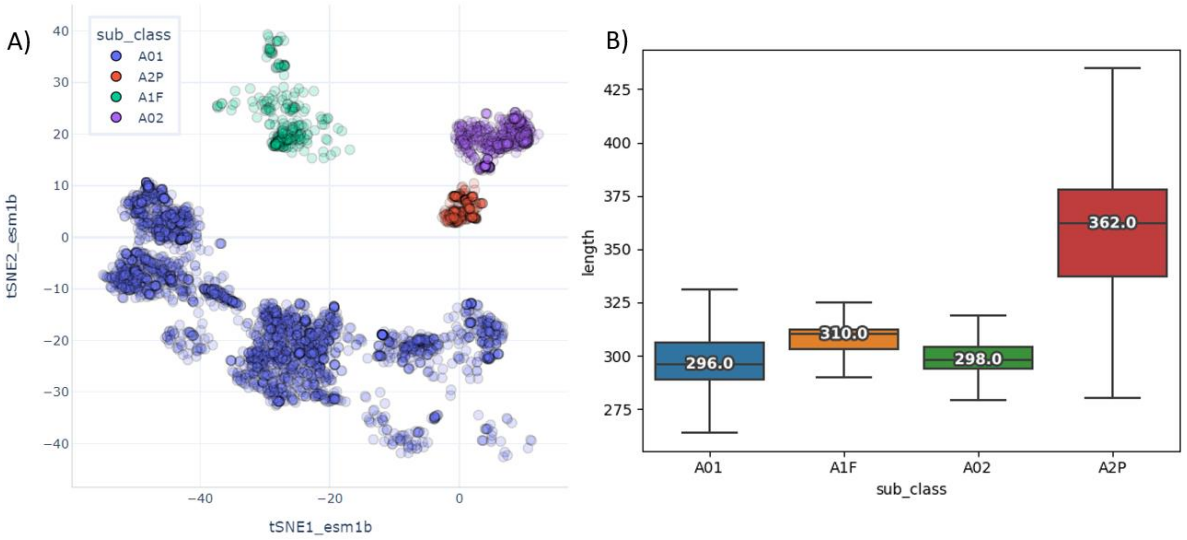


Figura suplementaria 18. Identificación de la putativa subclase A3. (A) Ubicación de las secuencias representativas por cada grupo. Solo se muestran las coordenadas de las secuencias representativas. (B) Distribución del número de residuos. Se señala con números el valor de la media. Para una mejor visualización se removieron los puntos de los valores atípicos. Si solo se consideran las secuencias de la familia Sphingomonadaceae del grupo 2AP, el valor de la media es de 370.

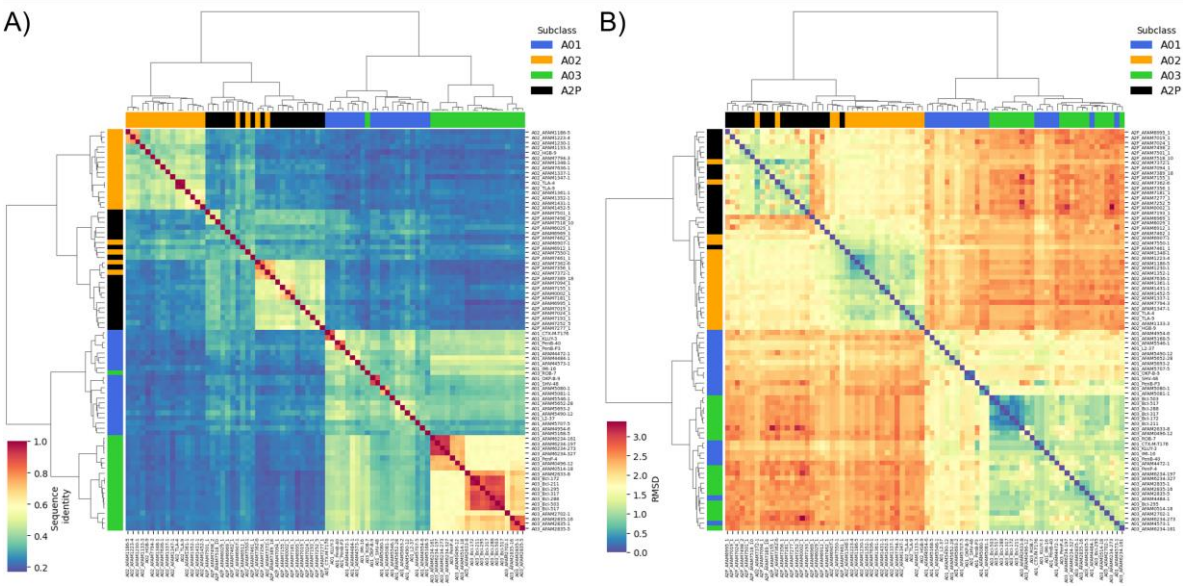


Figura suplementaria 19. Similitud a nivel de secuencia y estructura de los tres grupos de serinbetalactamasas clase A más 20 betalactamasas del grupo 2AP. Este análisis es el mismo que el presentado en la Figura 22, pero añadiendo 20 estructuras predichas con ESMFold del grupo de betalactamasas 2AP. Las secuencias fueron seleccionadas por la presencia de las tres hélices a partir de un conjunto de 100 secuencias tomadas al azar a partir de las enzimas representativas dentro de este grupo (i.e. agrupadas a 90% de identidad) que abarcaran todos los filos. Siendo las secuencias de Sphingomonadaceae las que muestran considerables diferencias respecto al resto. (A) Matriz de similitud de secuencia. (B) Matriz de similitud en función de la identidad de secuencia. Se muestra una leyenda de cuatro colores asociada cada uno de los cuatro grupos y una barra de color que indica los valores de identidad de secuencia o RMSD según el caso. Las matrices de similitud se construyeron con las estimaciones por pares de los respectivos valores y posteriormente, se visualizaron con la función clustermap (method = ward, metric = euclidean) de la librería Seaborn.

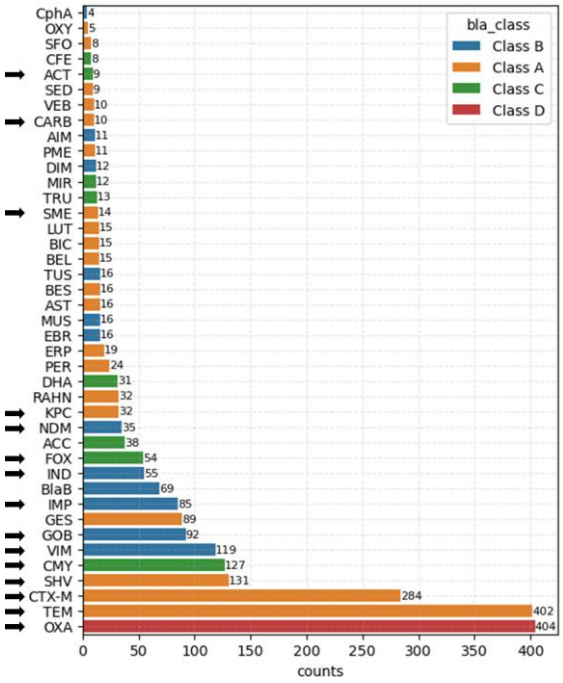


Figura suplementaria 20. Número de MICs por clase y familia de betalactamasas. Considerando el número de familias reconocidas por la BLDB como referencia (Fig. 5), la clase A solo tiene datos para el 16.67% de las familias (20 familias), la clase B tiene datos para el 11.88% de las familias (12 familias), la clase C tiene datos para el 16% de las familias (8 familias) y clase D solo tiene datos para el 5.26% de las familias (1 familia). Se indican con una flecha a las familias que tienen enzimas representativas dentro de la clasificación funcional de betalactamasas[11], las cuales representan un total de 1,853 datos de MICs (Fig. 2).

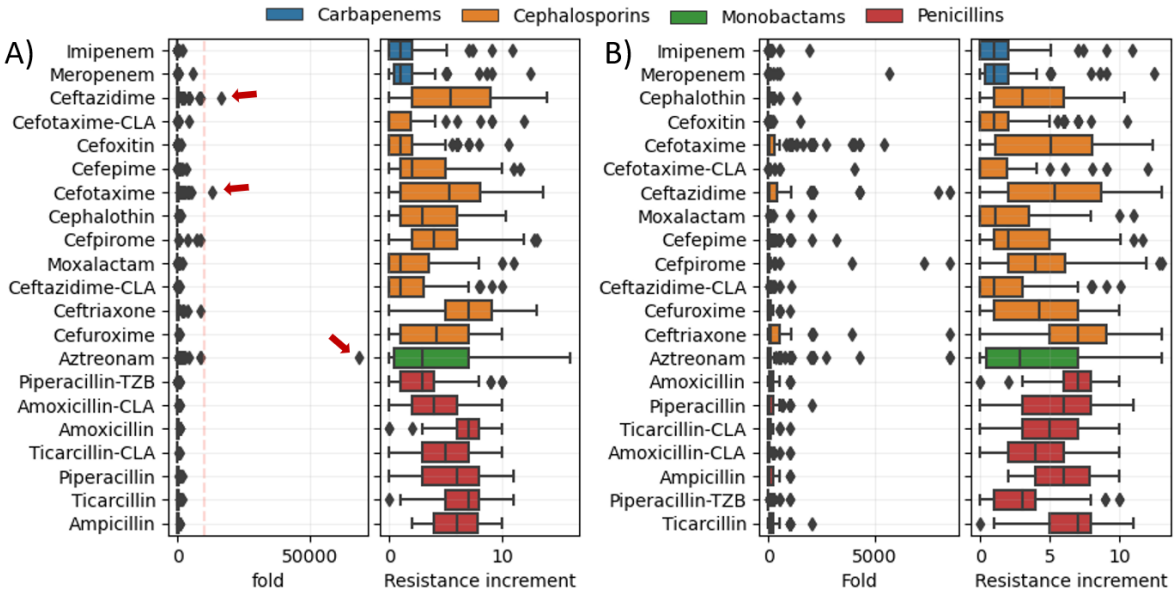


Figura suplementaria 21. Distribución de los valores de fold e incremento de resistencia. (A) Distribución antes de la remoción de los valores extremos. Se indica en una línea roja punteada el valor límite de fold (10,000) que fue considerado la definición de puntos extremos señalados con una flecha roja. Aunque dichos valores están reportados en sus respectivas publicaciones, se consideraron atípicos en esta tesis dado sus valores extremos. (B) Distribución de los valores después de la remoción de los valores extremos. Ver el Jupyter notebook "Functional_datasets_analysis" para más detalles.

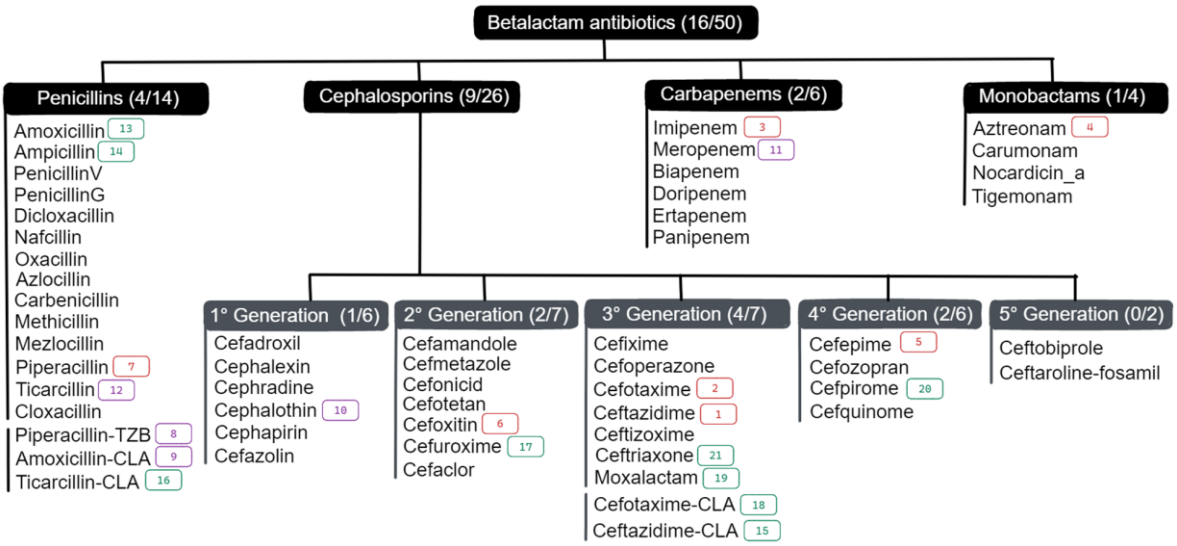


Figura suplementaria 22. Clases de antibióticos betalactámicos. Se consideraron 50 antibióticos clasificados en cuatro clases: monobactamas, carbapenemas, penicilinas y cefalosporinas (subdividida por generaciones). Los datos provienen principalmente de la base de datos CBMAR[103], sin embargo, noté varios errores en sus clasificaciones así que curé estos datos al corroborar su clase caso por caso. Se muestran como líneas discontinuas aparte a las cinco combinaciones con inhibidores. En el encabezado de cada clase se indica entre paréntesis el número de antibióticos presentes en mis datos respecto al total, por ejemplo, tengo datos para una de las cuatro carbapenemas. Se señala con recuadros a los antibióticos presentes en mis datos, y con un número su posición al ordenarlos por conteo, siendo ceftazidima y ceftriaxona los antibióticos con más y menos datos, respectivamente (Fig. 25). Los recuadros rojos señalan antibióticos con >150 datos, morado con ≥100 y <150, y en verde <100 datos. Abreviaciones: CLA, ácido clavulánico; TZB, tazobactam. Esquema modificado de la ref.[103] y realizado en https://excalidraw.com/.

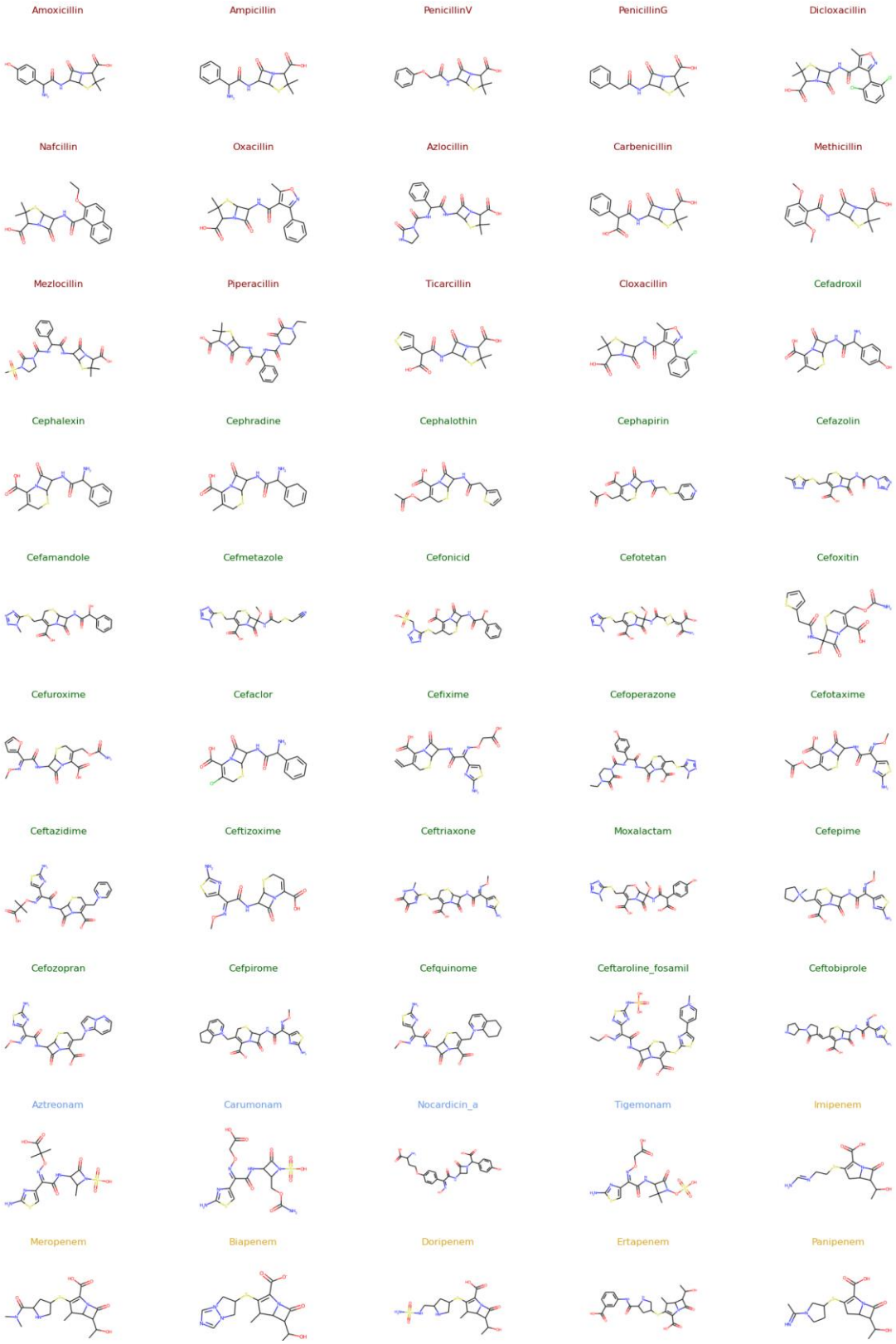


Figura suplementaria 23. Estructuras químicas de los 50 antibióticos betalactámicos. Se señala con color rojo, verde, azul y amarillo a los antibióticos de las clases penicilinas, cefalosporinas, monobactamas y carbapenemas, respectivamente. Las estructuras fueron visualizadas usando sus respectivas SMILES canónicas con la librería RDKit[182].

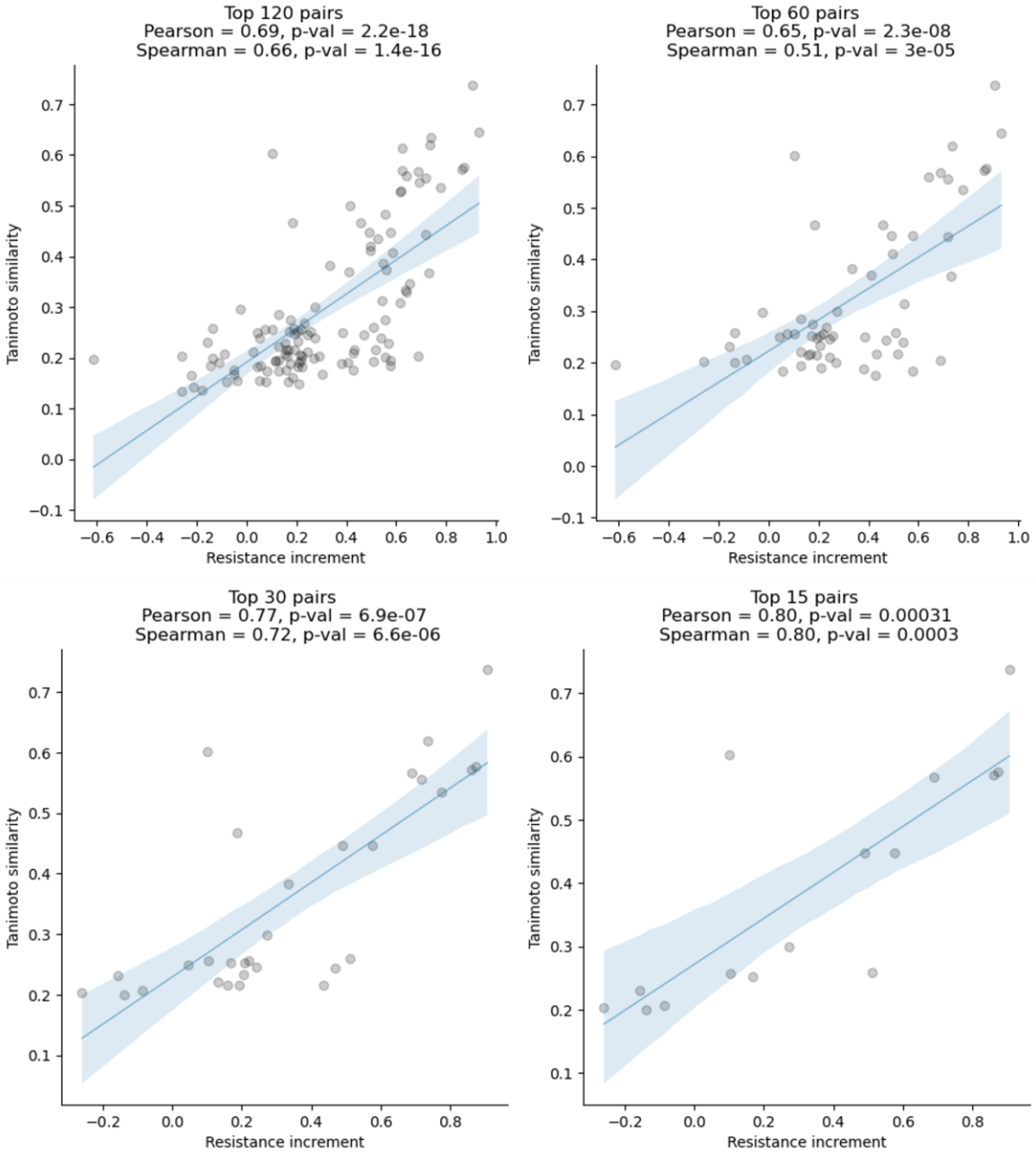


Figura suplementaria 24. Correlación entre datos de similitud de Tanimoto y correlación de incremento de resistencia. Se indica en el encabezado de cada panel el número de datos con los que se computó la correlación de Pearson (y Spearman). Los tres paneles que siguen a la correlación con los 120 pares totales siguen un patrón de submuestreo donde conservo el 50% superior de los datos ordenados por incremento de resistencia (i.e. 60 pares, 30 pares y 15 pares). Las valores de las correlaciones son bastante similares si considero los 120 pares y los filtro por sus p-values, resultando en 57 pares con un valor menor a 0.05. Para más detalles ver el Jupyter notebook "Functional_datasets_analysis".

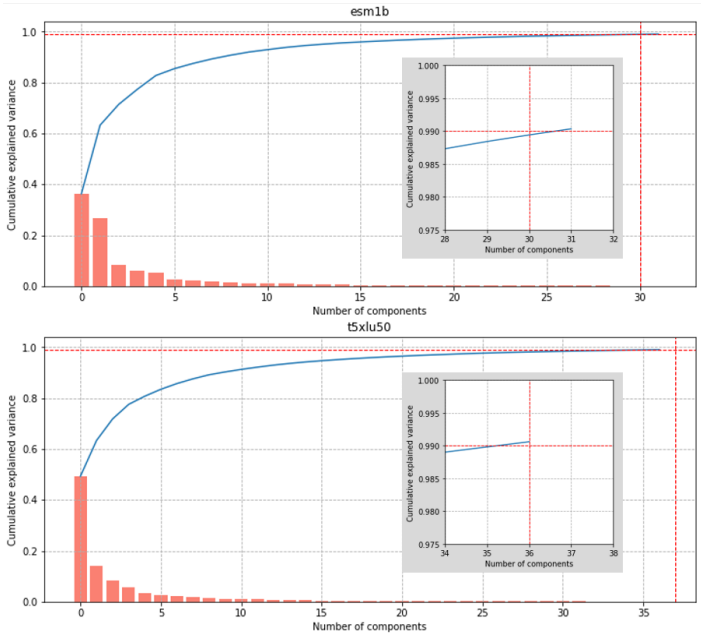


Figura suplementaria 25. Representaciones de baja dimensión con PCA de Prot-T5-XLU50 y ESM-1b. Se muestra la varianza explicada acumulada en función del número de componentes para ambos modelos de lenguaje de proteínas. Se indica el respectivo modelo en el encabezado de cada panel. Se muestra un acercamiento de los componentes en los que se acumula un 99% de la varianza (número de componentes: ESM-1b = 32; Prot-T5-XLU50 = 37).

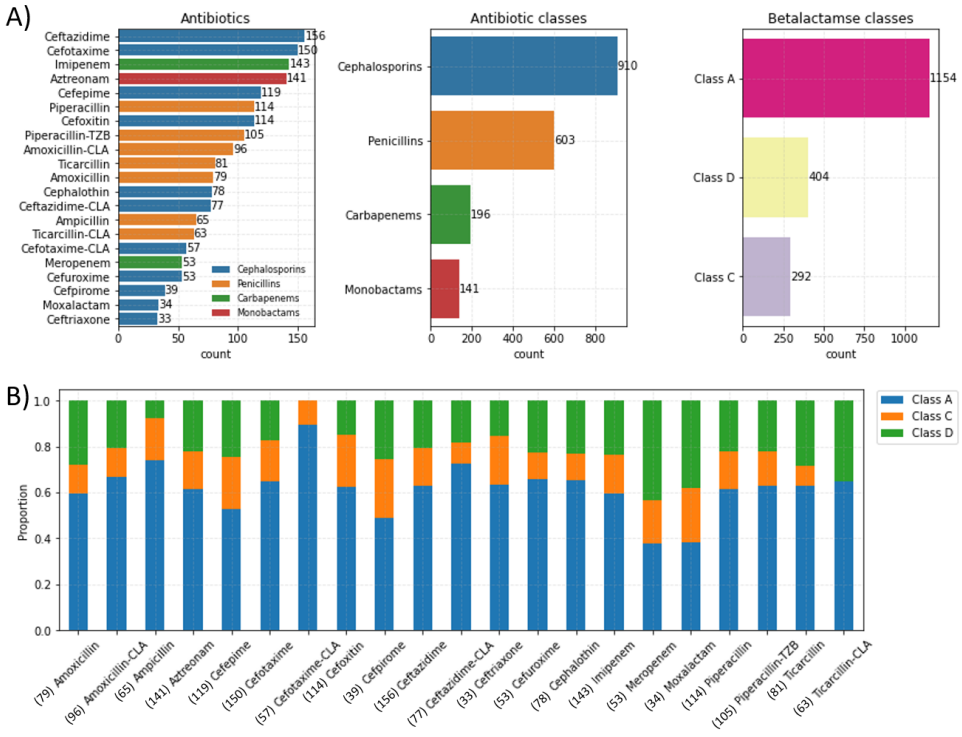


Figura suplementaria 26. Cantidad de datos de entrenamiento de serinbetalactamasas. (A) Cantidad de datos por antibiótico, clase de antibiótico y de betalactamasa. (B) Proporción de las clases de serinbetalactamasas por conjunto de antibióticos. Se indica en paréntesis la cantidad de datos. Abreviaciones: CLA, ácido clavulánico; TZB, tazobactam.

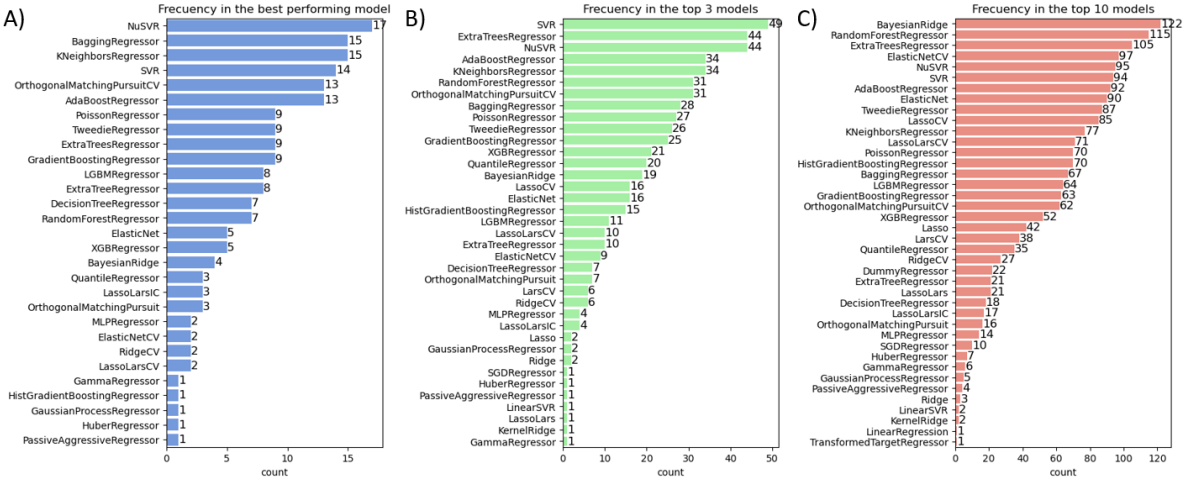


Figura suplementaria 27. Selección del mejor algoritmo de regresión. La librería LazyPredict permite comparar 42 algoritmos de regresión distintos (usando parámetros por defecto en cada caso), los cuales se probaron por cada una de las nueve representaciones (siete PLMs y dos representaciones de PCA de ESM-1b y ProtT5-XL-U50). (A) Frecuencias de los algoritmos en el top 1 mejores modelos. (B) Frecuencias de los algoritmos en el top 3 mejores modelos. (C) Frecuencias de los algoritmos en el top 10 mejores modelos. Se señala en cada barra el número de conteos de cada regresor.

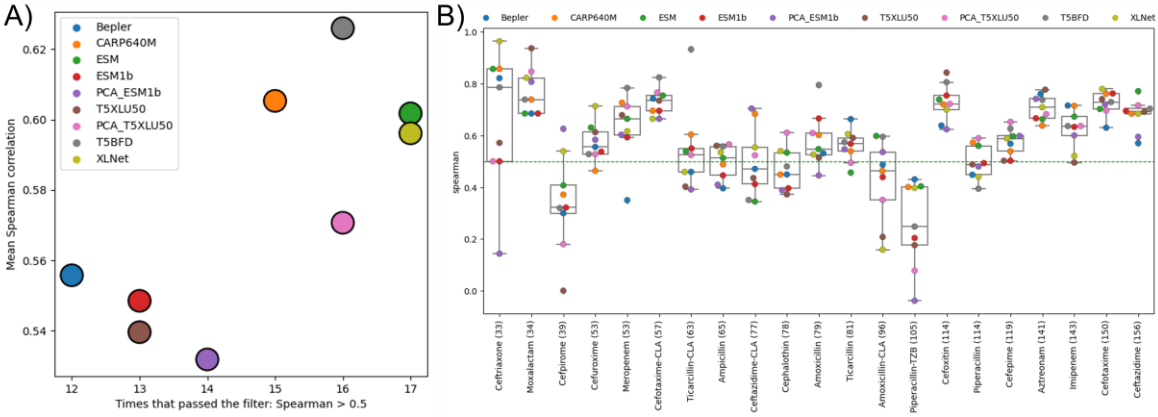


Figura suplementaria 28. Selección del mejor modelo de lenguaje de proteínas. (A) Criterios de evaluación. El valor de correlación de Spearman = 0.5 fue arbitrariamente asignado como un punto de corte. El valor promedio de correlación de Spearman se computa por cada codificación a lo largo de los 21 conjuntos de antibióticos. (B). Comparación de la correlación de Spearman de las 10 codificaciones por cada uno de los 21 antibióticos en la tarea de regresión. Se muestra en verde punteado el límite arbitrario a un valor de correlación de Spearman = 0.5. Por cada antibiótico se indica entre paréntesis el número total de datos disponibles, por ejemplo, para la Ceftriaxona hay disponibles 33 valores de incremento de resistencia, ocupando 26 datos para el entrenamiento (80%) y 7 para el conjunto de prueba (20%). Los antibióticos están ordenados de menor a mayor cantidad de datos de izquierda a derecha.

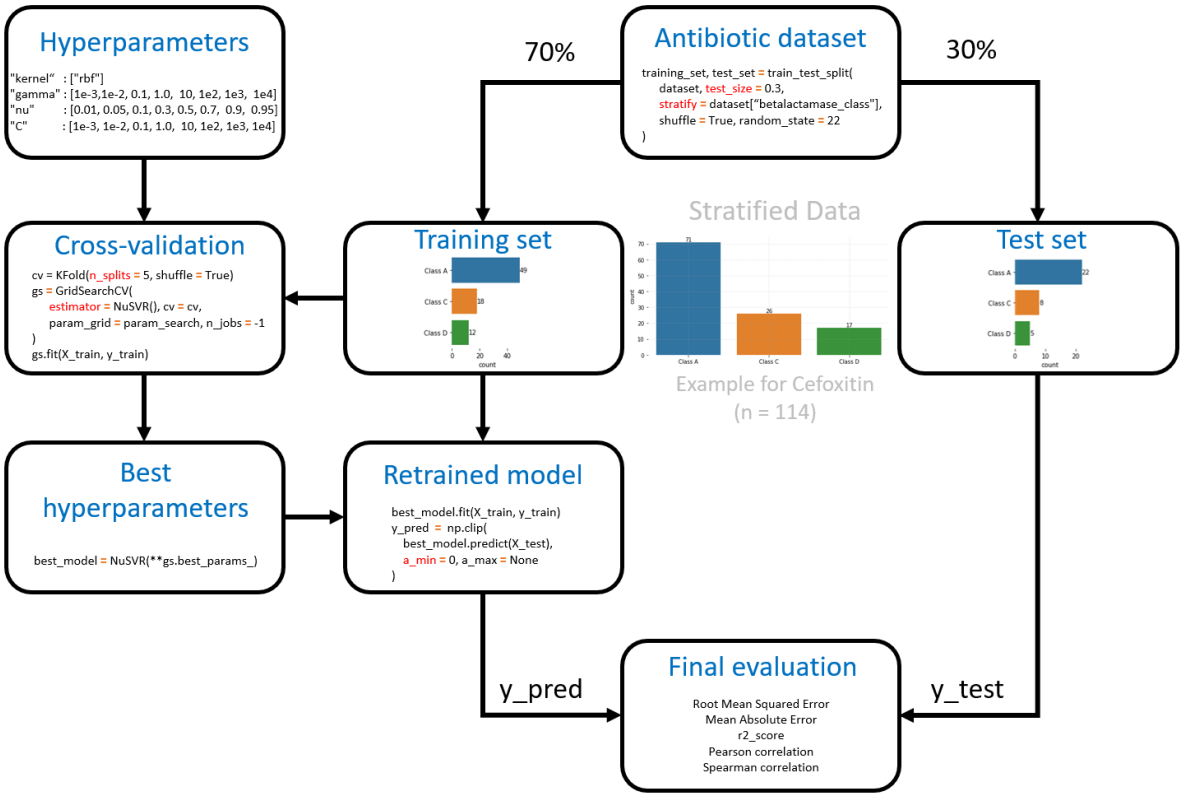


Figura suplementaria 29. Esquema de entrenamiento. Se muestra en cada panel el código de Python asociado. Se usan los datos de Cefoxitina para ilustrar como el conjunto de datos de entrenamiento y prueba cuentan con una proporción similar de las clases de betalactamasas. Esquema modificado de: https://scikit-learn.org/stable/modules/cross_validation.html

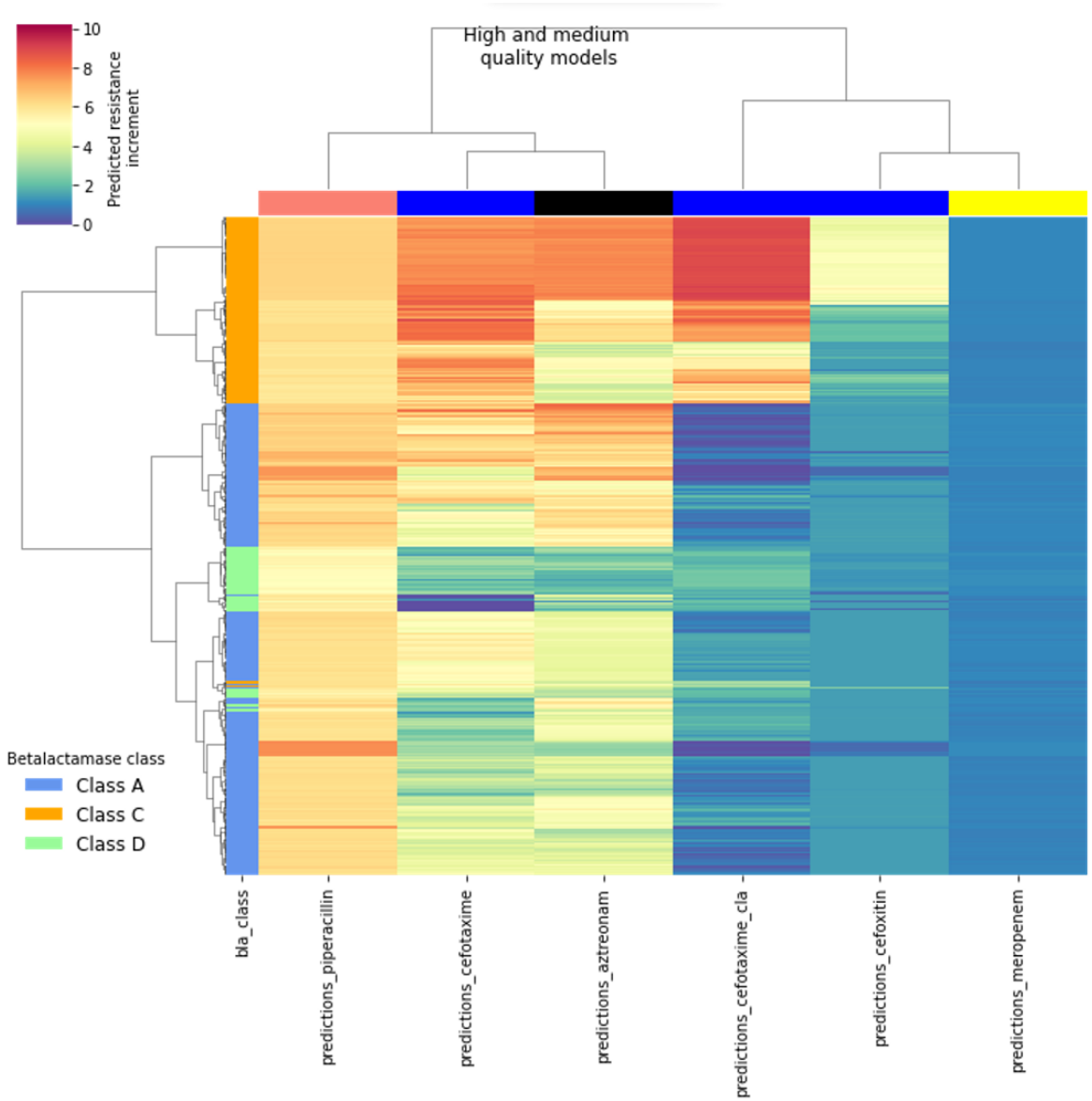


Figura suplementaria 30. Agrupación de las serinbetalactamasas en función de las predicciones de incremento de resistencia. Solo se consideraron los regresores de alta y mediana calidad. Se indica en la legenda las clases de betalactamasas y con barras de colores el eje horizontal la clase de antibiótico (rojo = penicilinas, azul = cefaloporinas, negro = monobactamas, amarillo = carbapenemas. Los datos se visualizaron con la función clustermap (method = ward, metric = euclidean) de la librería Seaborn.