# PHYperImpute: Imputación Filogenética Adaptativa mediante Selección Automática de Modelos

José M. Torres-Domínguez[1], Sergio Santander-Jiménez[1][0000−0002−2862−2026] y Miguel A. Vega-Rodríguez[1][0000−0002−3003−758X]

Universidad de Extremadura (https://ror.org/0174shg90), Escuela Politécnica, Campus Universitario s/n, Cáceres 10003, España.
jtorresgn@alumnos.unex.es, sesaji@unex.es, mavega@unex.es

**Resumen** Las distancias filogenéticas perdidas representan un problema que puede afectar significativamente a la precisión de los métodos bioinformáticos de reconstrucción de relaciones evolutivas. Para abordar dicho problema, este trabajo propone PHYperImpute, un método adaptativo de imputación de distancias filogenéticas basado en la selección automática de modelos de aprendizaje automático. PHYperImpute integra estrategias específicas para el tratamiento de conjuntos de datos filogenéticos, comparando distintos modelos candidatos en base a la información disponible para adaptar la imputación a las características de cada conjunto. La propuesta ha sido evaluada frente a otras diez técnicas de imputación y tres métodos filogenéticos del estado del arte, utilizando cuatro conjuntos de datos que engloban un total de 320 escenarios experimentales con distintos tamaños del problema y porcentajes de datos perdidos del 5 % al 60 %. Los resultados obtenidos, según la métrica *Normalized Robinson-Foulds* y el tiempo de ejecución, demuestran que PHYperImpute alcanza en tiempos competitivos la mejor precisión global en comparación al resto de técnicas.

**Keywords:** Filogenética · Imputación de datos perdidos · Matrices de distancias · Aprendizaje automático · Selección de modelos.

## 1. Introducción

La filogenética es una disciplina de amplio interés en bioinformática que estudia las relaciones evolutivas entre organismos, genes y especies. Dichas relaciones se plasman a través de árboles filogenéticos, los cuales representan una herramienta fundamental para comprender la historia evolutiva de la biodiversidad, y su aplicabilidad se extiende a problemas de gran relevancia, como el análisis de las variantes del virus SARS-CoV-2 durante la pandemia de COVID-19 [3]. En este contexto, disponer de filogenias precisas permite rastrear de manera efectiva el origen y la propagación de enfermedades infecciosas.

Entre los diversos enfoques para la inferencia de árboles filogenéticos, los métodos basados en distancias representan una aproximación ampliamente utilizada, ya que ofrecen un equilibrio razonable entre el coste computacional y la

calidad de los árboles filogenéticos obtenidos [14]. Estos métodos operan sobre matrices de distancias evolutivas calculadas a partir de secuencias biológicas alineadas. Sin embargo, la presencia de datos perdidos en dichas secuencias puede comprometer el cálculo de las distancias, dando lugar a matrices de distancias incompletas que dificultan la aplicación de los algoritmos de reconstrucción [7,11].

Ante esta problemática, surge el interés en aplicar soluciones basadas en técnicas de aprendizaje automático [6,9] con el objetivo de imputar valores perdidos en matrices de distancias filogenéticas. Entre las aproximaciones más destacadas aplicadas al contexto filogenético se encuentran los métodos propuestos en [4], basados en autocodificadores (*autoencoders*) y factorización matricial (*matrix factorization*). Otra propuesta alternativa es PhyloMissForest [13], que emplea *random forest* como modelo de imputación. No obstante, estas aproximaciones previas comparten una limitación común: utilizan una única técnica de aprendizaje automático durante todo el proceso de imputación, la cual puede no resultar precisa para *datasets* con distintas características o distribuciones numéricas en sus distancias. Es por tanto necesario explorar en filogenética estrategias basadas en enfoques híbridos y adaptativos [2] que permitan abordar esta limitación.

Este trabajo propone PHYperImpute, un método de imputación filogenética adaptativo, basado en el algoritmo HyperImpute [10], que realiza una selección automática de modelos en función de las características de la matriz de distancias de entrada. Para afrontar esta tarea, PHYperImpute integra, entre otros, modelos de imputación basados en *gradient boosting*, *random forest* y regresión por vectores de soporte como *Linear Support Vector Regression* (LinearSVR) y *Nu-Support Vector Regression* (NuSVR), esto es, modelos que destacan por su rendimiento en tareas de imputación de datos perdidos [1]. La idea clave reside en la evaluación comparativa de los modelos utilizando los datos observados, imputando las distancias perdidas seleccionando adaptativamente aquel modelo que mejor se ajuste a la distribución observada.

El resto de este artículo se organiza de la siguiente manera. La Sección 2 introduce la definición del problema abordado. La Sección 3 detalla el método propuesto PHYperImpute. La Sección 4 presenta los resultados experimentales obtenidos y comparativas con otros métodos. Finalmente, la Sección 5 presenta conclusiones y líneas de trabajo futuro.

## 2. Definición del Problema

Los análisis filogenéticos estudian los eventos evolutivos que explican la diversidad observada en un conjunto de $N$ secuencias biológicas alineadas. Cada secuencia contiene $M$ caracteres, cuyos estados vienen dados por el alfabeto IUPAC. Aquellos caracteres de los que se desconoce su estado se denotan con símbolos especiales, como '?', y pueden inducir datos perdidos.

Los métodos basados en distancias usan medidas de distancia entre secuencias para realizar la reconstrucción del árbol filogenético. Para calcular distancias, el primer paso es determinar la disimilitud $P_{ab}$ entre cada par de secuencias $a$ y $b$,

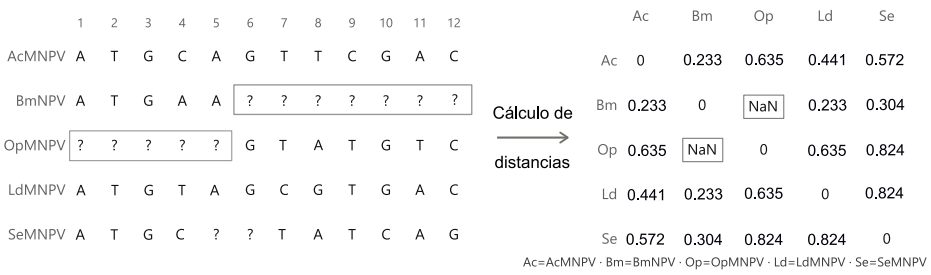



**Figura 1.** Ejemplo de matriz de distancias con datos perdidos. A la izquierda, las secuencias alineadas con estados observados y desconocidos ('?'). A la derecha, la matriz de distancias, donde las entradas perdidas se denotan como 'NaN'.

conforme a lo expresado en la Ecuación 1:

$$P_{ab} = \frac{\sum_{i=1}^{M_{obs}} \text{Diff}(a_i, b_i)}{M_{obs}}, \tag{1}$$

donde $\text{Diff}(a_i, b_i) = 1$ si $a_i \neq b_i$, y 0 en caso contrario. $M_{obs}$ representa el número de caracteres con estados conocidos.

A partir de las disimilitudes, se calculan las distancias $D_{ab}$ aplicando un factor de corrección como el de Jukes–Cantor [14] de acuerdo a la Ecuación 2:

$$D_{ab} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} P_{ab}\right). \tag{2}$$

Las distancias resultantes se almacenan en una matriz $N \times N$ simétrica, con diagonal nula. A partir de esta matriz, se pueden aplicar algoritmos como *neighbor-joining* [14] para reconstruir la filogenia agrupando iterativamente los pares de organismos $a, b$ que minimicen el valor de la Ecuación 3:

$$Q_{ab} = (n-2)D_{ab} - \sum_{k=1}^{n} D_{ak} - \sum_{k=1}^{n} D_{bk}, \tag{3}$$

donde $n$ se inicializa con el valor $N$ (número de secuencias) y va decrementándose en cada iteración de *neighbor-joining* a medida que se van agrupando organismos.

En presencia de caracteres desconocidos, las distancias se calculan solo sobre los estados observados. En el caso extremo, cuando dos secuencias no comparten ningún carácter conocido, la distancia no puede estimarse y da lugar a una entrada perdida en la matriz, tal y como se indica en la Ecuación 4:

$$\text{Matriz}[a, b] = \begin{cases} \text{Perdido} & \text{si } \forall i \in [1, M] : a_i = \text{'?'} \vee b_i = \text{'?'}, \\ D_{ab} & \text{si } \exists i \in [1, M] : a_i \neq \text{'?'} \wedge b_i \neq \text{'?'}. \end{cases} \tag{4}$$

La Figura 1 muestra un ejemplo de matriz con distancias perdidas. Este problema es frecuente en conjuntos de datos del mundo real [7,15] y tiene un impacto directo en la calidad de los análisis filogenéticos.

## 3. Descripción del Método

En esta sección se describe PHYperImpute, el método propuesto para la imputación adaptativa de matrices de distancias filogenéticas incompletas.

### 3.1. Ideas Básicas

El método PHYperImpute recibe como entrada una matriz de distancias incompleta $\mathbf{M}$ de tamaño $N \times N$, donde las entradas perdidas se representan mediante valores *NaN*. La idea clave del algoritmo radica en implementar un proceso iterativo de imputación (hasta un máximo de $\mathbf{Max_{It}}$ iteraciones), procesando $\mathbf{M}$ por columnas y utilizando la información disponible (datos observados) como referencia para realizar la selección de un modelo de imputación a partir de un conjunto $\mathbf{S}$ de modelos candidatos. El modelo que mejor se adecúe a la distribución de distancias observadas será aplicado para realizar la imputación de los valores perdidos sobre la columna procesada.

El proceso de imputación se complementa con una serie de pasos adicionales para garantizar que la matriz resultante cumple las propiedades de una matriz de distancias filogenéticas (esto es, distancias positivas y estructura simétrica). La salida del método vendrá dada por la matriz imputada y el árbol filogenético obtenido a partir de dicha matriz.

El Algoritmo 1 presenta el pseudocódigo de PHYperImpute. Las siguientes subsecciones explican en mayor detalle el proceso de imputación diseñado.

### 3.2. Preprocesado de la Matriz

Antes de iniciar la imputación, se valida y normaliza el formato de la matriz de distancias $\mathbf{M}$ a procesar. Concretamente, el método acepta matrices procedentes de ficheros de entrada bajo el formato PHYLIP [5], el cual establece en la primera fila del fichero las dimensiones de la matriz. Cada una de las filas restantes denotan las distancias de cada organismo con respecto al resto. Las distancias observadas se codifican en formato numérico (números reales) utilizando estructuras matriciales *NumPy*. En lo que respecta a las distancias perdidas, dado que estas pueden denotarse bajo diferentes representaciones ('N', '.', etc.), se comprueban dichas entradas y se codifican internamente mediante valores *NaN* dentro de la estructura matricial.

### 3.3. Procesado de la Matriz - Imputación

El proceso de imputación comienza realizándose una primera estimación inicial $\mathbf{M}^*$ consistente en reemplazar en cada columna los valores perdidos por la media de los valores observados en dicha columna, generando así una primera aproximación de imputación (línea 1 en el Algoritmo 1). A continuación, el algoritmo entra en el bucle de imputación iterativa (líneas 3–14), donde se realizará la selección de modelos y su aplicación para refinar la estimación.

**Algoritmo 1** Pseudocódigo del método PHYperImpute

**Entrada:** $\mathbf{M}$: matriz de distancias incompleta, $\mathbf{S}$: modelos de imputación candidatos, $\mathbf{Max_{It}}$: número máximo de iteraciones.
**Salida:** $\mathbf{M}^*$: matriz de distancias imputada, $\mathbf{T}$: árbol filogenético estimado.

1: $\mathbf{M}^* \leftarrow$ Realizar imputación inicial por media ($\mathbf{M}$)
2: */* Búsqueda de modelos e imputación iterativa de datos perdidos */*
3: $it \leftarrow 0$
4: **mientras** $it < \mathbf{Max_{It}}$ && $\mathbf{M}^*$ no haya convergido **hacer**
5: **para** cada columna $j \in \mathbf{M}$ con datos perdidos **hacer**
6: $Dat_{obs}, Reg_{obs}, Dat_{lost}, Reg_{lost} \leftarrow$ Identificar datos y regresores ($\mathbf{M}$, $\mathbf{M}^*$, $j$)
7: **si** No se ha detectado patrón hacia un modelo concreto **entonces**
8: $modelo_sel \leftarrow$ Seleccionar modelo ($\mathbf{S}$, $Dat_{obs}$, $Reg_{obs}$)
9: **fin si**
10: $modelo_sel \leftarrow$ Ajustar modelo ($modelo_sel$, $Dat_{obs}$, $Reg_{obs}$)
11: $\mathbf{M}^* \leftarrow$ Imputar datos perdidos en $j$ ($modelo_sel$, $Dat_{lost}$, $Reg_{lost}$)
12: **fin para**
13: $it \leftarrow it + 1$
14: **fin mientras**
15: */* Posprocesado de la matriz y generación del árbol filogenético */*
16: $\mathbf{M}^* \leftarrow$ Corregir distancias negativas ($\mathbf{M}^*$)
17: $\mathbf{M}^* \leftarrow$ Simetrizar matriz de distancias ($\mathbf{M}^*$)
18: $\mathbf{T} \leftarrow$ Reconstruir árbol filogenético ($\mathbf{M}^*$)
19: **devolver** $\mathbf{M}^*$, $\mathbf{T}$

En cada iteración del bucle se procesa cada columna $j$ con datos perdidos en la matriz de la siguiente manera. En primer lugar, se identifican para $j$ las entradas observadas y perdidas, junto con sus regresores (línea 6). Utilizando la información disponible, se realiza la evaluación y comparación de los modelos candidatos del conjunto $\mathbf{S}$, seleccionando aquel que mejor se adapte a la distribución de los datos observados (líneas 7–9). En caso de que en las columnas anteriores se haya identificado un mismo patrón de selección de modelo, este proceso de comparación puede omitirse para reducir los tiempos de cómputo, utilizando el mismo modelo seleccionado en las columnas previas para el procesamiento de las columnas restantes de la misma iteración.

La selección de modelos considera un conjunto heterogéneo de aproximaciones candidatas. En concreto, el repertorio de modelos candidatos incluye regresión lineal de mínimos cuadrados con regularización l2, *random forest*, *gradient boosting* y regresores basados en vectores de soporte en sus vertientes lineales (LinearSVR) y nu con kernel de base radial (NuSVR). Estos modelos compiten en el proceso de selección, evaluándose el rendimiento predictivo de cada aproximación sobre los datos observados de la columna activa para identificar aquella que proporciona un resultado más satisfactorio, bajo criterios RMSE y $R^2$.

Una vez seleccionado el modelo a utilizar, este se ajusta con los datos disponibles específicos de la columna y a continuación se utiliza para imputar los valores perdidos (líneas 10 y 11). Este proceso se repite hasta haber procesado todas las columnas con datos perdidos, procediéndose a hacer una nueva ite-

ración del algoritmo hasta que se satisfaga la condición de parada (alcanzar el número máximo de iteraciones o bien alcanzar convergencia).

### 3.4. Posprocesado de la Matriz

Tras haber concluido el proceso de imputación, se lleva a cabo un posprocesado de la solución para garantizar que la matriz estimada $\mathbf{M}^*$ se ajuste a las características propias de una matriz de distancias filogenéticas, comprobando posibles valores negativos o asimetría en la matriz. Con este propósito, se aplican los dos siguientes pasos. En primer lugar, se corrigen posibles distancias estimadas negativas reemplazándolas por su valor absoluto (línea 16 en el Algoritmo 1). Tras ello, en un segundo paso se realiza la simetrización de la matriz, calculando la media aritmética de cada par de entradas $(i, j)$ y $(j, i)$, esto es, $\mathbf{M}^*[i, j] = \mathbf{M}^*[j, i] = (\mathbf{M}^*[i, j] + \mathbf{M}^*[j, i])/2$ (línea 17).

Como tarea final, se emplea la matriz resultante $\mathbf{M}^*$ para realizar la inferencia del árbol filogenético $\mathbf{T}$ asociado, aplicando con este propósito el método de reconstrucción *neighbor-joining* (línea 18). Se devuelve la matriz completa e imputada $\mathbf{M}^*$ junto con la filogenia $\mathbf{T}$ en la salida del método (línea 19).

## 4. Resultados Experimentales

En esta sección se presenta la evaluación de PHYperImpute, detallando las condiciones experimentales, conjuntos de datos, métricas y resultados obtenidos.

### 4.1. Configuración Experimental

Los experimentos se han llevado a cabo sobre un sistema con procesador *AMD Ryzen 5 3600* a 3.6 GHz y 16 GB de memoria RAM, bajo el sistema operativo *Windows 10 Profesional (10.0.19045 N/D Compilación 19045)*. El software empleado incluye *Python 3.11.9*, junto con las librerías *NumPy 1.26.4*, *pandas 2.3.3*, *scikit-learn 1.8.0*, *natsort 8.4.0*, *BioPython 1.86* y *ETE3 3.1.3*.

PHYperImpute se ha comparado con otros métodos de imputación incluidos en el framework HyperImpute [10]: MIWAE, MIRACLE, GAIN, Sinkhorn, EM, SoftImpute, MICE y HyperImpute. Junto con estos ocho, se han realizado comparativas con las dos aproximaciones base de nuestro conjunto de modelos candidatos que presentaron mejor rendimiento individual, LinearSVR y NuSVR.

En estos métodos de referencia, se emplearon sus configuraciones paramétricas por defecto al no observarse desviaciones significativas con respecto a otras configuraciones, salvo en el caso de LinearSVR que precisó establecer un número máximo de iteraciones de 3000. Por su parte, la configuración de nuestro método PHYperImpute adoptó también los parámetros por defecto de sus modelos candidatos (salvo en el caso de LinearSVR, donde se aplicó la misma condición de 3000 iteraciones). El número de iteraciones del bucle de imputación de PHYperImpute fue establecido a un máximo de $Max_{It}$=2, al encontrarse que esta configuración proporcionaba un equilibrio adecuado entre la calidad de la imputación efectuada y el tiempo de ejecución requerido.

### 4.2. Conjuntos de Datos

En nuestros experimentos, se han utilizado cuatro conjuntos de datos filogenéticos con matrices de diferentes dimensiones, procedentes de la literatura [4,13]:

- $9 \times 9$: matrices procedentes de 9 secuencias de baculovirus. Incluye un total de 120 instancias del problema con porcentajes de datos perdidos del 5 % al 60 %.
- $40 \times 40$: matrices procedentes de 40 secuencias sintéticas de nucleótidos. Incluye un total de 60 instancias del problema con porcentajes de datos perdidos del 5 % al 60 %.
- $55 \times 55$: matrices procedentes de 55 secuencias del gen rbcL. Incluye un total de 120 instancias del problema con porcentajes de datos perdidos del 5 % al 60 %.
- $201 \times 201$: matrices procedentes de 201 secuencias sintéticas de nucleótidos. Incluye un total de 20 instancias del problema con porcentajes de datos perdidos del 5 % al 20 %.

En suma, se han empleado un total de 320 instancias del problema (matrices) diferentes para la evaluación experimental de la propuesta.

### 4.3. Métricas de Evaluación

Se han utilizado dos métricas para evaluar el comportamiento de PHYperImpute en comparación al resto de métodos de imputación. La primera de ellas es el *Normalized Robinson-Foulds* (NRF) [14], una métrica que permite cuantificar la calidad de la imputación midiendo la distancia topológica normalizada entre el árbol filogenético reconstruido a partir de la matriz imputada y el árbol filogenético verdad de cada conjunto de datos. Valores más bajos de NRF indican mayor similitud topológica entre ambos árboles y, por tanto, una imputación más satisfactoria. La segunda métrica considerada es el tiempo de ejecución en segundos de cada método, en línea con recomendaciones recientes para la evaluación comparativa de imputadores [9].

La métrica NRF resulta especialmente adecuada en este contexto porque evalúa directamente el impacto de la imputación sobre el objetivo final del proceso, la reconstrucción de la filogenia. En contraste a métricas puramente numéricas como el error cuadrático medio, que se limita a medir la diferencia entre valores imputados y verdaderos, el NRF cuantifica la precisión topológica del árbol filogenético resultante. Dada su importancia, el NRF constituye la métrica primaria de evaluación en este estudio desde una perspectiva biológica.

Por otra parte, el tiempo de ejecución representa una métrica adicional de interés práctico. En el ámbito de los análisis filogenéticos, los flujos de trabajo a menudo involucran tareas computacionales costosas, por lo que la inclusión de métodos de imputación con tiempos de ejecución excesivos puede limitar la aplicabilidad práctica de las soluciones propuestas. Por ello, se empleará el tiempo de ejecución como una métrica complementaria al NRF, teniendo en cuenta que la precisión de la inferencia filogenética prevalece sobre el coste computacional cuando ambos aspectos se encuentran dentro de rangos razonables.

### 4.4. Análisis de Resultados

La Tabla 1 presenta los valores medios de NRF obtenidos por cada método en los cuatro conjuntos de datos evaluados, ordenados de mejor a peor según su media global. PHYperImpute alcanza el mejor resultado medio global, con un NRF de 0,260, lo que lo sitúa como el método que presenta el comportamiento más satisfactorio en términos generales. Su calidad resulta especialmente destacable en el conjunto de datos con matrices de mayores dimensiones, $201 \times 201$, donde obtiene el mejor NRF (0,132) de la comparativa. Este resultado da cuenta de la capacidad de PHYperImpute para adaptarse a escenarios de imputación complejos que involucren grandes tamaños de matriz.

Cabe señalar que, si bien métodos individuales como NuSVR o LinearSVR obtienen resultados competitivos en conjuntos de datos específicos, estos presentan una notable variabilidad de un escenario a otro. Por ejemplo, NuSVR alcanza un NRF de 0,218 en el conjunto de datos $40 \times 40$, pero la calidad de su imputación se degrada considerablemente en el caso de $55 \times 55$ (NRF de 0,692). De forma similar, LinearSVR obtiene un buen resultado en $55 \times 55$ (0,220) pero es menos competitivo en $40 \times 40$ (0,350). PHYperImpute, por el contrario, consigue mantener una calidad de imputación más consistente en todos los escenarios evaluados, justificando así la relevancia de la aproximación adaptativa.

Con respecto al método de referencia HyperImpute, nuestra propuesta PHYperImpute consigue ir un paso más allá, mejorando los resultados en todos los conjuntos de datos considerados. Así, se consigue una reducción en la media global de NRF de 0,315 a 0,260. Los métodos restantes (Sinkhorn, SoftImpute, GAIN, MICE, MIWAE, EM y MIRACLE) dan lugar a peores resultados.

**Tabla 1.** Valores medios de NRF por conjunto de datos y media global. El mejor resultado por columna se indica en negrita. Los métodos se ordenan de menor (mejor) a mayor (peor) media global.

| Método | $9 \times 9$ | $40 \times 40$ | $55 \times 55$ | $201 \times 201$ | Media global |
|---|---|---|---|---|---|
| PHYperImpute | 0,371 | 0,282 | 0,253 | **0,132** | **0,260** |
| LinearSVR | 0,376 | 0,350 | **0,220** | 0,191 | 0,284 |
| HyperImpute | 0,415 | 0,294 | 0,356 | 0,194 | 0,315 |
| NuSVR | **0,361** | **0,218** | 0,692 | 0,161 | 0,358 |
| Sinkhorn | 0,464 | 0,358 | 0,484 | 0,246 | 0,388 |
| SoftImpute | 0,546 | 0,402 | 0,417 | 0,255 | 0,405 |
| GAIN | 0,490 | 0,617 | 0,803 | 0,283 | 0,548 |
| MICE | 0,499 | 0,558 | 0,763 | 0,396 | 0,554 |
| MIWAE | 0,574 | 0,654 | 0,771 | 0,561 | 0,640 |
| EM | 0,488 | 0,764 | 0,869 | 0,773 | 0,723 |
| MIRACLE | 0,854 | 0,840 | 0,969 | 0,832 | 0,874 |

La Tabla 2 presenta los tiempos medios de ejecución observados para cada conjunto de datos. PHYperImpute registra un tiempo medio de 14,95 segundos, un resultado razonable de acuerdo a los costes computacionales frecuentemente

observados en los análisis filogenéticos. En comparación con los otros métodos, PHYperImpute requiere tiempos medios notablemente inferiores a los de EM (163,70 segundos), MICE (117,51 segundos) y LinearSVR ejecutado de forma aislada (77,29 segundos). Además, PHYperImpute también resulta significativamente más rápido que HyperImpute (34,58 segundos). Los métodos con tiempos más reducidos, como NuSVR (3,46 segundos) o GAIN (5,04 segundos), no alcanzan la consistencia o precisión demostrada por PHYperImpute, como se ha indicado previamente. Por tanto, podemos concluir que PHYperImpute ofrece un buen equilibrio entre calidad de imputación y tiempo computacional.

**Tabla 2.** Tiempos medios de ejecución (en segundos) por conjunto de datos. El mejor resultado por columna se indica en negrita. Se remarcan en color los métodos que obtuvieron mejor calidad de imputación según NRF.

| Método | $9 \times 9$ | $40 \times 40$ | $55 \times 55$ | $201 \times 201$ | Media global |
|---|---|---|---|---|---|
| PHYperImpute | 1,12 | 1,18 | 2,23 | 55,29 | 14,95 |
| LinearSVR | **0,04** | 2,05 | 4,19 | 302,87 | 77,29 |
| HyperImpute | 13,35 | 14,65 | 85,02 | 25,29 | 34,58 |
| NuSVR | 0,05 | **0,41** | **0,58** | 12,78 | **3,46** |
| Sinkhorn | 8,40 | 11,47 | 9,10 | 15,19 | 11,04 |
| SoftImpute | 2,59 | 4,45 | 4,80 | 14,45 | 6,57 |
| GAIN | 2,69 | 3,08 | 3,36 | **11,01** | 5,04 |
| MICE | 2,18 | 11,58 | 19,44 | 436,85 | 117,51 |
| MIWAE | 1,49 | 3,02 | 3,19 | 31,11 | 9,71 |
| EM | 0,60 | 12,86 | 26,64 | 614,68 | 163,70 |
| MIRACLE | 3,32 | 12,99 | 17,75 | 75,51 | 27,39 |

### 4.5. Comparativa con el Estado del Arte

Además de la comparación con los métodos de imputación generalistas, es importante contrastar la calidad de la imputación de PHYperImpute con la obtenida por otras propuestas del estado del arte específicamente diseñadas para la imputación filogenética. A continuación, se presenta una comparativa con tres métodos de referencia en el problema: las aproximaciones basadas en *autoencoder* y *matrix factorization* de [4] y PhyloMissForest [13].

La Tabla 3 presenta la comparativa de NRF medios entre PHYperImpute y los métodos del estado del arte. En el caso de PhyloMissForest, se consideran las dos variantes de configuración reportadas por sus autores (sin y con bootstrap) [13]. Para calcular la media global, los casos "N/A" visualizados en la Tabla 3 para los métodos competidores se contabilizan con valor 1, ya que representan ejecuciones que no consiguieron alcanzar una solución válida.

PHYperImpute obtiene el mejor resultado medio de NRF en tres de los cuatro conjuntos de datos evaluados ($40 \times 40$, $55 \times 55$ y $201 \times 201$), alcanzando la mejor media global de la comparativa. Caso por caso, para el conjunto de

**Tabla 3.** Comparativa con el estado del arte: valores medios de NRF por conjunto de datos y media global. El mejor resultado por columna se indica en negrita. PhyloMissForest-nb y PhyloMissForest-b hacen referencia, respectivamente, a las configuraciones de ese método sin bootstrap y con bootstrap. "N/A" indica que el método correspondiente no consiguió una solución válida, implicando un valor de NRF = 1.

| Método | $9 \times 9$ | $40 \times 40$ | $55 \times 55$ | $201 \times 201$ | Media global |
|---|---|---|---|---|---|
| PHYperImpute | 0,371 | **0,282** | **0,253** | **0,132** | **0,260** |
| Autoencoder | 0,483 | 0,432 | 0,541 | 0,402 | 0,465 |
| Matrix factorization | 0,445 | 0,333 | 0,414 | N/A | 0,548 |
| PhyloMissForest-nb | 0,356 | 0,330 | 0,355 | 0,325 | 0,342 |
| PhyloMissForest-b | **0,348** | 0,347 | 0,305 | N/A | 0,500 |

datos más pequeño, $9 \times 9$, PHYperImpute obtiene un NRF de 0,371, superando al *autoencoder* (0,483) y a *matrix factorization* (0,445), pero quedando por detrás de la media de PhyloMissForest (0,348–0,356). No obstante, PHYperImpute verifica para este conjunto de datos un comportamiento más satisfactorio que PhyloMissForest en instancias con porcentajes elevados de datos perdidos. Por ejemplo, en las matrices $9 \times 9$ con 50 % de datos perdidos, PHYperImpute alcanzó un NRF medio de 0,420, mejorando los valores conseguidos por PhyloMissForest en dichas matrices (0,513 sin bootstrap y 0,500 con bootstrap).

En el conjunto de datos $40 \times 40$, PHYperImpute (0,282) mejora en término medio a todos los competidores, incluyendo PhyloMissForest sin bootstrap (0,330) y con bootstrap (0,347). Similarmente, en el conjunto $55 \times 55$, PHYperImpute consolida su ventaja con un NRF de 0,253, representando la aproximación más satisfactoria. Es digno señalar que, para las matrices $55 \times 55$, las soluciones generadas por PHYperImpute mejoraron los resultados del resto de métodos en todos los porcentajes considerados de datos perdidos.

Finalmente, el conjunto de datos $201 \times 201$ constituye el escenario donde las diferencias son más pronunciadas. PHYperImpute alcanza un NRF medio de 0,132, mejorando al *autoencoder* (0,402) y a PhyloMissForest sin bootstrap (0,325) por márgenes amplios. Los otros dos métodos, *matrix factorization* y PhyloMissForest con bootstrap, no pudieron completar su ejecución en una ventana temporal de 24 horas, lo que da cuenta de los problemas de escalabilidad de estos métodos al procesar matrices más complejas. PHYperImpute, por el contrario, completa las imputaciones asociadas a este escenario en un tiempo medio de 55,29 segundos, superando pues las limitaciones temporales de los competidores.

De acuerdo a los resultados obtenidos, se puede concluir que la estrategia adaptativa implementada en PHYperImpute representa una aproximación prometedora para resolver el problema de la imputación filogenética, mejorando en distintos escenarios la precisión de las estimaciones en comparación a otros métodos del estado del arte. Además, los tiempos reducidos de PHYperImpute, que completa sus imputaciones en segundos frente a aproximaciones que pueden requerir horas [13], lo posicionan como una herramienta con interés práctico para su integración dentro de los flujos de trabajo usados en filogenética computacional.

## 5. Conclusiones y Trabajo Futuro

En este trabajo se ha propuesto PHYperImpute, un método de imputación adaptativa de matrices de distancias filogenéticas incompletas basado en la selección automática de modelos de aprendizaje automático. PHYperImpute afronta el proceso de imputación mediante una aproximación iterativa donde se verifica el comportamiento de un conjunto de modelos candidatos en base a los datos observados, seleccionando la técnica que mejor se adecúe a la información disponible para afrontar la imputación y refinamiento de los datos perdidos en cada columna de la matriz. Entre los modelos soportados se incluyen técnicas de mínimos cuadrados, *random forest*, *gradient boosting* y regresores basados en vectores de soporte.

Los resultados experimentales, obtenidos a partir de cuatro conjuntos de datos que engloban un total de 320 matrices con dimensiones y porcentajes de datos perdidos diferentes, mostraron que PHYperImpute representa una aproximación robusta para abordar el problema. Utilizando NRF como medida de la calidad de la imputación, la propuesta alcanzó la mejor media global en comparación a otros diez métodos de imputación, presentando un comportamiento consistente y afrontando exitosamente la resolución del conjunto de datos de mayores dimensiones. En cuanto a los tiempos de ejecución, PHYperImpute registró un tiempo medio de ejecución de aproximadamente 15 segundos, ofreciendo un equilibrio adecuado entre calidad de imputación y coste temporal. Finalmente, las comparativas con métodos del estado del arte específicamente diseñados para imputación filogenética confirmaron la bondad de la propuesta, observándose un aumento en las mejoras obtenidas a medida que se incrementaba el tamaño de las instancias del problema consideradas.

Como líneas de investigación futuras, se plantea la integración de estrategias de computación paralela para acelerar el proceso de selección de modelos. Una segunda línea de trabajo consiste en afrontar la adaptación del método para procesar conjuntos de datos de naturaleza heterogénea, que combinen información procedente de distintos niveles biológicos (morfológicos, nucleótidos, aminoácidos, etc.) [8,12]. Asimismo, resulta de interés investigar la aplicabilidad del enfoque propuesto a otros dominios de la bioinformática donde la imputación de datos perdidos constituya un reto significativo.

## Agradecimientos

Este trabajo, respaldado por la AEI (Agencia Estatal de Investigación, España), el MICIU (Ministerio de Ciencia, Innovación y Universidades, España) y el FEDER (Fondo Europeo de Desarrollo Regional, Unión Europea), es parte del proyecto PID2022-137275NA-I00 (X-BIO) financiado por MICIU/AEI/10.13039-/501100011033 y por FEDER, UE. Este trabajo también ha sido parcialmente financiado por la Junta de Extremadura y el FEDER mediante la ayuda GR24017.

## Referencias


1. Alabadla, M., Sidi, F., Ishak, I., et al.: Systematic Review of Using Machine Learning in Imputing Missing Values. IEEE Access **10**, 44483–44502 (2022). https://doi.org/10.1109/ACCESS.2022.3160841
2. Asanka, P.D., Rajapakshe, C., Takahashi, M.: Adaptive Multi-agent System for Dynamic Missing Value Imputation in Time Series. New Generation Computing **44**, 9 (2026). https://doi.org/10.1007/s00354-025-00312-6
3. Attwood, S.W., Hill, S.C., Aanensen, D.M., et al.: Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nature Reviews Genetics **23**(9), 547–562 (2022). https://doi.org/10.1038/s41576-022-00483-8
4. Bhattacharjee, A., Bayzid, M.S.: Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. BMC Genomics **21**, 497 (2020). https://doi.org/10.1186/s12864-020-06892-5
5. Christensen, H., Olsen, J.E.D.: Introduction to phylogenetic analysis of molecular sequence data. In: Introduction to Bioinformatics in Microbiology, pp. 111–130. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-45293-2_6
6. Emmanuel, T., Maupong, T., Mpoeleng, D., et al.: A survey on missing data in machine learning. Journal of Big Data **8**, 140 (2021). https://doi.org/10.1186/s40537-021-00516-9
7. Featherstone, L.A., McGaughran, A.: The effect of missing data on evolutionary analysis of sequence capture bycatch, with application to an agricultural pest. Molecular Genetics and Genomics **299**, 11 (2024). https://doi.org/10.1007/s00438-024-02097-7
8. Hunt, R., Reyes-Hernández, J.L., Shaw, J.J., et al.: Integrating Deep Learning Derived Morphological Traits and Molecular Data for Total-Evidence Phylogenetics: Lessons from Digitized Collections. Systematic Biology **74**(3), 453–468 (2025). https://doi.org/10.1093/sysbio/syae072
9. Jäger, S., Allhorn, A., Bießmann, F.: A Benchmark for Data Imputation Methods. Frontiers in Big Data **4**, 693674 (2021). https://doi.org/10.3389/fdata.2021.693674
10. Jarrett, D., Cebere, B., Liu, T., et al.: HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. In: Proceedings of the 39th International Conference on Machine Learning. pp. 9913–9937 (2022), https://proceedings.mlr.press/v162/jarrett22a/jarrett22a.pdf
11. Kocot, K.M., Struck, T.H., Merkel, J., et al.: Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. Systematic Biology **66**(2), 256–282 (2017). https://doi.org/10.1093/sysbio/syw079
12. Moura, M.R., Ceron, K., Guedes, J.J.M., et al.: A phylogeny-informed characterisation of global tetrapod traits addresses data gaps and biases. PLOS Biology **22**(7), e3002658 (2024). https://doi.org/10.1371/journal.pbio.3002658
13. Pinheiro, D., Santander-Jiménez, S., Ilic, A.: PhyloMissForest: a random forest framework to construct phylogenetic trees with missing data. BMC Genomics **23**, 377 (2022). https://doi.org/10.1186/s12864-022-08540-6
14. Singh, G.B.: Introduction to Phylogeny. In: Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB, pp. 265–281. Springer (2025). https://doi.org/10.1007/978-3-031-75694-8_13
15. Yan, B., Dietrich, C.H., Yu, X.F., et al.: Missing data and model selection in phylogenomics: A re-evaluation of Cicadomorpha (Hemiptera: Auchenorrhyncha) superfamily level relationships. Journal of Systematics and Evolution **63**(5), 1203–1216 (2025). https://doi.org/10.1111/jse.13190