
**UNIVERSIDAD DE BUENOS AIRES**

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

# Técnicas de aprendizaje profundo en la predicción del poder de frenado de iones en materiales

Tesis presentada para optar al título de Magíster de la Universidad de Buenos Aires en Explotación de Datos y Descubrimiento de Conocimiento

Tesista: Lic. Felipe Bivort Haiek

Director de tesis: Dr. Darío Mitnik

Co-directora de tesis: Dra. Claudia Montanari

Buenos Aires, 12 de Septiembre 2025

# Contenidos

## Resumen

El estudio del poder de frenado (*stopping power*) de partículas cargadas al atravesar la materia es fundamental en múltiples áreas científicas y tecnológicas. Es esencial para comprender y modelar procesos de interacción ion-materia, con aplicaciones en campos como la terapia con iones, la modificación de materiales mediante bombardeo iónico y la simulación de transporte en entornos de altas energías. Sin embargo, calcular con precisión este parámetro representa un desafío considerable debido a la complejidad de los procesos de interacción entre iones y electrones, y a la escasez de datos experimentales sistematizados.

La base de datos de poder de frenado de la Agencia Internacional de Energía Atómica (IAEA) constituye una fuente integral que recopila mediciones experimentales realizadas durante casi un siglo. No obstante, a pesar de su valor para la comunidad científica global, su estructura original presentaba diversos problemas, tales como formatos heterogéneos, duplicación de registros, carencia de metadatos normalizados y presencia de valores atípicos u obsoletos, que dificultaban su utilización.

En este trabajo presentamos un enfoque sistemático para procesar, limpiar y aprovechar esta base de datos mediante técnicas de aprendizaje automático (*Machine Learning*). En primer lugar, se realizó un reformateo completo para garantizar su compatibilidad con herramientas analíticas contemporáneas. Posteriormente, se implementó un protocolo automatizado de limpieza de datos, basado en Autoencoders de Eliminación de Ruido (*Denoising Autoencoders*) y un método heurístico construido sobre DBSCAN, que permitió identificar y eliminar valores inconsistentes preservando la integridad de los datos experimentales.

Con el conjunto de datos depurado, desarrollamos un modelo de red neuronal profunda denominado ESPNN (*Electronic Stopping Power Neural Network*), capaz de predecir con alta precisión los poderes de frenado electrónicos para cualquier combinación ion-blanco dentro de un amplio rango de energías. Al ser validado con datos de prueba reservados y conjuntos de validación temporal, ESPNN superó a los mejores modelos teóricos existentes, mostrando mejoras promedio superiores al 20 % en el

error porcentual absoluto medio (MAPE). Además, ampliamos nuestra metodología para incluir blancos multielementales, desarrollando un enfoque original que incorpora representaciones vectoriales mediante la arquitectura de redes neuronales de paso de mensajes SCHNET (*Message-Passing Neural Network*). Esta extensión permitió obtener resultados igualmente prometedores.

El código de inferencia de ESPNN y la base de datos procesada se encuentran disponibles públicamente, ofreciendo a la comunidad científica una herramienta moderna y precisa para la predicción del poder de frenado en materiales mono y multielementales.

**Deep Learning Techniques towards the prediction of Stopping Power in materials. Abstract**

The study of the stopping power of charged particles in matter is fundamental in multiple scientific and technological fields. It is essential for understanding and modeling ion–matter interaction processes, with applications in areas such as ion therapy, materials modification by ion bombardment, and transport simulations in high-energy environments. However, accurately calculating this parameter poses a considerable challenge due to the complexity of ion–electron interaction processes and the scarcity of systematic experimental data.

The stopping power database of the International Atomic Energy Agency (IAEA) constitutes a comprehensive source compiling experimental measurements collected over nearly a century. Despite its value to the global scientific community, the original structure presented several issues, including heterogeneous formats, duplicated records, a lack of standardized metadata, and the presence of outliers or obsolete values, which hindered its practical use.

In this work, we present a systematic approach to process, clean, and leverage this database using Machine Learning techniques. First, a complete reformatting was performed to ensure compatibility with contemporary analytical tools. Subsequently, an automated data-cleaning protocol was implemented, based on Denoising Autoencoders and a heuristic method built upon DBSCAN, which enabled the identification and removal of inconsistent values while preserving the integrity of the experimental data.

Using the cleaned dataset, we developed a deep neural network model named ESPNN (Electronic Stopping Power Neural Network), capable of accurately predicting electronic stopping powers for any ion–target combination within a wide range of energies. When validated against held-out test data and temporal validation sets, ESPNN outperformed existing state-of-the-art theoretical models, achieving average improvements of over 20 % in the mean absolute percentage error (MAPE). Moreover, we extended our methodology to include multielement targets by developing an original approach that incorporates vector representations through the message-passing neural network architecture SCHNET. This extension yielded equally promising results.

The inference code ESPNN and the processed database are publicly available, providing the scientific community with a modern and accurate tool for predicting stopping power in both mono- and multielemental materials.

A Darío por enseñarme lo que es un ampersand y las charlas y la motivación durante el desarrollo de esta investigación.

A Claudia por sus conocimientos de Stopping Power que tan generosamente compartió, y por su paciencia.

A Coco por enseñarme el valor de la ciencia y los libros.

A $\rho$ por que sin ella a $\phi$ le falta un eje.

A Epicteto, Marco Aurelio, Séneca y Rúfus por sus enseñanzas.

”Interpolar es humano,
extrapolar es divino.”

---

# Capítulo 1

# Introducción

## 1.1. Aprendizaje automatizado

En los últimos años se ha notado un creciente interés en el uso de herramientas de aprendizaje automatizado, *machine learning* (ML), para resolver problemas físicos en sistemas complejos de difícil solución. Estas técnicas se implementan en diversas áreas, como por ejemplo en cosmología, física cuántica de muchos cuerpos, computación cuántica, y en física de materiales [1]. El objetivo no consiste únicamente en mejorar los tiempos de cómputo, sino también en intentar obtener nuevos métodos de resolución de las ecuaciones diferenciales básicas [2].

El aprendizaje por datos ha producido importantes cambios en los paradigmas de diversas disciplinas. En este proceso, no ha resultado inmune la química computacional, que está incorporando métodos de ML en la predicción de energías y propiedades de moléculas y sólidos, con aplicaciones que crecen en popularidad dramáticamente. Estos trabajos han cambiado el rol de los métodos de ML, que dejaron de ser una mera herramienta de clasificación, para involucrarse directamente en las raíces teóricas de la química computacional. Los avances en este área son muy promisorios, hasta el punto en el que "se vislumbra un futuro en el cual el diseño, la síntesis, la caracterización y las aplicaciones de moléculas y materiales será acelerado por la inteligencia artificial" [3].

La naturaleza cuántica de las interacciones hace que la evaluación computacional de las estructuras atómicas y moleculares resulte sumamente costosa. Esto convierte a ML en un excelente candidato para tratar este tipo de cálculos. Los usos de redes neuronales para la simulación de potenciales e interacciones atomísticas han sido examinados y compilados por Behler [4]. Podemos destacar entre ellos el estudio de sistemas cuánticos de muchos cuerpos [5], las novedosas técnicas de modelado y representación de las interacciones atómicas [6, 7], la identificación de estructuras con propiedades específicas, el cálculo de superficies energéticas, y los cálculos de funcionales de densidades, entre otros [4].

En el área específica de la física atómica, si bien existen algunos intentos de resolución de la ecuación de Schrödinger empleando redes neuronales profundas (*deep neural networks* – DNN) aún no se han visto trabajos que involucren estados del continuo y procesos colisionales resueltos mediante aprendizaje automatizado. Mills *et al.* [8] emplean una red convolucional para resolver potenciales bidimensionales. Hermann *et al.* [7] obtienen funciones de onda introduciendo modificaciones en la estructura de capas, y utilizando orbitales de Hartree–Fock como base de partida, lo que permite incorporar de manera natural las restricciones de espín y simetrías. El modelo logra reproducir correlaciones interelectrónicas con precisión química, incluso en sistemas complejos de hasta 30 electrones. Comparado con los cálculos típicos realizados mediante métodos de Monte Carlo variacional, que requieren de cientos o miles de determinantes, esta DNN utiliza sólo una decena de ellos, mostrando excelentes perspectivas para su aplicación en problemas más complejos. Existen además herramientas como el SchNetPack [6], basada en redes neuronales para grafos, con una capa de convolución de filtros continuos (*continuous filter convolutional layer*) que integra información del entorno de cada electrón. Esta herramienta ha sido utilizada para evaluar diversas propiedades atómicas y moleculares (energías, momento dipolar, momento cuadrupolar magnético), y también se ha aplicado en sistemas periódicos (cristales), alcanzando resultados de vanguardia (*state-of-the-art*, SOTA).

Dado el éxito obtenido por estos modelos en cálculos de estructura atómica, resulta interesante explorar la posibilidad de adaptar sus arquitecturas al estudio de procesos de colisión. En estos casos, la complejidad es considerablemente mayor, ya

que se involucran estados del continuo y una gran cantidad de procesos resultantes de las interacciones entre el proyectil y los electrones del blanco. Estos fenómenos físicos comprenden sistemas de muchos cuerpos, múltiples procesos colisionales e interacciones, lo que dificulta la aplicación de cálculos desde primeros principios. Además, las teorías existentes suelen estar restringidas a ciertos regímenes energéticos: algunas describen adecuadamente colisiones a bajas energías, mientras que otras son válidas solo en el régimen de altas energías. Por estos motivos, en la presente Tesis nos proponemos utilizar técnicas de aprendizaje profundo para la predicción del poder de frenado de iones en materiales.

## 1.2. Stopping Power

### 1.2.1. Concepto físico

El poder de frenado o *stopping power* (SP) se define como la pérdida de energía del proyectil por unidad de camino al ser disparado contra un blanco denso. Esta cantidad depende tanto de las características materiales del proyectil y del blanco, como de la energía del impacto. La definición formal del poder de frenado $S$ es

$$S(E_0) = -\left.\frac{dE}{dx}\right|_{E_0}, \tag{1.1}$$

donde $x$ representa la longitud de penetración y $E_0$ es la energía de la partícula incidente [9]. También es usual expresarlo como sección eficaz de frenado

$$\sigma(E_0) = \frac{S(E_0)}{\delta}, \tag{1.2}$$

siendo $\delta$ la densidad del blanco.

La Figura 1.1 representa lo que ocurre en el trayecto de un ion de alta energía incidente, mientras se frena al atravesar un medio material. La primera parte (la más larga del trayecto, antes de llegar al máximo), está asociada a las interacciones inelásticas del proyectil con los electrones del blanco (por ejemplo, excitaciones, io-

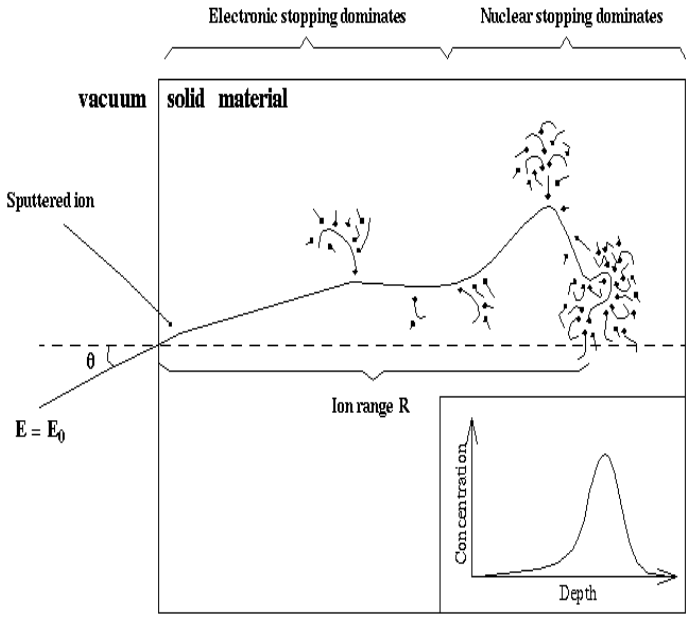



Figura 1.1: Ilustración del fenómeno de colisiones que da lugar al stopping power. Un proyectil ingresa con ángulo $\theta$ a un medio denso donde es repetidamente frenado por la interacción con electrones y núcleos del material.

nización, etc.). Esto es lo que se conoce como stopping electrónico. A medida que el ion desarrolla más interacciones con el medio, la pérdida de energía aumenta hasta llegar a un máximo. En esa región, la pérdida de energía es suficientemente elevada como para hacer detener al ion incidente. Cuando el ion se frena lo suficiente, puede interactuar con los núcleos, produciendo lo que se conoce como *stopping nuclear*. Estos procesos combinados producen como resultado un frenado abrupto, que detiene completamente al ion. La curva que describe la pérdida de energía en función de la penetración se conoce como curva de Bragg. La misma se ilustra en la parte inferior de la Figura 1.1. Se caracteriza por un máximo claro y asimétrico (pico de Bragg) y una caída abrupta, hasta la posición en que el ion se detiene.

La propiedad peculiar que muestra la curva de Bragg tiene una enorme importancia en diversas aplicaciones, tanto tecnológicas como médicas: implantación iónica en materiales, evaluación del daño en dispositivos de reactores de fisión nuclear, pre-

dicción de la penetración de iones en equipos espaciales y exposición a radiación en tejidos biológicos [10], entre otras. Este último aspecto dio origen a la terapia de cáncer por protones, en la cual se busca direccionar el proyectil con la energía adecuada para que el pico de Bragg ocurra exactamente en el lugar donde se encuentra el tumor [11]. La característica caída casi vertical del *stopping power* ofrece una gran ventaja frente a terapias como la radioterapia convencional. Esta permite minimizar el daño a tejidos más profundos o circundantes. La Figura 1.2 muestra la comparación entre la pérdida de energía de fotones (*photon beam*, curva verde) al penetrar tejido biológico, y la pérdida de energía altamente localizada de protones (*proton beam*, curva roja). La curva azul en la Figura 1.2 ilustra cómo, mediante la combinación de haces de protones que atraviesan materiales (films) previos al tejido, es posible "dibujar.[el] espesor del tumor a tratar.

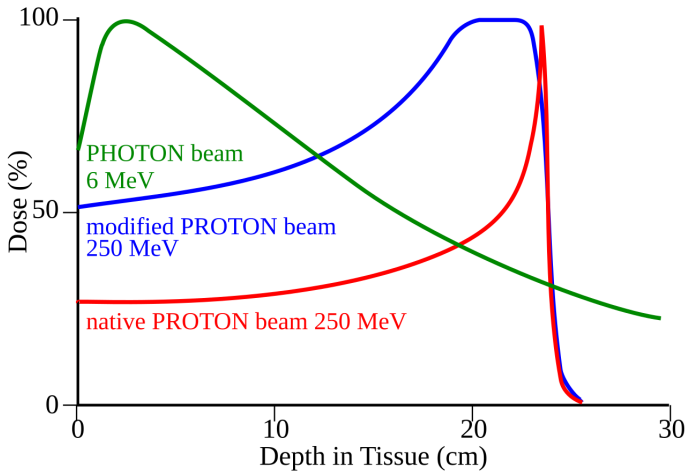



Figura 1.2: Pérdida de energía en un medio biológico en función de la profundidad. Comparación entre terapia de rayos (*photon beam*) y protonterapia (*proton beam*)

### 1.2.2. Estado del arte

Dado que el *stopping power* (SP) total es la suma de las contribuciones electrónica y nuclear, es necesario distinguir claramente entre ambos procesos. Para que un ion incidente pueda excitar o ionizar electrones del blanco, su energía debe superar un

cierto umbral (*threshold*). Por lo tanto, a muy bajas energías, solo se considera el frenamiento debido a colisiones con los núcleos del blanco. Sin embargo, a energías del orden de unos pocos miles de eV, el frenamiento nuclear se vuelve completamente despreciable frente al electrónico. El *stopping* electrónico es el principal responsable de la aparición del pico de Bragg —el máximo de la curva de pérdida de energía—, por lo que en esta Tesis nos concentraremos exclusivamente en esta contribución. En lo sucesivo, toda mención al *stopping power* hará referencia únicamente a su componente electrónica. Del mismo modo, todos los datos experimentales y modelos considerados en este trabajo corresponden a *stopping* electrónico.

La pérdida de energía de iones en materiales es el resultado de distintos procesos colisionales inélasticos. Su cálculo desde primeros principios implica la determinación de probabilidades de transición de los electrones del blanco a cualquier estado distinto del inicial. Estas transiciones se producen debido a la transferencia de energía del ion incidente a los electrones del blanco. Dada la enorme complejidad que conlleva el cálculo *ab initio* de estos procesos, la mayoría de las predicciones teóricas se obtienen a través de aproximaciones semi–empíricas.

El interés que despierta este proceso, ha dado lugar a una enorme cantidad de mediciones experimentales, produciendo datos que están diseminados en diversas publicaciones. En una tarea monumental, un pionero en el tema, el Dr Helmut Paul [12], inició a comienzos de 1990 la recopilación de datos experimentales de SP. De esta manera generó una importantísima base de datos abierta, con sede en la Universidad de Linz, que ha sido utilizada mundialmente. Esta base de datos, disponible en la página web de la Universidad de Linz [13], incluye resultados desde las primeras mediciones que se hicieron en 1928 (Rosemblum, 1928 [14]), hasta experimentos reportados en 2015. En ese año, Paul se retiró, y delegó la curaduría y responsabilidad del mantenimiento y actualización de la base, a la Dra. Claudia Montanari, co–directora de esta Tesis. Desde entonces la base de datos pasó a estar bajo la tutela de la *International Atomic Energy Agency* (IAEA) [15], y el sitio web correspondiente [16] se ha constituido en el portal de referencia mundial de datos de SP [17, 18], el cual tomaremos como base para este trabajo de Tesis de Maestría.

Podríamos mencionar una muy vasta bibliografía para referirnos a SP. Algunas

reseñas recientes pueden encontrarse en [9, 18–21]. Entre los numerosos algoritmos diseñados con fines tecnológicos, se destacan los programas computacionales que se utilizan para el estudio espacial o para aplicaciones médicas [10, 22, 23]. Respecto a las aproximaciones semi–empíricas para el cálculo de SP, el código mas utilizado es, sin duda, el SRIM (*Stopping and Range of Ions in Matter*) [24, 25], que con decenas de miles de descargas anuales proporciona una idea del impacto que puede tener la presente investigación y sus resultados. En general, existe concordancia entre los diferentes resultados experimentales a altas energías, y entre ellos y las predicciones de SRIM, que son considerados como datos de referencia. En las regiones de energías intermedias (alrededor del máximo de SP) y bajas, la tendencia no es tan clara, notándose diferencias entre los datos más antiguos y otros posteriores. Por otro lado, hemos señalado [18] experimentos recientes que muestran discrepancias tanto con datos previos como con las predicciones de SRIM, lo que pone en evidencia la necesidad de buscar nuevas perspectivas y acercamientos al tema, tales como las que se propone en este trabajo.

## 1.3. Objetivos de esta Tesis

El objetivo *principal* de esta tesis es contribuir a mejorar los sistemas de predicción de SP electrónico de iones en materiales utilizando modelos de aprendizaje profundo a partir de la base de datos experimentales de IAEA [16]. Los objetivos *específicos* para lograrlo son los siguientes:

- **Analizar la conformación de la base de datos:**
  - Realizar un análisis exploratorio exhaustivo de los datos (*EDA*)
  - Identificar patrones y correlaciones entre variables
  - Evaluar la calidad y completitud de los datos
  - Determinar la distribución estadística de las variables
  - Detectar valores atípicos y anomalías (*outliers*)

- **Desarrollar técnicas para limpiar el ruido en los datos:**
  - Implementar y evaluar sistemas de eliminación de ruido (*denoising*) basados en codificadores automáticos (*autoencoders*)
  - Desarrollar un método de heurísticas basado en DBSCAN para la detección de valores atípicos
  - Comparar diferentes técnicas de imputación de datos faltantes
  - Establecer métricas de calidad para evaluar la efectividad de la limpieza
  - Validar los resultados mediante técnicas de validación cruzada (*cross-validation*)
- **Desarrollar modelos predictivos para el caso de blancos mono-elementales:**
  - Realizar un análisis comparativo de diferentes arquitecturas de redes neuronales
  - Optimizar hiperparámetros mediante técnicas de búsqueda sistemática
  - Implementar técnicas de regularización y prevención del sobreajuste (*overfitting*)
  - Evaluar el rendimiento mediante múltiples métricas de desempeño
  - Comparar resultados con modelos de referencia (*baseline*) y estado del arte
  - Analizar la interpretabilidad de los modelos desarrollados
- **Desarrollar modelos predictivos para el caso de blancos complejos:**
  - Diseñar arquitecturas de redes neuronales adaptadas a predicciones multiobjetivo
  - Implementar técnicas de aprendizaje por transferencia (*transfer learning*) cuando sea aplicable
  - Optimizar la selección de características mediante técnicas de importancia de características (*feature importance*)

- Evaluar la relación compromiso-desempeño (*trade-off*) entre complejidad del modelo y rendimiento
- Realizar un análisis exhaustivo de resultados comparativos con el estado del arte
- Documentar las mejores prácticas y lecciones aprendidas

# Capítulo 2

# Base de datos

## 2.1. Estado de la base y procesamiento

El presente trabajo se basa en los resultados experimentales de *Stopping Power* recopilados en la base de datos de la IAEA, disponibles en la página web [26]. Para poder utilizar estos datos como entradas en los procedimientos de aprendizaje automático, tuvimos que realizar un arduo trabajo de curado, mediante el cual logramos transferir la información de la base de datos a tablas de lectura simple. Dada la característica de estos datos, por ejemplo, que no proceden de una única fuente, que no están publicados con nomenclatura uniformada, y que tampoco se expresan en las mismas unidades, tuvimos que desarrollar métodos iterativos, que detallamos a continuación. La base de datos original estaba organizada en directorios clasificados según el ion incidente, nombrados con el formato `ANombre`, donde `A` representa la masa del ion (un número entero en unidades atómicas) y `Nombre` corresponde al símbolo del elemento según la tabla periódica. Por ejemplo: '01H' o '04He'. No obstante, se detectaron algunas excepciones a esta convención, como el directorio "40AR", que fue renombrado en el código como "40Ar"para asegurar la compatibilidad con otras bases de datos. Esta estandarización fue necesaria ya que, durante el entrenamiento de los modelos, se incorporan datos complementarios provenientes de otras fuentes que generalmente utilizan la nomenclatura estándar de la tabla periódica. Dentro de cada

carpeta de un dado proyectil, se encontraban los archivos correspondientes a cada blanco, nombrados en forma no homogénea: `[proyectil][blanco]`, otros solamente `[blanco]` sin mencionar el proyectil, y otros `[blanco][número de referencia]`. En el caso de blancos atómicos, el uso de la nomenclatura basada en los símbolos de la tabla periódica resultó, en general, unívoco; la única excepción fue el uso de "OS"para oxígeno sólido, que podría confundirse con "Os"(osmio). En cambio, en el caso de blancos moleculares, los nombres presentaban una gran variabilidad: algunos se daban como fórmulas estequiométricas, otros como el nombre de la molécula (a veces completo, otras abreviado), e incluso se utilizaban nombres comerciales no científicos. Algunos ejemplos: *Dmam* para dimetilamino, la fórmula estequiométrica CH4 para el Metano, *Pen* para Pentano, *Mylar* y *Kapton* (marcas registradas de DuPont Coorporation) para los films de polyethylene terephthalate y polyimide, respectivamente. Incluso algunas daban origen a ambigüedades o duplicaciones, como "Hexav"por Hexane vapor en la carpeta He. Otra consideración importante que tuvimos fue que no todas los archivos respondían ni a la misma cantidad, ni ordenamiento de columnas, ni mantenían un formato de la separación decimal consistente. Para resolver esto hicimos un procesamiento línea a línea de los archivos teniendo en cuenta las peculiaridades de todos los archivos, generando así nuevas tablas `[proyectil]consolidado.csv`. La estructura de la base para los siguientes procesamientos queda dada por la Figura 2.1.

Otro aspecto tedioso con el que hemos tenido que lidiar fue la falta de consistencia en las unidades, tanto para la energía del proyectil como para el poder de frenado. Típicamente, la energía estaba expresada en `keV`, `MeV`, o en `keV/u` y `MeV/u` (energía por unidad de masa). Incluso, en algunos casos, se utilizaban unidades como energía equivalente de protones o de helio. Esta variabilidad se debe a la inclusión de datos correspondientes a proyectiles isotópicos. Por ejemplo, en el caso del hidrógeno (H), se incluían datos para $^{1}$H, $^{2}$H (deuterio) y $^{3}$H (tritio), cuyos valores de SP pueden compararse más directamente cuando se expresan en función de la energía por unidad de masa. Por otro lado, en los valores de SP también encontramos diversidad de unidades: `eV/Å` (SP por unidad de longitud, como en la ecuación (1.1)), `eV cm`$^2$`/atom` o `eV cm`$^2$`/mg` (sección eficaz de SP por átomo o por unidad de masa, como en la ecuación (1.2)). Se crearon tablas con el formato `[tipo de proyectil]unidades.csv`

para asociar unidades a cada par blanco–proyectil, a partir del análisis de los gráficos publicados en el sitio web de la IAEA [26], considerando el título, las unidades y las etiquetas de los ejes. Fue necesario tener en cuenta que, en el eje $x$ (correspondiente a la energía incidente en todos los gráficos), no se utiliza una nomenclatura uniforme. Por ejemplo, aparecen combinaciones como "`Energía (keV/amu)`.º "`Energía por protón (keV)`", que en realidad hacen referencia a la misma unidad, lo que añade una complejidad adicional al momento de normalizar los datos en un sistema de unidades coherente.

Otros archivos de gran utilidad presentes en la base original [26] son `hscsc.txt`, `hescs.txt` y `heavyscs.txt`. Cada entrada de estas tablas corresponde a un experimento para un determinado par blanco–proyectil, e incluye la masa del proyectil (lo que permite distinguir entre isótopos), el nombre del proyectil y un código de referencia a la publicación original. Este último debe buscarse en el archivo `referencias.txt`, el cual contiene las citas completas de las publicaciones científicas de las que fueron extraídos los datos.

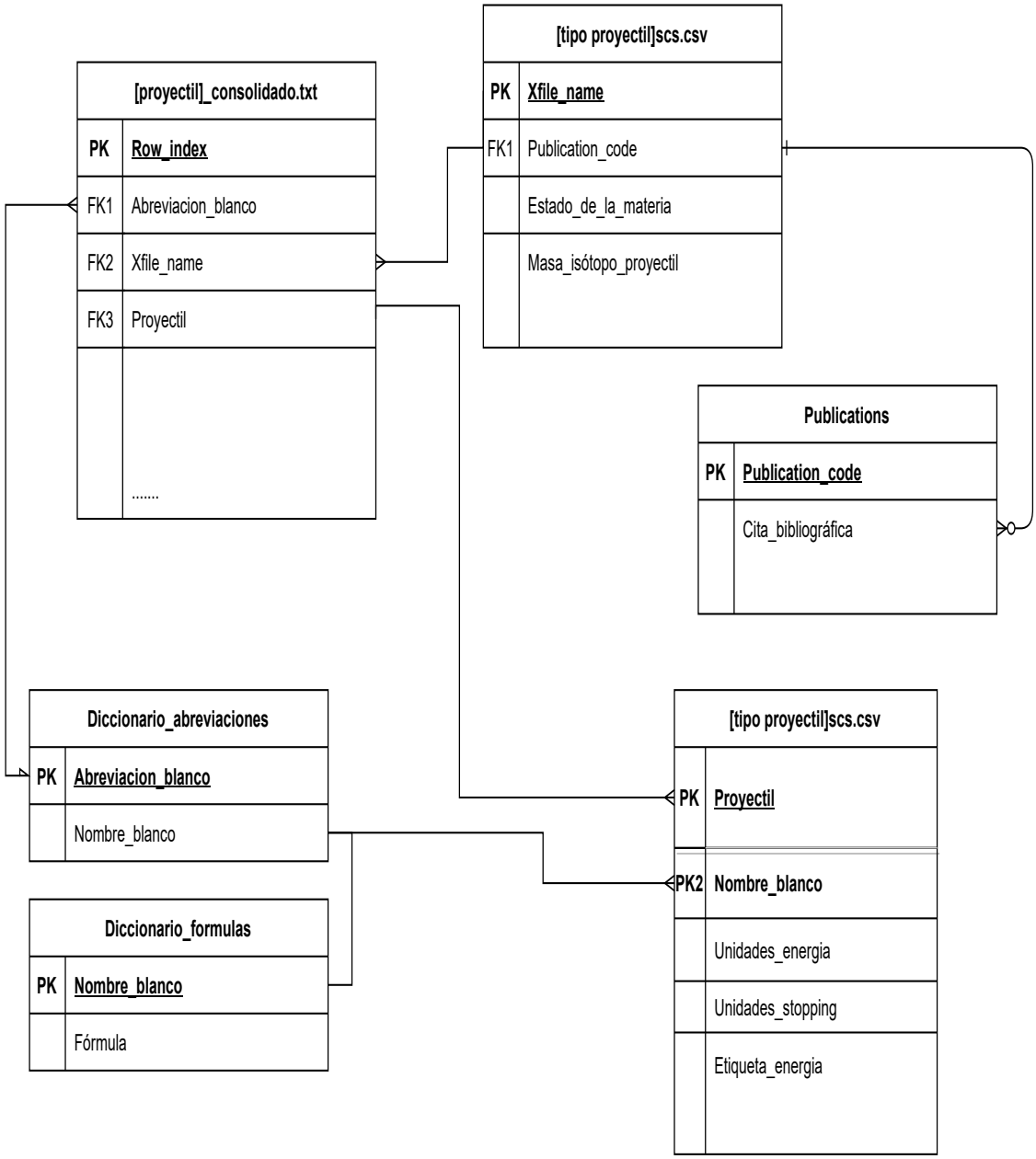



Figura 2.1: Diagrama de entidad relación.

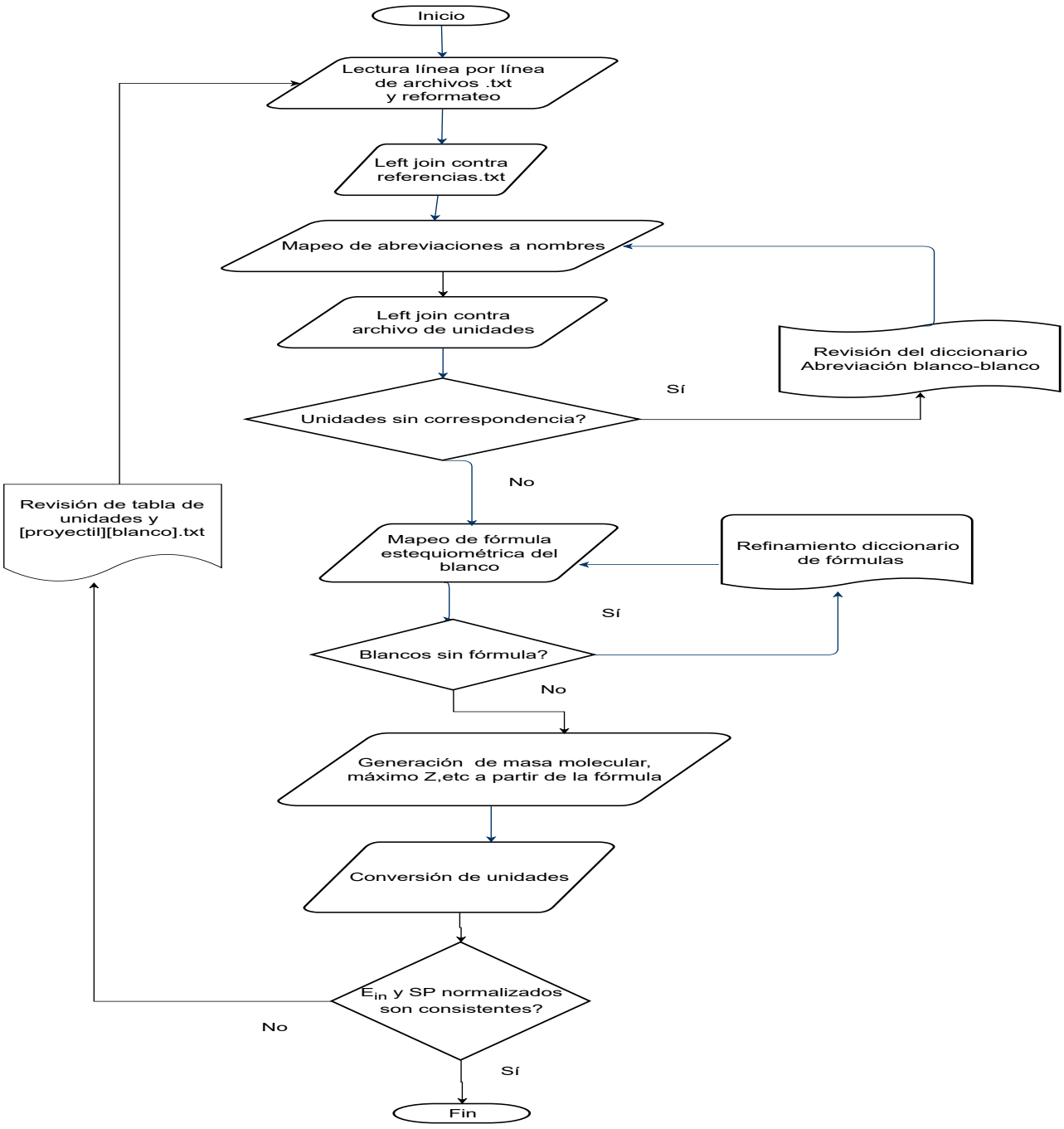



Figura 2.2: Diagrama de flujo resumido del manejo de la base de datos.

Con el fin de procesar todos los datos en igualdad de condiciones, se decidió homogeneizar las unidades de poder de frenado a $\texttt{MeV·cm}^2\texttt{/mg}$, y las unidades de energía incidente a `keV/amu`. Dado que los proyectiles no siempre correspondían al mismo isótopo, fue necesario documentar cuidadosamente la masa del proyectil asociada a cada experimento, para poder normalizar correctamente la energía incidente, $E_{in}$, en unidades de energía por unidad de masa. Aun así, persisten casos problemáticos, como el del LR-115, un compuesto muy utilizado en aceleradores para el trazado de trayectorias de partículas, que presenta distintas conformaciones estructurales. Esta ambigüedad impide determinar con certeza la masa molecular de la muestra utilizada en el experimento correspondiente. El proceso completo, así como los puntos de control de consistencia utilizados para construir las bases de datos específicas de cada proyectil, se presentan en la Figura 2.2. Una vez obtenidas las tablas con los valores de $E_{in}$ y SP ya normalizados, en la etapa final del procesamiento se verificaron valores extremos tanto de energía incidente como de SP. Esto permitió identificar errores en la carga de datos, como por ejemplo comas mal colocadas, que derivaban en curvas anómalas como las mostradas en la Figura 2.3. Estas anomalías fueron detectadas analizando la máxima diferencia entre puntos consecutivos de SP. Finalmente, algunas discrepancias entre los datos numéricos de los archivos de la base y los gráficos publicados en el sitio de la IAEA sólo pudieron resolverse consultando directamente las publicaciones originales.

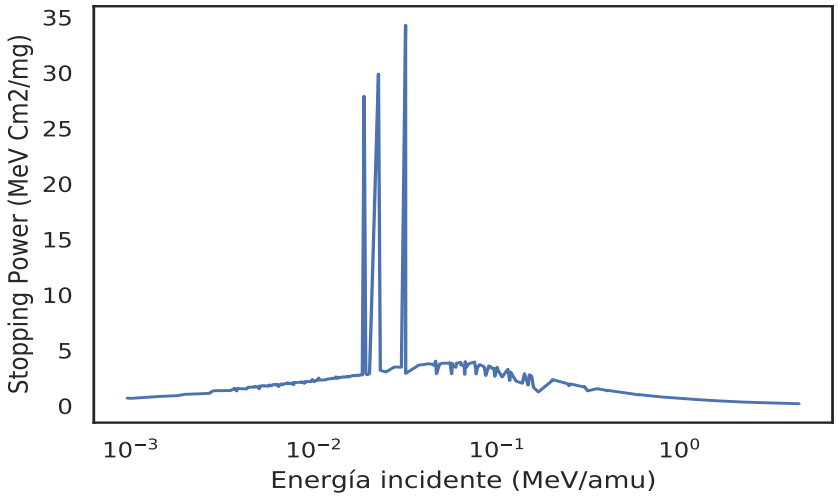



Figura 2.3: Curva de *Stopping Power* con datos mal entrados entre las energías $10^{-2}$ y $10^{-1}$ MeV/amu

La base limpia sirvió de sustento para la modernización de la base de SP de IAEA, según se describe en [18]. La versión original [26] ha sido reemplazada por la versión actual [16].

La modernización de la base de datos de poder de frenado (SP) fue parte de una pasantía realizada en la sede de la IAEA, en el Vienna International Center. Durante esta estancia, también se incorporaron herramientas de búsqueda y visualización que permiten explorar la base por combinación ion–blanco (más de 1400 combinaciones) o por autor (más de 3000, correspondientes a unas 700 publicaciones). Además, se vincularon directamente las referencias bibliográficas mediante identificadores DOI. Por otro lado, se generaron gráficos para todos los sistemas ion–blanco, ya que en la base original [26] solo se graficaban aquellos con al menos dos conjuntos de datos. Las nuevas visualizaciones, disponibles en [16], están implementadas con figuras interactivas de PLOTLY. Un ejemplo se muestra en la Figura 2.4, correspondiente al sistema H sobre Au. Como se observa, es posible seleccionar los puntos por publicación o estado, inspeccionar sus valores, excluirlos del gráfico o cambiar las unidades de visualización. Para el poder de frenado, pueden elegirse `eV·cm²/atom` o `MeV·cm²/mg`; para la energía incidente, `MeV` o `MeV/amu`. En el panel inferior, se listan las publica-

ciones correspondientes a los datos incluidos en el gráfico, junto con sus enlaces DOI, y se ofrece la posibilidad de descargar las tablas con los datos originales.

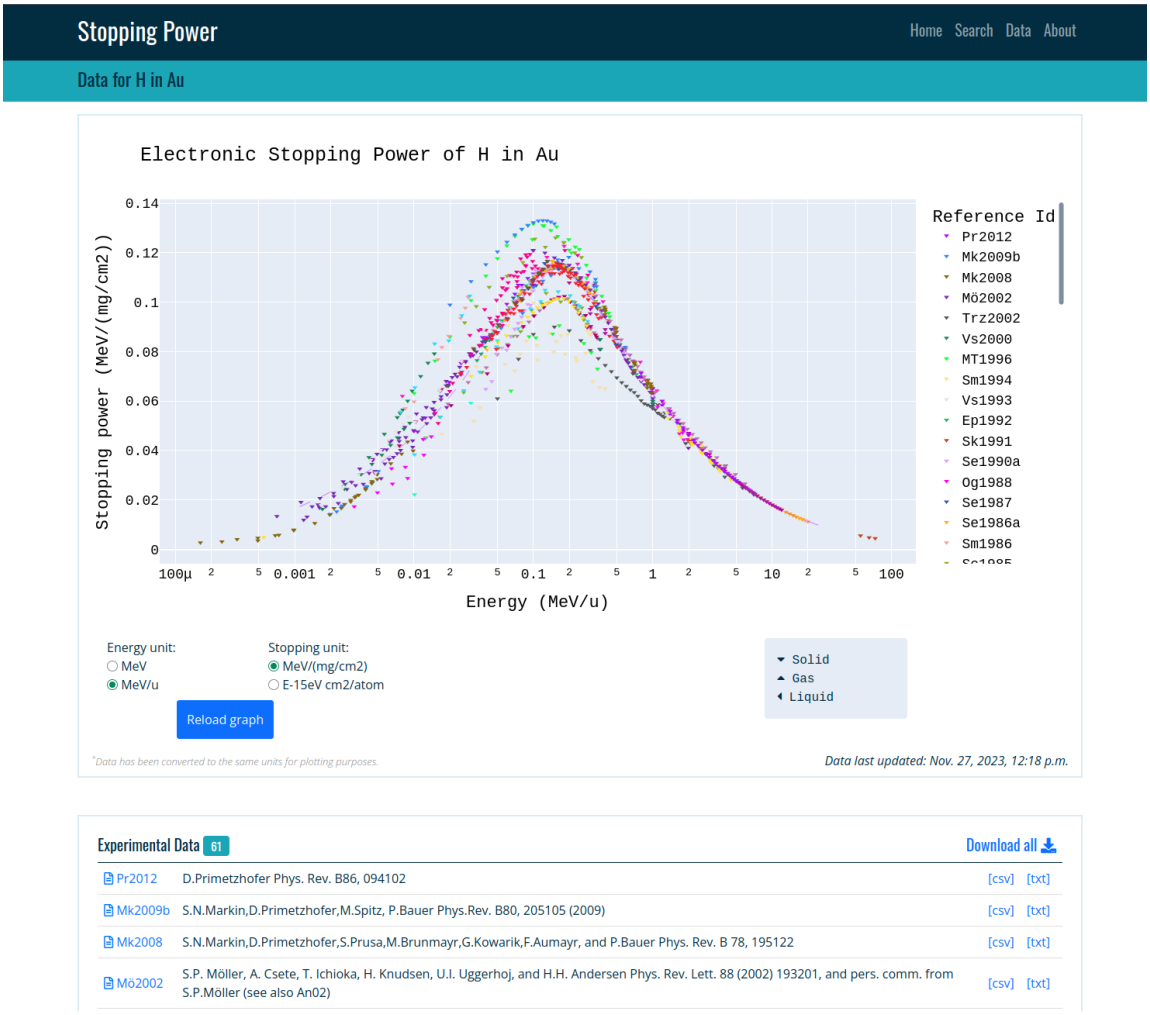



Figura 2.4: Ejemplo de búsqueda en la versión actual de la base de *Stopping Power* (interactiva).

## 2.2. Análisis estadístico de la base de datos

La base de datos actualizada de la IAEA recopila resultados experimentales de poder de frenado (SP) publicados entre 1928 y 2025, conteniendo registros de colisiones que involucran 283 blancos y 49 proyectiles distintos. El conjunto incluye un total de 1491 pares proyectil–blanco con mediciones disponibles. En este capítulo presentamos un análisis estadístico, que tiene como objetivo caracterizar la composición y calidad de los datos experimentales, identificar posibles sesgos en la distribución de estos valores, y evaluar la completitud de la información disponible. Esta caracterización es fundamental para entender las limitaciones y fortalezas de nuestro conjunto de datos, así como para orientar el desarrollo y validación de modelos predictivos.

### 2.2.1. Distribución de datos por características del proyectil

Para comprender los sesgos inherentes en los datos experimentales disponibles, comenzamos analizando la distribución de información según las características del proyectil incidente. Las Figuras 2.5 y 2.6 muestran la distribución de datos por proyectil incidente. En la primer figura se puede apreciar la cantidad de blancos diferentes, en la segunda, el número de mediciones recopiladas, y en ambas, en función de la masa del proyectil. Ambas figuras revelan una fuerte concentración de datos experimentales en proyectiles de baja masa atómica. Esta distribución refleja tanto la facilidad experimental para generar y acelerar iones ligeros, como el interés histórico en proyectiles como H y He, fundamentales para aplicaciones en física nuclear y medicina.

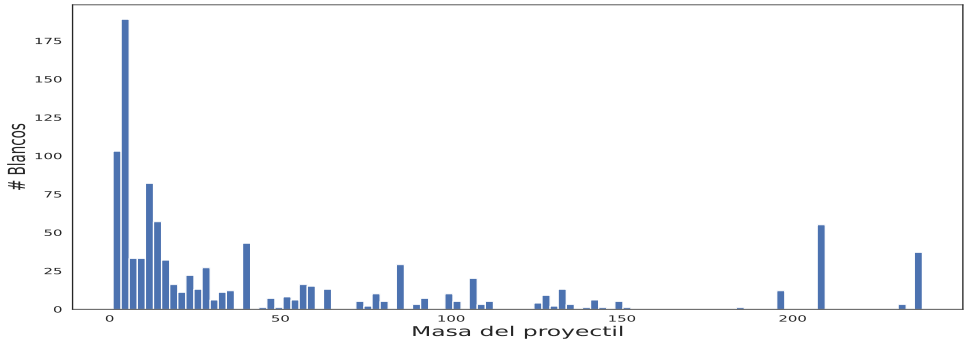



Figura 2.5: Cantidad de blancos distintos en función de la masa del proyectil.

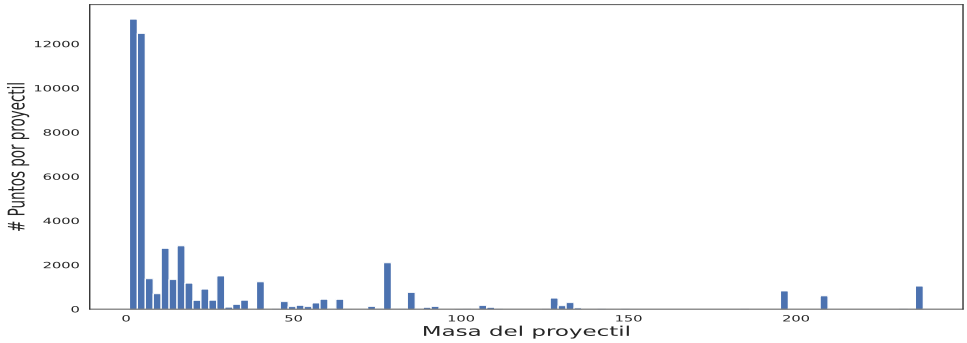



Figura 2.6: Cantidad de datos experimentales disponibles en función de la masa del proyectil.

### 2.2.2. Características de los blancos experimentales

El análisis de la composición de blancos revela patrones complementarios que influyen en la generalización de modelos predictivos. La Figura 2.7 muestra la cantidad de valores experimentales reportados en función del número de elementos del blanco. Dicha figura muestra que más del 60 % de los datos corresponden a blancos monoelementales. Esta predominancia justifica el enfoque en sistemas simples adoptado en trabajos previos, incluyendo nuestra publicación anterior [27] y el estudio de Guo *et al.* [28]. Sin embargo, también indica una oportunidad para extender los modelos hacia sistemas más complejos.

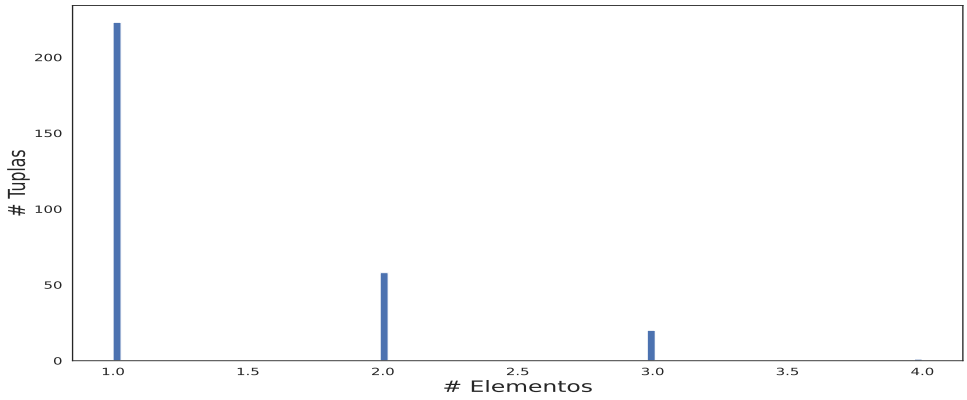



Figura 2.7: Distribución de datos según el número de elementos en el blanco.

En contraste, la composición de los blancos explorados en los distintos experimentos presentan una distribución significativamente más diversa, como se observa en la Figura 2.8, donde se muestra el número de datos experimentales en función de la masa del blanco. Esta diversidad refleja distintos intereses científicos y tecnológicos, generalmente vinculados a las aplicaciones específicas que requieren datos precisos de poder de frenado.

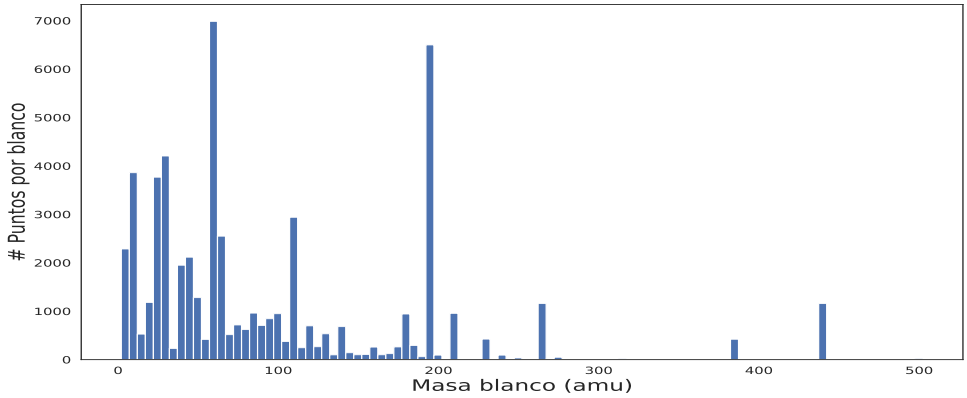
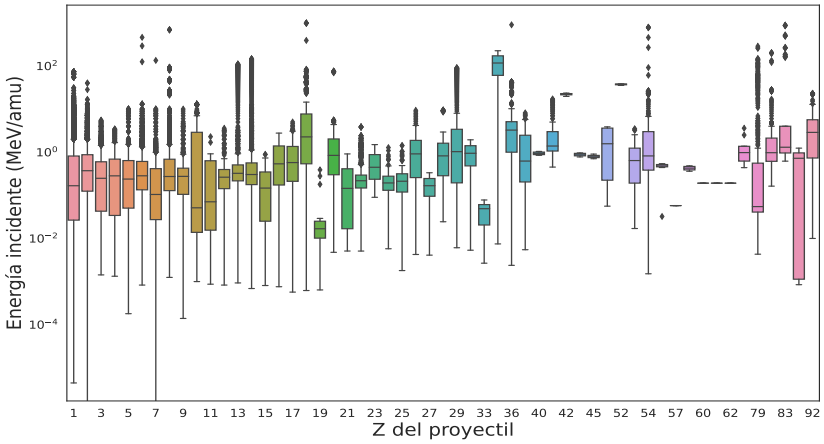


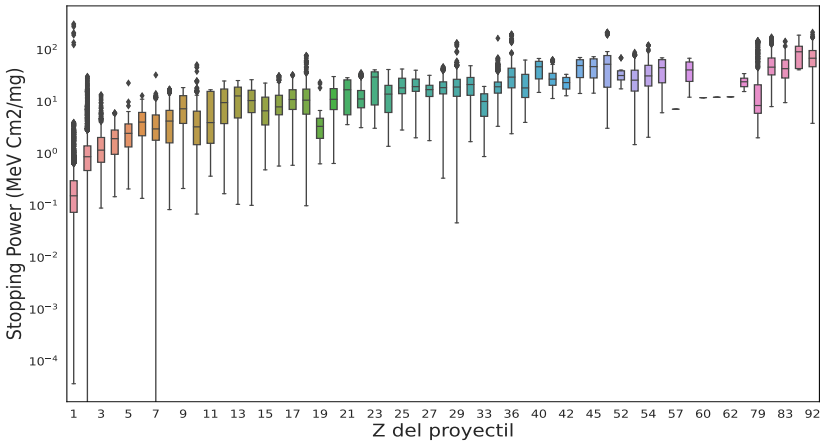

Figura 2.8: Cantidad de datos experimentales por masa del blanco.

### 2.2.3. Rangos de medición explorados

Realizamos un análisis de los rangos de energías y del poder de frenado (SP) que han sido más comúnmente explorados en los experimentos. Este tipo de análisis ofrece una forma rápida y eficaz de identificar posibles errores en los datos de la base, en particular aquellos relacionados con cambios o inconsistencias en las unidades. En las Figuras 2.9 se presentan diagramas de cajas correspondientes a los rangos de energía incidente (a) y poder de frenado (b), ambos normalizados y representados en función del número atómico $Z$ del proyectil. En ambas figuras, la línea central de la caja indica la mediana; el borde inferior, el primer cuartil; y el borde superior, el tercer cuartil. Los "bigotes"se extienden hasta los valores mínimo y máximo, siempre que se encuentren dentro de 1.5 veces la distancia intercuartílica respecto a los cuartiles primero y tercero, respectivamente. Los valores que se encuentran fuera de ese rango se representan como puntos individuales y corresponden a casos extremos. En cuanto a los rangos de energía, se observa que los experimentos realizados a energías más bajas suelen involucrar proyectiles con bajo número atómico $Z$. Asimismo, al considerar la escala logarítmica del eje vertical, puede advertirse que la media de los valores de SP tiende a seguir una función creciente con respecto al número atómico del proyectil.




(a)




(b)

Figura 2.9: Diagramas de cajas para los rangos de (a) energía incidente y (b) poder de frenado, normalizados, en función del número atómico $Z$ del proyectil.

### 2.2.4. Clasificación de sesgos en la distribución energética

Para caracterizar cómo se distribuyen los puntos experimentales a lo largo de la curva respecto de la posición del valor máximo, utilizamos los cuantiles de energía.

Un cuantil es un valor que divide un conjunto de datos ordenados en proporciones específicas. Por ejemplo, el cuantil 10 % ($q_{0,1}$) representa el valor de energía por debajo del cual se encuentra el 10 % de todas las mediciones, mientras que el cuantil 90 % ($q_{0,9}$) corresponde al valor por debajo del cual se encuentra el 90 % de los datos. Definimos el carácter de una curva comparando la posición de su máximo con la de estos cuantiles. Como ejemplo del uso de esta clasificación, se muestra en la Figura 2.10, una curva de valores de SP en función de la energía incidente. En ella, las líneas verdes verticales representan las posiciones de los cuantiles $q_{0,1}$ y $q_{0,9}$, mientras que la línea azul indica la energía $E_{\max}$ correspondiente al máximo de la curva de frenamiento. Esta curva ilustra un caso típico de "muestreo mayoritario a la derecha", en la cual la mayoría de los datos experimentales se encuentran a energías mayores que la del pico (es decir, se concentran a la derecha del máximo). Matemáticamente, esto se define como:

$$\text{Mayoritario a la derecha}: \qquad E_{\max} - q_{0,1} < 0$$

Un caso de "muestreo mayoritario a la izquierda" ocurre cuando la mayoría de las mediciones se concentran a energías menores que la del pico, es decir, en la cola izquierda de la curva. En este caso:

$$\text{Mayoritario a la izquierda}: \qquad E_{\max} - q_{0,9} > 0$$

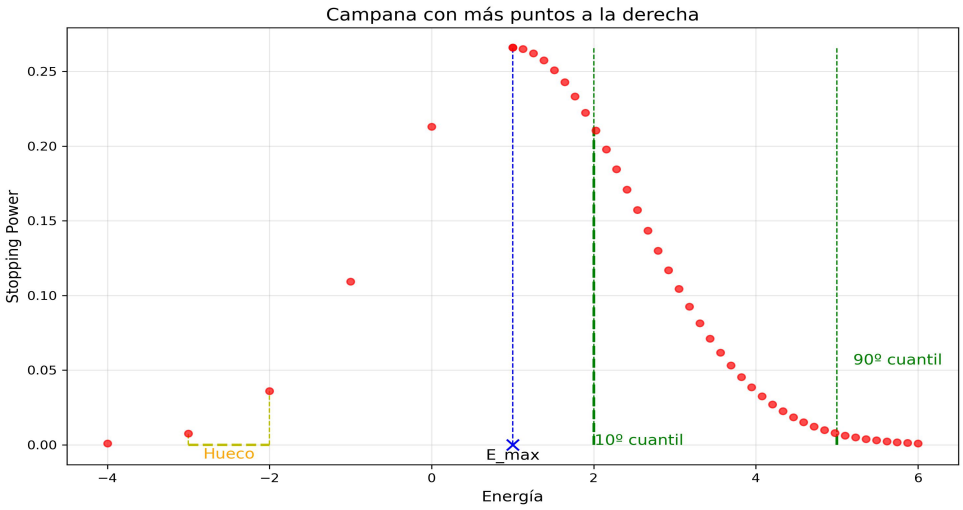



Figura 2.10: Ejemplo de curva típica con muestreo mayoritario a la derecha. Puntos rojos: valores experimentales. $E_{\max}$: Energía en la cual se produce el máximo en el SP. Líneas verdes: cuantiles $q_{0,1}$ y $q_{0,9}$.

La detección de sesgos sistemáticos proporciona información relevante para el entrenamiento de redes neuronales y sus posteriores predicciones. A partir de esta clasificación, identificamos 317 curvas con muestreo mayoritario hacia la izquierda y 305 hacia la derecha. En nuestro análisis observamos que ciertos proyectiles presentan sesgos sistemáticos en sus distribuciones. En particular, hay varios elementos cuyas mediciones se realizaron exclusivamente a bajas energías, lo que refleja tanto limitaciones experimentales como intereses de investigación específicos. Además, es posible que influya el hecho de que, a altas energías, los valores experimentales suelen ajustarse mediante la ley de Bethe. Definimos las fracciones de curvas con muestreo mayoritario a la derecha $f_d(Z)$ o izquierda $f_i(Z)$ como

$$f_d(Z) \equiv \frac{N_d(Z)}{N(Z)} \quad ; \quad f_i(Z) \equiv \frac{N_i(Z)}{N(Z)} \,,$$

donde $N(Z)$ denota la cantidad total de curvas con mediciones del elemento $Z$, mientras que $N_d(Z)$ ($N_i(Z)$) indica la cantidad de estas curvas que presentan un muestreo mayoritario hacia la derecha (izquierda). La Tabla 2.1 muestra las fracciones

correspondientes encontradas en la base de datos.

| Sesgo izquierda | | Sesgo derecha | |
|---|---|---|---|
| Proyectil | $f_i(Z)$ | Proyectil | $f_d(Z)$ |
| Zr | 1.00 | H | 0.58 |
| Sc | 1.00 | He | 0.55 |
| As | 1.00 | Ar | 0.32 |
| Co | 1.00 | Kr | 0.25 |
| K | 1.00 | O | 0.23 |
| W | 1.00 | Fe | 0.23 |
| I | 0.90 | C | 0.22 |
| P | 0.88 | Si | 0.22 |
| Br | 0.78 | Cu | 0.18 |
| Mn | 0.71 | Na | 0.17 |
| Cr | 0.70 | | |
| Pb | 0.61 | | |
| Be | 0.61 | | |
| Nb | 0.57 | | |
| B | 0.54 | | |

Tabla 2.1: Fracción de curvas con muestreo mayoritario a la izquierda $f_i(Z)$ y a la derecha $f_d(Z)$ para distintos proyectiles.

### 2.2.5. Evaluación de completitud: análisis de huecos en los datos

Una limitación crítica para el entrenamiento de modelos predictivos es la presencia de regiones con datos faltantes (*gaps*) en las curvas de poder de frenado. Para identificarlas, desarrollamos métricas específicas que permitan caracterizar estas lagunas y evaluar su impacto potencial. Definimos, para cada curva, la métrica de hueco relativo $G_R$:

$$G_R \equiv \frac{\max[E_{i+1} - E_i]}{\text{Rango}(E)},$$

que mide el tamaño de la discontinuidad máxima entre puntos consecutivos, en relación con el rango total de energías. En la Figura 2.10, esto correspondería a medir la distancia horizontal entre los puntos consecutivos más separados (por ejemplo, entre los puntos rojos más distantes) y dividirla por el rango total de energías representado en el eje x. Aplicando esta métrica a curvas con más de 10 puntos (umbral elegido para garantizar significancia estadística), encontramos que 216 curvas presentan al menos un hueco superior a 0.3 del rango total, y 84 curvas presentan huecos superiores a 0.5, lo cual indica una carencia significativa de información en dichas mediciones.

Dado que las curvas abarcan varias décadas logarítmicas, definimos además una métrica hueco local $G_L$:

$$G_L \equiv \frac{E_{i+1} - E_i}{E_i} \,,$$

la cual pondera la importancia relativa de un salto en función de su posición energética. A diferencia del hueco relativo, esta métrica refleja que un mismo salto absoluto tiene mayor impacto cuando ocurre a energías bajas que cuando ocurre a energías altas. La Figura 2.11 muestra la distribución de los valores máximos de $G_L$ registrados en las curvas de la base de datos experimentales. Se identificaron 20 curvas con máximo de $G_L > 10^2$ y 6 curvas con $G_L > 10^3$, lo que representa discontinuidades extremas. Estas curvas constituyen tanto un desafío significativo para el aprendizaje automático como una oportunidad para evaluar la capacidad de extrapolación de los modelos.

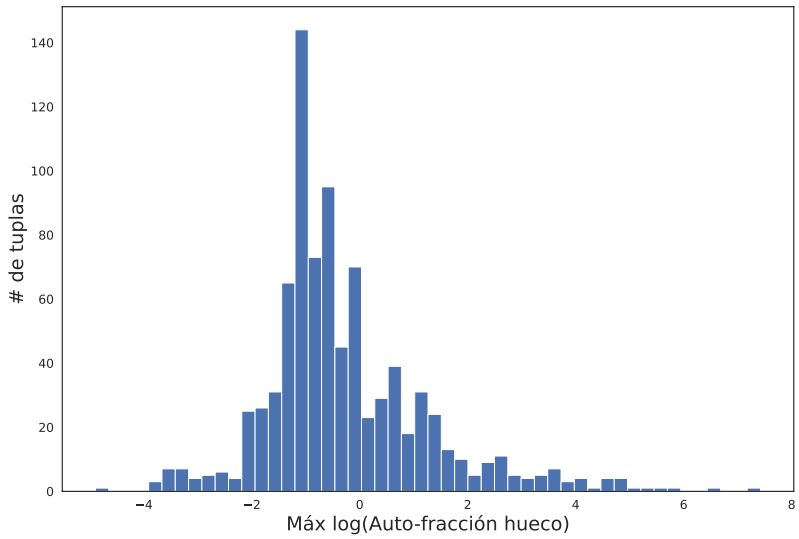



Figura 2.11: Distribución de los valores máximos de huecos locales $G_L$ en las curvas experimentales.

## 2.3. Evolución temporal de los datos

El análisis temporal de los datos proporciona perspectivas valiosas sobre la evolución del interés científico y las capacidades experimentales a lo largo del tiempo. Realizamos una serie de estudios centrados en la composición de los blancos utilizados en los experimentos, con el objetivo de comprender cómo estos han cambiado en el tiempo. En particular, la Figura 2.12 muestra la evolución del número de átomos presentes en los blancos experimentales. Se observa que recién en las dos últimas décadas comenzaron a estudiarse sistemas complejos, lo cual posiblemente esté vinculado al desarrollo de nuevas técnicas radiológicas y aplicaciones médicas.

Por su parte, la Figura 2.13 presenta la diversidad de elementos en los blancos a lo largo del tiempo. Como puede apreciarse, hasta principios del siglo XXI predominaban los blancos mono-elementales. El creciente interés en materiales multi-elementales refleja el desarrollo de aplicaciones más sofisticadas y específicas.

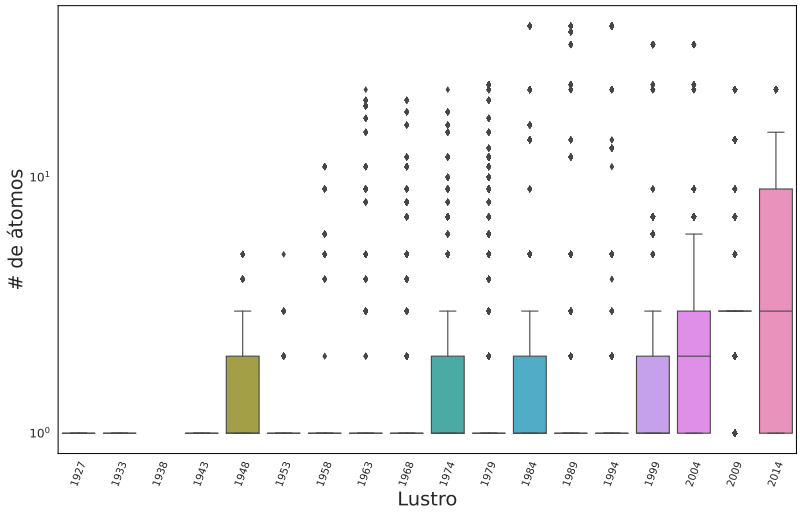



Figura 2.12: Evolución temporal del número de átomos por blanco estudiado.

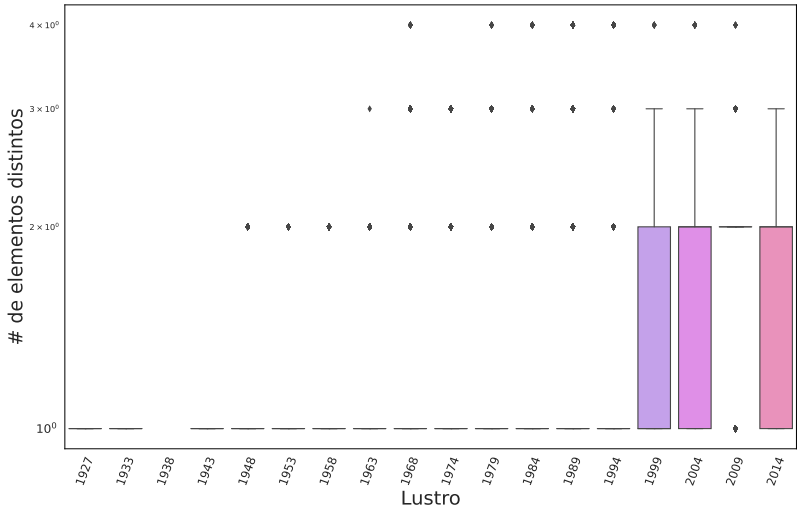



Figura 2.13: Evolución temporal de la diversidad de los blancos estudiados.

Finalmente, la Figura 2.14 muestra los patrones temporales específicos asociados a cada proyectil. Se observa que el interés por proyectiles ligeros ha disminuido relativamente con el tiempo. También se identifican picos de actividad asociados a elementos pesados ($Z$ = 82) durante la década de 2000, y a elementos intermedios

($Z = 9 - 13$) en el período 2018–2020, reflejando avances tecnológicos y el surgimiento de nuevas aplicaciones en distintos momentos históricos.

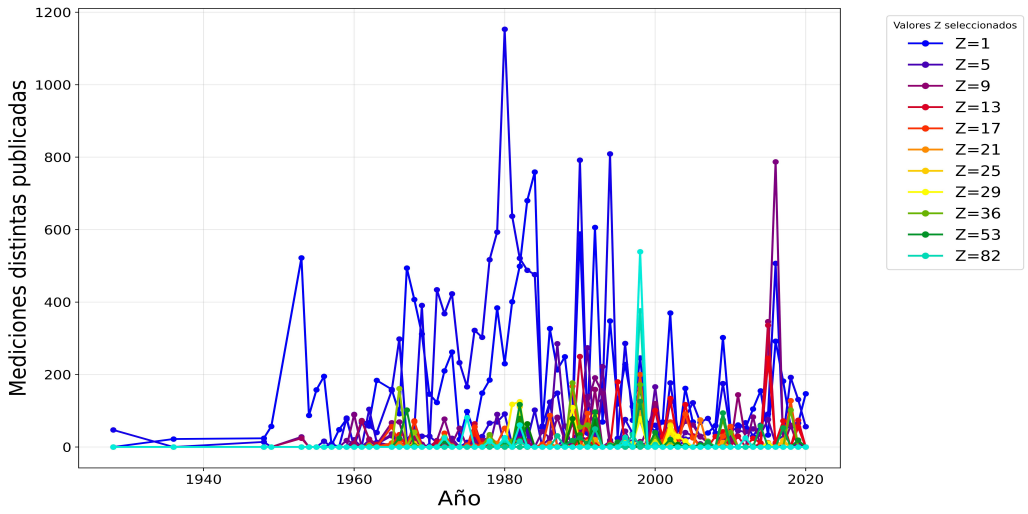



Figura 2.14: Evolución temporal del volumen de datos por proyectil.

## 2.4. Implicaciones para el modelado

Los distintos análisis realizados sobre la base de datos revelan varios aspectos críticos que deben considerarse en el desarrollo de modelos predictivos:

- **Sesgos en la distribución de datos**:
  La predominancia de proyectiles ligeros y blancos mono-elementales sugiere que los modelos alcanzarán mayor precisión en esas regiones del espacio de parámetros, mientras que la extrapolación a sistemas más complejos requerirá una validación cuidadosa.

- **Desafíos de completitud**:
  Los huecos identificados, en particular aquellos con auto-fracciones locales extremas, corresponden a regiones con escasa información experimental. Estas áreas destacan la necesidad de generar nuevos datos y representan un desafío considerable para la capacidad de extrapolación de los modelos.

- **Evolución temporal**:
  La creciente complejidad de los sistemas estudiados en años recientes pone de manifiesto la necesidad de desarrollar modelos capaces de manejar blancos multi-elementales y proyectiles pesados de manera robusta.

Estas características de la base de datos deben guiar tanto las estrategias de entrenamiento como los protocolos de validación en el desarrollo de nuestros modelos predictivos.

# Capítulo 3

# Remoción de Ruido

Dadas las diferencias en la toma de datos, ya sea a lo largo del tiempo por el surgimiento de nuevas técnicas experimentales, por la aplicación de metodologías distintas en diferentes rangos de energía, por limitaciones físicas de los dispositivos y métodos de medición, o bien por impurezas presentes en los blancos experimentales, los datos no se presentan como una línea continua. En cambio, se observan regiones densamente pobladas por puntos, curvas separadas que eventualmente se entrecruzan, y nubes de puntos aisladas.

Retirar el ruido de los datos sin disponer de un criterio robusto para determinar cuales son los resultados esperados verdaderos (*ground truth*), y sin introducir sesgos derivados de modelos preexistentes, como por ejemplo de SRIM, plantea un dilema metodológico. En las siguientes secciones, exploramos el uso de un codificador-decodificador automático para eliminación de ruido (DAE - *Denoising Autoencoder*) y un algoritmo de filtrado que desarrollamos para complementar al método de agrupamiento espacial basado en densidad para aplicaciones con ruido (DBSCAN).

## 3.1. Denoising Auto Encoder

El método de DAE fue propuesto por primera vez en [29]. La idea principal consiste en tomar los datos tal como se encuentran, agregarles ruido de manera controlada

y luego entrenar un modelo para que aprenda a recuperar los datos originales sin ruido. En nuestro caso, debido a la amplia diferencia de escalas entre los experimentos correspondientes a distintos pares proyectil-blanco, la manera más sencilla de introducir ruido es mediante permutaciones de los valores de SP. Concretamente, se seleccionan aleatoriamente pares de puntos `Energía-Frenamiento` $(E_i, S_i)$ y $(E_j, S_j)$ dentro de una misma curva, y se intercambian sus componentes de SP, generando así nuevos puntos $(E_i, S_j)$ y $(E_j, S_i)$. Este procedimiento se repite sobre distintos pares aleatorios a lo largo del conjunto de datos. Esta forma de introducir ruido permite mantener las escalas propias de cada curva de SP, que, como se analizó en la sección 2.2, varían ampliamente según el par proyectil-blanco considerado.

### 3.1.1. Arquitectura y selección de tamaño de la ventana

La arquitectura de nuestra red consiste en un embudo de capas convolucionales unidimensionales. En la primera parte del modelo, denominada *encoder*, se aplican capas de convolución 1D [30] junto con capas de *pooling*, que incrementan progresivamente la cantidad de canales mientras reducen el tamaño de la secuencia de entrada. En la segunda parte, denominada *decoder*, se emplean capas de convolución transpuesta 1D y capas de *upsampling*, que sucesivamente reducen la cantidad de canales y aumentan nuevamente el tamaño de la secuencia.

Este tipo de arquitectura requiere que los tensores de entrada y salida tengan el mismo tamaño. Dado que las redes se entrenan en lotes (*batches*) mediante el algoritmo de descenso del gradiente estocástico (SGD – *Stochastic Gradient Descent*) [31], un tensor típico de entrada tiene la forma `[batch size, features, window size]`, donde `features` representa las características de cada punto, en nuestro caso, energía y SP. A mayor tamaño de ventana, mayor es la cantidad de información que el DAE puede utilizar para revertir las permutaciones introducidas como ruido. Sin embargo, usar ventanas muy grandes obliga a descartar curvas de SP correspondientes a ciertos pares proyectil-blanco con pocas mediciones, o bien completarlas artificialmente con datos sintéticos. Como compromiso entre cobertura y capacidad de aprendizaje, se optó por un tamaño de ventana de 20 puntos. Esta elección implica que no es posible aplicar el modelo de remoción de ruido a aproximadamente

el 30 % de los pares proyectil-blanco, que tienen menos de 20 mediciones.

Para determinar la arquitectura óptima y los hiperparámetros de entrenamiento, se utilizó el paquete OPTUNA [32], el cual implementa algoritmos de búsqueda automática. En particular, se empleó el método *Tree-structured Parzen Estimator* (TPE) [33]. La métrica utilizada para la optimización fue el *Mean Absolute Percentage Error* (MAPE), definido como:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i^{\text{pred}} - y_i^{\text{true}}|}{y_i^{\text{true}}}\,. \tag{3.1}$$

Una de las ventajas de OPTUNA es que permite modificar dinámicamente la región de búsqueda en función de los valores ya explorados, lo cual resulta fundamental al ajustar hiperparámetros sensibles como la tasa de aprendizaje (*learning rate*) o el factor de decaimiento (*weight decay*). La arquitectura final obtenida se describe en la Tabla 3.1. Las capas 1 a 3 forman el *encoder*, mientras que las capas 4 a 6 conforman el *decoder*. Las capas 2 y 3 son convolucionales, y las capas 4 y 5 son convolucionales transpuestas. La capa 6 realiza una operación de *flattening* seguida por una capa lineal. También se detallan los tamaños de los *kernels* convolucionales y la cantidad de canales presentes en cada etapa del procesamiento.

| **Capas** | **Operación** | **Tamaño del Kernel** | **Canales** | **Activación** |
|---|---|---|---|---|
| 0 | Entrada | - | 2 | - |
| 1 | BatchNorm1d | - | 20 | - |
| 2 | Conv1d | 7 | $2 \to 5$ | ReLU |
| 3 | Conv1d | 8 | $5 \to 6$ | ReLU |
| 4 | ConvTranspose1d | 8 | $6 \to 5$ | ReLU |
| 5 | ConvTranspose1d | 7 | $5 \to 2$ | - |
| 6 | Flatten + Linear | - | $2 \times L_{out} \to 20$ | - |
| | Salida | - | 20 | - |

Tabla 3.1: Descripción de la arquitectura de nuestro DAE convolucional en 1D.

### 3.1.2. Evaluación de DAEs

Como se observa en la Figura 3.1, la convergencia obtenida en términos del error cuadrático medio (*Mean Squared Error* – MSE) durante el entrenamiento es muy buena. No se evidencian signos de sobreajuste (*overfitting*) y el MAPE en el conjunto de validación alcanza un valor de 0.1. Sin embargo, este resultado no es suficiente por sí solo para demostrar que la red está efectivamente eliminando el ruido de manera robusta.

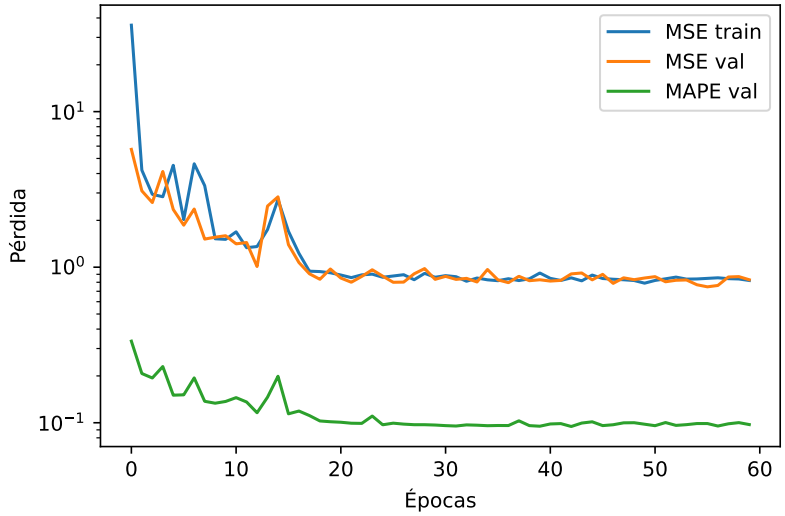



Figura 3.1: Curvas de entrenamiento para el DAE, obtenidos con los mejores hiperparámetros sugeridos por OPTUNA. El ruido consiste en permutaciones de SP con una probabilidad de 0.01.

Una vez entrenado el modelo convolucional, surge la necesidad de establecer criterios que permitan evaluar la calidad de sus resultados. Dado que se espera que el método no sólo reduzca el ruido artificial introducido durante el entrenamiento, sino también el ruido inherente al conjunto de datos original, no es posible determinar directamente si efectivamente está eliminando este último. Esto se debe a la ausencia de datos de referencia generales que permitan comparar las salidas del modelo con un valor verdadero conocido. Sin embargo, contamos con dos heurísticas que permiten evaluar su eficacia. La primera, que se analiza en este capítulo, consiste simplemente

en verificar si la diferencia entre las predicciones es menor que el ruido artificialmente introducido. La segunda, que se abordará más adelante, se basa en analizar el desempeño del algoritmo de predicción entrenado sobre los datos previamente filtrados por nuestro DAE. Para estimar el nivel total de ruido introducido mediante permutaciones, se calcula el valor del MAPE sobre $10^4$ permutaciones aleatorias del conjunto completo de datos. En cada permutación, cada punto tiene una probabilidad de 0.01 de ser intercambiado. La distribución de los valores de MAPE resultantes se muestra en la Figura 3.2. De dicha figura se deduce que un MAPE de 0.1, obtenido al aplicar el DAE sobre los datos artificialmente distorsionados, corresponde a un caso con distorsión relativamente baja. El valor de *p-value* asociado es menor a 0.01, lo cual indica que la probabilidad de que el proceso aleatorio de permutación genere una configuración más "limpia"que la recuperada por el DAE es inferior al 1 %.

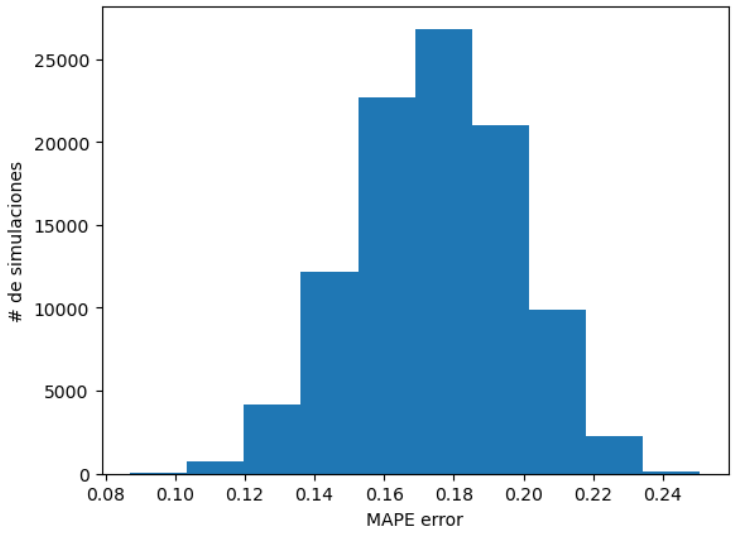



Figura 3.2: Distribución de MAPE obtenida al aplicar ruido mediante $10^4$ permutaciones de SP con una probabilidad de 0.01.

A pesar de estos resultados aparentemente favorables, un MAPE de 0.1 sigue siendo superior al valor cercano a 0.05 reportado en [34], que consideramos como referencia a superar. Más aún, una inspección cualitativa de los datos filtrados sugiere la necesidad de contar con un sistema de limpieza que permita tener un mayor control

sobre qué tipo de publicaciones se prioriza preservar. En particular, se requiere un método que facilite incorporar un sesgo explícito hacia mediciones más modernas.

## 3.2. Agrupamiento y Limpieza con DBSCAN

Como alternativa a los métodos de limpieza de ruido desarrollados en la sección anterior, diseñamos un algoritmo alternativo, que tiene como objetivo remover por completo referencias experimentales sospechosas. Para ello, empleamos el algoritmo DBSCAN [35, 36], el cual permite identificar agrupamientos con geometría no globular, lo que representa una ventaja significativa a la hora de distinguir entre mediciones que siguen el contorno característico de las curvas de SP y aquellas que se desvían de dicho comportamiento.

El algoritmo DBSCAN, cuyo nombre proviene de *Density-Based Spatial Clustering of Applications with Noise*, es un método de agrupamiento espacial basado en la densidad de los puntos, y tiene capacidad para identificar y excluir ruido. La idea principal detrás de esta técnica radica en lo siguiente: dado un conjunto de puntos en un espacio métrico, el algoritmo agrupa aquellos que se encuentran densamente conectados (es decir, que poseen muchos vecinos cercanos), y clasifica como valores atípicos aquellos puntos ubicados en regiones de baja densidad, donde los vecinos más próximos se encuentran demasiado alejados. DBSCAN requiere dos parámetros de entrada: el radio del vecindario $\epsilon$, y el número mínimo de puntos dentro de dicho vecindario $N_{\min}$, alcanzables (dentro de una distancia $\epsilon$), que deban ser tomados en cuenta para que una región sea considerada densa. Los puntos que cumplen estas condiciones forman un grupo (*cluster*), mientras que los que no son alcanzables desde ningún otro punto se etiquetan como valores atípicos o puntos de ruido (*outliers*).

Entre las principales ventajas de DBSCAN se destacan que no requiere especificar previamente la cantidad de grupos presentes en los datos (a diferencia de métodos como *K-means*), puede identificar agrupamientos de forma arbitraria posee una noción explícita de ruido, es robusto frente a valores atípicos y es poco sensible al orden de los datos. Estas características lo han convertido en un algoritmo ampliamente utilizado y uno de los más citados en la literatura de aprendizaje no supervisado. En

particular en este trabajo resaltamos la capacidad de DBSCAN de encontrar clusters no globulares que pueden seguir las formas de curvas arbitrarias frente a *K-means* que solo encuentra clusters globulares.

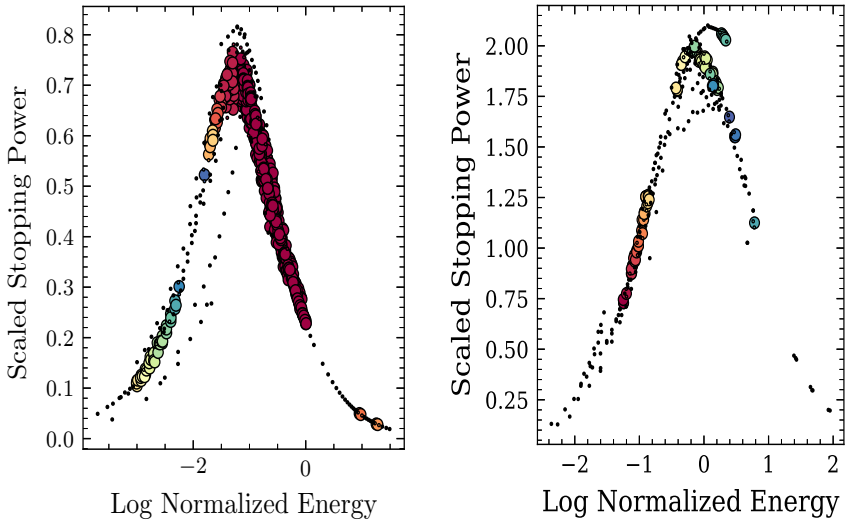



Figura 3.3: Secciones eficaces del poder de frenado escaladas para H en Si (izquierda) y O en Au (derecha). Los diferentes colores representan grupos detectados por el algoritmo DBSCAN con $\epsilon$ = 0,025 y $N_{\text{min}}$ = 3. Los puntos atípicos se indican con puntos pequeños.

Como ejemplo de las particiones generadas por DBSCAN, en la Figura 3.3 se muestran las secciones eficaces del poder de frenado para H en Si (izquierda) y O en Au (derecha). En todos los casos, las secciones eficaces se han escalado para mantener ambos ejes de tamaño similar.

Encontramos conveniente establecer los parámetros de entrada como $\epsilon$ = 0,025 y $N_{\text{min}}$ = 3. En el caso de H en Si, el algoritmo no supervisado identificó 19 grupos (representados con distintos colores) y clasificó 139 puntos como atípicos. En el caso de O en Au, se detectaron 26 grupos y 181 valores atípicos. Recordemos que los puntos atípicos son aquellos que no llegan a tener $N_{\text{min}}$ puntos vecinos dentro de un círculo de radio $\epsilon$.

En aplicaciones estándar del algoritmo DBSCAN para la limpieza de datos, los pun-

tos marcados como atípicos suelen eliminarse directamente del conjunto de datos. Sin embargo, en nuestro contexto esto implicaría una pérdida excesiva de información. Por ello, desarrollamos e implementamos un algoritmo de filtrado secuencial en tres etapas, en el cual, una vez identificados los grupos y los puntos atípicos, el algoritmo decide qué observaciones conservar y cuáles eliminar.

### 3.2.1. Algoritmo de filtrado: descripción general

El algoritmo de filtrado se aplica individualmente para cada sistema colisional. Primero, se escalean los datos y se ejecuta el algoritmo DBSCAN, mediante el cual se determinan diferentes grupos y valores atípicos. Luego, se hace un análisis de estos datos, considerando ciertas características concernientes a su publicación. En particular, se consideran el año de publicación, el rango de energía que cubre dicha presentación, y se analiza si este rango tiene solapamiento con otras publicaciones. Estos criterios son relevantes para determinar la eliminación definitiva de puntos de ruido, como detallaremos a continuación. Por ejemplo, en la Figura 3.3, algunos puntos que DBSCAN señala como atípicos se encuentran en regiones donde no hay otras mediciones disponibles. En esos casos, a pesar de su clasificación inicial, se decide conservar dichos puntos como válidos. En una segunda etapa, se inspeccionan puntos que están superpuestos en energía pero aislados en sección eficaz. Si bien, al estar aislados, deberían considerarse como valores atípicos, implementamos un criterio *ad hoc* que explicaremos a continuación para determinar si efectivamente son descartados como ruido, o considerados como valores válidos de entrada. Por último, se examina la composición de cada grupo. Se evalúa si los datos experimentales del grupo provienen de una única fuente o de múltiples publicaciones, lo que permite ajustar el criterio de limpieza con mayor precisión.

A continuación, describiremos nuestro algoritmo con más detalle, incluyendo el pseudocódigo correspondiente. Algunos ejemplos del filtrado se presentan en la Figura 3.4.

### 3.2.2. Algoritmo de filtrado: detalles

Para cada sistema colisional, evaluamos inicialmente cada publicación según su solapamiento energético con otras referencias. Para ello, determinamos el número de valores experimentales $n_i$ reportados en la publicación $P_i$, distribuidos en un rango de energía $\Delta E_i$. Si los datos de la publicación cubren una región energética sin superposición con otras mediciones, dichos datos se conservan en la base de datos de entrada. En caso contrario, la conservación o descarte de una publicación (siempre considerando el mismo sistema colisional) se determina considerando un criterio temporal. En primer lugar, se identifican aquellos valores experimentales que:

1. se encuentran dentro del rango de energía $\Delta E_i$, y
2. fueron publicados con posterioridad a $P_i$.

Denotamos por $\Delta E_i^p$ el rango total de energía cubierto por estas publicaciones más recientes. A partir de ello, evaluamos qué fracción del rango de energía de $P_i$ también está cubierta por publicaciones posteriores. Si esta fracción no es suficientemente grande, implica que no hay demasiados datos más modernos que cubran la región, por lo que mantenemos como válidos los valores publicados en la publicación $P_i$. Es decir, escogemos un valor $\sigma_\Delta$ tal que:

$$\text{Si} \quad \frac{\Delta E_i^p}{\Delta E_i} \leq \sigma_\Delta \quad \rightarrow \quad \textbf{CONSERVAR} \quad \text{la publicación } P_i\,. \tag{3.2}$$

El umbral de superposición $\sigma_\Delta$ se fija típicamente en 0.6, excepto para los casos en los que el número de referencias que abordan el sistema colisional específico es muy pequeño; en tales casos, $\sigma_\Delta \approx 1$. La condición (3.2) garantiza que se conserven publicaciones que aportan datos en regiones energéticas previamente inexploradas, incluso si reportan valores que han sido clasificados atípicos por DBSCAN. Podemos ver un ejemplo en el caso de colisiones O en Au. Según se observa en el panel derecho de la Figura 3.3, los puntos pertenecientes a la región de bajas energías han sido clasificados como valores atípicos. Sin embargo, como se observa en el panel derecho inferior de la Figura 3.4, estos valores se mantienen y se incluyen en la entrada de la

red neuronal, sin importar si pertenecen a un grupo o si son clasificados como valores atípicos.

A continuación, se procesan las publicaciones restantes que, al no satisfacer el criterio (3.2), comparten regiones superpuestas en el dominio energético. Para un sistema ion-objetivo dado, esta segunda etapa de filtrado consiste en descartar publicaciones que reportan resultados "aislados en SP", es decir, mediciones realizadas en el mismo rango de energía que en otras publicaciones pero con valores de secciones eficaces significativamente diferentes. Para ello, se calcula el número de valores atípicos $N^i_{\text{outl}}$ detectados en $P_i$ y se evalúa:

$$\text{Si} \quad \frac{N^i_{\text{outl}}}{n_i} > \sigma_{\text{out}} \quad \rightarrow \quad \textbf{ELIMINAR} \quad \text{la publicación } P_i \tag{3.3}$$

donde el parámetro $\sigma_{\text{out}} \approx 0{,}45$, por defecto, excepto cuando el número de publicaciones es escaso; en tales casos, $\sigma_{\text{out}} \approx 1$. Esta condición permite descartar publicaciones en las cuales una proporción sustancial de sus resultados es considerada sospechosa. Si se supera cierto umbral, es preferible dirimir las diferencias conservando las publicaciones más recientes, que se suponen más fiables.

Finalmente, se aplica una evaluación semejante a este último criterio, pero en lugar de examinar valores atípicos, intentaremos analizar si dentro del conjunto de grupos determinados por DBSCAN existen grupos atípicos. Para ello, en cada publicación $P_i$, se selecciona al grupo más numeroso $C_i$. Este grupo contiene $t^i$ puntos en total, de los cuales $l^i$ provienen exclusivamente de $P_i$. Se evalúa entonces:

$$\text{Si} \quad \frac{l^i}{t^i} > \sigma_{\text{clu}} \quad \rightarrow \quad \textbf{ELIMINAR} \quad \text{la publicación } P_i \tag{3.4}$$

Este criterio, con $\sigma_{\text{clu}} \approx 0{,}45$ en general (y cercano a 1 en conjuntos escasos), busca evitar que una publicación forme un grupo compuesto exclusivamente por resultados provenientes de ese mismo trabajo, sin la presencia de datos de múltiples fuentes que respalden sus observaciones, lo que podría indicar una falta de consistencia con el resto de la evidencia disponible.

**Procedure** Clusters and Outliers
**Input:**
  $\epsilon$ (radio del vecindario)
  $N_{\text{min}}$ (número de puntos alcanzables)
Ejecutar DBSCAN para seleccionar diferentes clusters y outliers
**Output:**
  $N_{\text{clust}}$ (número de clusters)
  $N_{\text{outl}}$ (número de puntos de ruido)
**End Procedure**

**Procedure** Drop Publications
**for** cada publicación $P_i$ en la lista de publicaciones **do**
  Contar $n_i$ (número de resultados experimentales en $P_i$)
  Definir $\Delta E_i$ (rango de energía cubierto en $P_i$)
  **for** cada publicación $P_j^n$ **más nueva que** $P_i$ **do**
    Definir $\Delta E_{ji}^n \equiv \Delta E_j^n \bigcap \Delta E_i$
  **end for**
  Definir $\Delta E^{\text{new}} \equiv \bigcup_j \Delta E_{ji}^n$
  **if** $\frac{\Delta E^{\text{new}}}{\Delta E_i} \leq \sigma_\Delta$ **then**
    **PRESERVAR** publicación $P_i$ y **BREAK**
  **end if**
  Contar $N_{\text{outl}}^i$ (número de outliers en $P_i$)
  **if** $\frac{N_{\text{outl}}^i}{n_i} > \sigma_{\text{outl}}$ **then**
    **BORRAR** publicación $P_i$
  **end if**
  Identificar $C^i$ (cluster más grande en $P_i$)
  Contar $l^i$ (número de valores $n_i$ desde $P_i \in C^i$)
  Contar $t^i$ (total de resultados $\sum n_j \in C^i$)
  **if** $\frac{l^i}{t^i} > \sigma_{\text{clu}}$ **then**
    **BORRAR** publicación $P_i$
  **end if**
**end for**
**End Procedure**

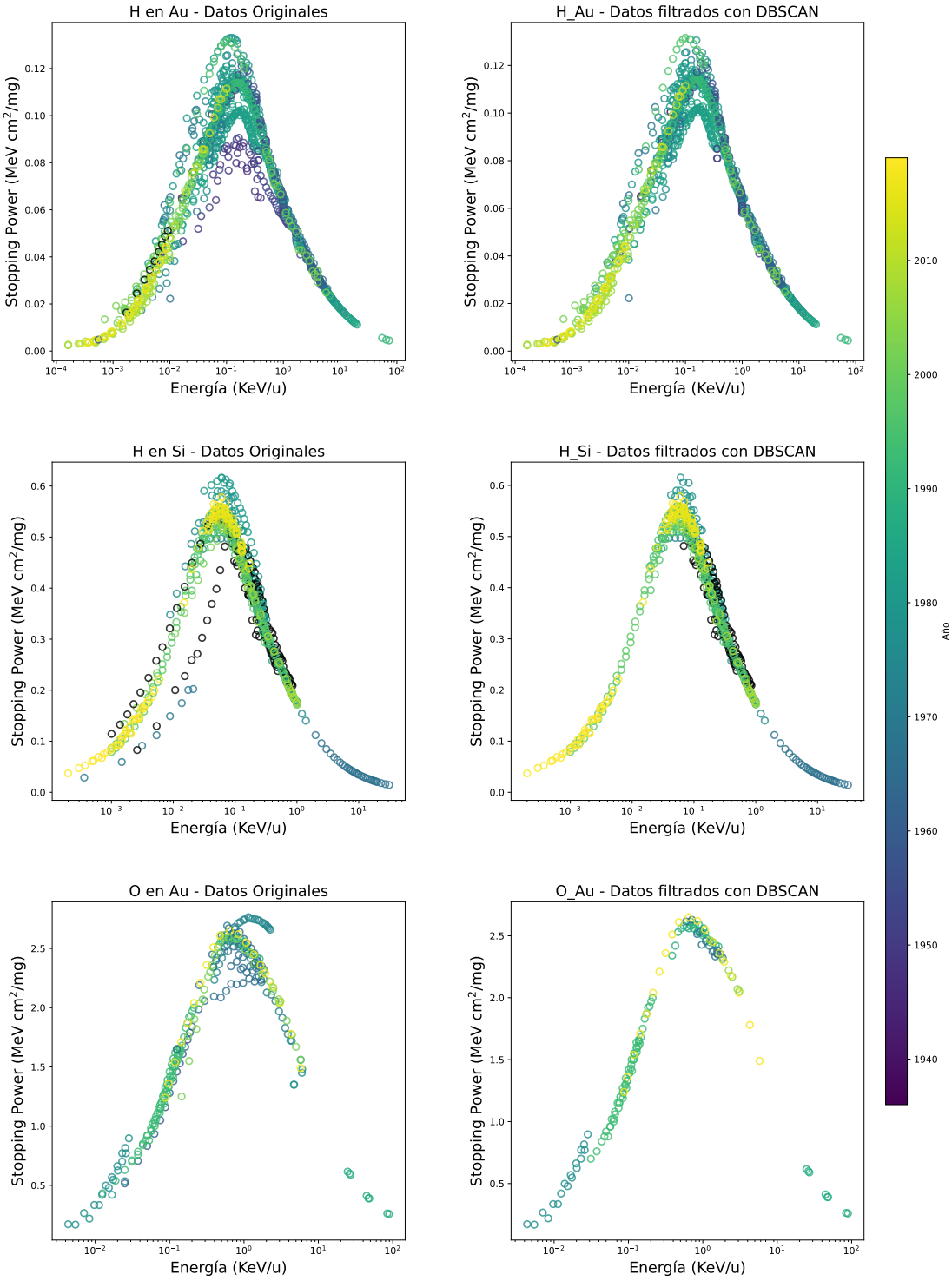



Figura 3.4: Secciones eficaces del poder de frenado escaladas para H en Au (arriba), H en Si (centro) y O en Au (abajo). En los paneles de la izquierda se muestran los valores experimentales originales. A la derecha, los datos filtrados con nuestro algoritmo. Los colores indican el año de publicación.

### 3.2.3. Evaluación del método heurístico

Para evaluar la eficacia del método de filtrado basado en DBSCAN, se desarrolló un protocolo de validación utilizando publicaciones sintéticas con niveles de ruido controlado. El procedimiento consiste en recorrer cada sistema colisional proyectil-blanco, seleccionando aleatoriamente una publicación entre todas las que contienen datos para dicho sistema. A los datos de esta publicación seleccionada se les introduce ruido artificial. La perturbación se implementa multiplicando todos los valores de poder de frenado (SP) por un factor $1 \pm \alpha$, donde $\alpha$ representa el nivel de ruido relativo introducido. Posteriormente, se aplican los dos métodos de filtrado: el método heurístico basado en DBSCAN y el método de autoencoder DAE. En el caso de DBSCAN, el algoritmo identifica de forma natural las publicaciones problemáticas y las elimina de la base de datos. Para evaluar el método DAE, se verifica si los datos de la publicación alterada permanecen alterados (en cuyo caso el método no logra detectarla como publicación ruidosa) o si son reemplazados por los valores originales, indicando así que el modelo ha reconocido la distorsión.

La Tabla 3.2 presenta los resultados obtenidos para distintos valores de $\alpha$, comparando el desempeño del método heurístico propuesto con el método DAE.

| **Nivel de ruido $\alpha$** | DBSCAN **(% detectados)** | **DAE (% detectados)** |
|---|---|---|
| 0.10 | 30 | 10 |
| 0.05 | 12 | 1 |
| 0.03 | 5 | 0 |

Tabla 3.2: Porcentaje de publicaciones ruidosas correctamente detectadas por los métodos de filtrado basados en DBSCAN y DAE, para distintos niveles de ruido $\alpha$.

Los resultados obtenidos muestran que la heurística basada en DBSCAN supera en forma consistente al método DAE en todos los niveles de ruido evaluados. Es particularmente destacable que, para $\alpha = 0{,}1$, el método heurístico logra identificar correctamente el 30 % de las publicaciones problemáticas, triplicando la eficacia del enfoque DAE. Si bien las tasas de detección alcanzadas por ambos métodos no son elevadas en términos absolutos, estos resultados deben interpretarse a la luz de la complejidad inherente del problema. La detección de publicaciones anómalas en bases

de datos experimentales constituye un desafío considerable, ya que implica distinguir entre variaciones legítimas, derivadas de diferentes condiciones experimentales, y errores sistemáticos genuinos. En este contexto, la mejora sustancial proporcionada por la heurística basada en DBSCAN resulta especialmente relevante en la región de mayor interés práctico: perturbaciones del orden de $\alpha$ = 0,1, que corresponden a desviaciones del 10 % en los valores de poder de frenado. Cabe señalar, además, que el protocolo de evaluación propuesto proporciona una estimación conservadora del desempeño real del método, ya que el ruido se introduce de forma uniforme sobre publicaciones completas. En escenarios reales, los errores pueden presentarse de forma más localizada o seguir patrones específicos que podrían facilitar su detección. Esto sugiere que la eficacia del método en aplicaciones prácticas podría ser incluso superior a la observada en estas pruebas controladas.

## 3.3. Conclusión

Se evaluaron distintos enfoques para la detección y eliminación de ruido en los datos experimentales. Se ha validado que el método basado en DBSCAN genera curvas verosímiles, caracterizadas por una mayor concentración de datos provenientes de mediciones modernas y de alta calidad. La prueba definitiva de cualquier método de limpieza de datos reside en su capacidad para producir modelos que se desempeñen adecuadamente en conjuntos de validación no utilizados durante el entrenamiento. Este análisis será desarrollado en el capítulo siguiente.

# Capítulo 4

# Modelos de Stopping Power

Una vez completado el estudio y la limpieza de la base de datos, emprenderemos la tarea de desarrollar un modelo de red neuronal para las predicciones de SP en colisiones con sistemas mono-elementales, es decir, para compuestos de un solo elemento químico. En este sentido, primero haremos un breve repaso de los trabajos previos, y luego nos dedicaremos a establecer la arquitectura de red y sistema de aprendizaje que utilizamos. Finalmente, observaremos los resultados del mejor modelo en distintos cortes temporales y compararemos contra el mejor modelo pseudo-empírico. Gran parte de lo que se explica en este capítulo ha sido publicado en [27].

## 4.1. Predicciones del poder de frenado

El cálculo del poder de frenado de un ion al interactuar con la materia implica modelar numerosos procesos y parámetros, resultando un desafío sumamente complejo. A pesar de los distintos modelos teóricos desarrollados para describir los datos experimentales, ninguno ha logrado reproducirlos con la suficiente precisión en una amplia región energética. Generalmente se utilizan métodos semiempíricos que ajustan los resultados conocidos, para proporcionar valores recomendados en simulaciones y aplicaciones multipropósito. Además, los datos experimentales presentan discrepancias significativas, lo que obliga a introducir métodos auxiliares para cla-

sificar y seleccionar información confiable. Considerando la abundante cantidad de resultados experimentales recopilados en la base de datos de la IAEA, esta constituye un recurso ideal para emplear métodos de aprendizaje automático que puedan ayudar a manejar la complejidad de los procesos involucrados e incluso descubrir relaciones ocultas entre los datos.

Mientras desarrollábamos este proyecto, se publicaron dos artículos independientes, que adoptaron un enfoque similar al nuestro. Parfitt y Jackman [37] entrenaron un modelo de algoritmo de regresión de bosques aleatorios utilizando miles de mediciones compiladas por Paul (con datos hasta 2015), que como mencionamos en 1.2.2 ha sido la antecesora de la base de IAEA. Mediante validación cruzada $k$-fold y frente a varias métricas de error, demostraron que su modelo se ajustaba bien a los datos de entrenamiento y ofrecía predicciones de bajo error sobre datos no vistos. Por su parte, Guo *et al.* [34] emplearon la misma base de datos para entrenar una red neuronal profunda, logrando reproducciones precisas de los resultados experimentales en objetivos mono-elementales.

A diferencia del enfoque adoptado por Guo *et al.* [28], en este trabajo se optó por no extender artificialmente la base de datos mediante la incorporación de valores teóricos o semi-empíricos. Nuestro objetivo principal es reproducir únicamente los datos experimentales considerados válidos, e introducir conocimiento físico exclusivamente a través de la estructura de la red neuronal y de la selección adecuada de características de entrada. Esta decisión responde tanto a criterios metodológicos como a consideraciones sobre la fidelidad del aprendizaje. Desde un punto de vista estadístico, la base de datos empleada contiene un total de 36 000 puntos experimentales. En contraste, en [28] se añadieron aproximadamente 32 000 puntos generados mediante una fórmula semiempírica, lo que implica que una proporción significativa de los datos de entrenamiento proviene de señales sintéticas. En tales casos, existe el riesgo de que el modelo aprenda a replicar la función generadora artificial más que los comportamientos reales observados en los datos experimentales, comprometiendo la generalización del modelo a escenarios físicamente relevantes.

Es importante destacar que ninguno de los otros modelos ofrece código de acceso libre que permita reproducir sus resultados en distintos conjuntos de prueba y vali-

dación, lo que hace imposible una comparación *ceteris paribus* con ellos. En nuestro caso, hemos publicado el código de inferencia en https://github.com/ale-mendez/ESPNN, el cual permite validar fácilmente el modelo contra otros enfoques utilizando nuevas mediciones experimentales.

## 4.2. Modelo de Red Neuronal

Durante el diseño de nuestra red neuronal (*Neural Network* NN), se realizó un análisis exhaustivo para establecer los mejores hiperparámetros a utilizar. Este estudio incluye no solo el diseño arquitectónico de la NN sino también la definición de la función de pérdida a minimizar, los parámetros involucrados en el proceso de minimización, la tasa de aprendizaje, la degradación de pesos y otros. Encontramos que el mejor diseño es el que se describe en la Tabla 4.1. Este consiste en un conjunto de cinco capas ocultas lineales y una función de activación Leaky-ReLU para cada capa, y el optimizador de estimación de momento adaptativo (ADAM) como algoritmo de minimización. El número de características de entrada (`num_features`) será discutido más adelante, y el número de salida (`num_targets`) es uno, correspondiendo a la predicción de SP para los valores de entrada dados. Se impusieron tasas de *dropout* de 0.5 para cada capa oculta. Los otros parámetros para el modelo de entrenamiento son la tasa de aprendizaje $\alpha$ = 0,001, el tamaño del lote (*batch* size) $b$ = 64, y la degradación de peso (*weight decay*) $\lambda = 10^{-10}$. Empleamos 300 épocas con una parada temprana de 50. Se utilizó la técnica de reparametrización Weight Normalization [38] para aumentar la velocidad de convergencia.

Vale la pena mencionar el siguiente aspecto técnico, que detallaremos más adelante. El poder de frenado debe disminuir monótonicamente hacia energías bajas. Sin embargo, para algunos sistemas colisionales, hemos encontrado que la red produce divergencias para bajos valores de energías. Intentamos agregar valores ficticios mínimos de poder de frenado en energías muy bajas, pero como era de esperar, el entrenamiento se volvió extremadamente inestable debido a su componente MAPE. Logramos un comportamiento correcto en el modelo final mediante un simple truco: eliminando los parámetros de sesgo de la primera capa lineal.

| Capa | Operación | Entrada | Salida | Activación | Dropout |
|---|---|---|---|---|---|
| | Entrada | `num_features` | | - | - |
| 1 | Linear + Weight Norm | `num_features` | 10 | Leaky-ReLU | 0.5 |
| 2 | Linear + Weight Norm | 10 | 24 | Leaky-ReLU | 0.5 |
| 3 | Linear + Weight Norm | 24 | 32 | Leaky-ReLU | 0.5 |
| 4 | Linear + Weight Norm | 32 | 24 | Leaky-ReLU | 0.5 |
| 5 | Linear + Weight Norm | 24 | 10 | Leaky-ReLU | 0.5 |
| | Salida: Linear | 10 | `num_targets` | - | |

Tabla 4.1: Arquitectura de la Red Neuronal para la predicción del poder de frenado.

Un estudio adicional se centró en la asignación de los datos para el proceso de aprendizaje. De los datos limpios (28 000 valores), separamos el 5 % (1400 resultados) para el `conjunto de prueba`. Estos datos no participaron en el proceso de entrenamiento. El 95 % restante de los datos, llamado conjunto de aprendizaje, fue dividido en cinco pliegues. Cuatro pliegues comprendían el `conjunto de entrenamiento`, mientras que el quinto pliegue constituía el `conjunto de validación`. Usando este diseño, entrenamos la red neuronal hasta asegurar la convergencia exitosa de la función de pérdida. Luego, los pesos de cada capa se almacenan para estimar las secciones eficaces del poder de frenado en una etapa posterior. Este procedimiento se lleva a cabo iterativamente rotando los pliegues de entrenamiento-validación. Como resultado, obtenemos cinco conjuntos de pesos y, por lo tanto, cinco estimaciones diferentes para cualquier entrada. Las predicciones finales se obtienen promediando los resultados obtenidos de los cinco pesos de NN, proporcionando también una estimación aproximada de la incertidumbre presente en el cálculo.

Como explicaron Parfitt y Jackman [37], se pueden utilizar diferentes métricas de error para cuantificar el rendimiento del modelo en términos de los valores predichos $y_{\text{pred}}$ frente a los valores experimentales verdaderos $y_{\text{true}}$. Para comparaciones con otros métodos, utilizamos la función de pérdida del error porcentual absoluto medio, o MAPE, dada por

$$\text{MAPE} \equiv \frac{100}{n} \sum \left| \frac{y_i^{\text{true}} - y_i^{\text{pred}}}{y_i^{\text{true}}} \right| . \tag{4.1}$$

Sin embargo, para el proceso de entrenamiento, la función de pérdida se define mejor añadiendo una pequeña contribución de otra métrica al MAPE. Esto evita que el entrenamiento presente inestabilidades, que pueden aparecer en valores pequeños de poder de frenado. Definimos la función de pérdida como una combinación lineal de dos métricas distintas,

$$L \equiv \text{MAPE} + \frac{\text{MSE}}{\beta} \,, \tag{4.2}$$

donde MSE es el error cuadrático medio definido por

$$\text{MSE} = \frac{\sum_i \left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2}{n} \,, \tag{4.3}$$

y $\beta$ actúa como un factor de escala entre diferentes unidades y se establece en 100. La Fig. 4.1 muestra la curva de aprendizaje de cuatro conjuntos de entrenamiento. En ella podemos observar que el procedimiento de aprendizaje alcanza un mínimo global para la función de pérdida. Este comportamiento demuestra que nuestro modelo NN no sufre de problemas de sesgo. Además, el hecho de que el conjunto de validación siga la misma tendencia que la curva de aprendizaje indica que los datos no están sobreajustados, lo que significa que los parámetros de la red neuronal se han seleccionado correctamente, evitando problemas de varianza.

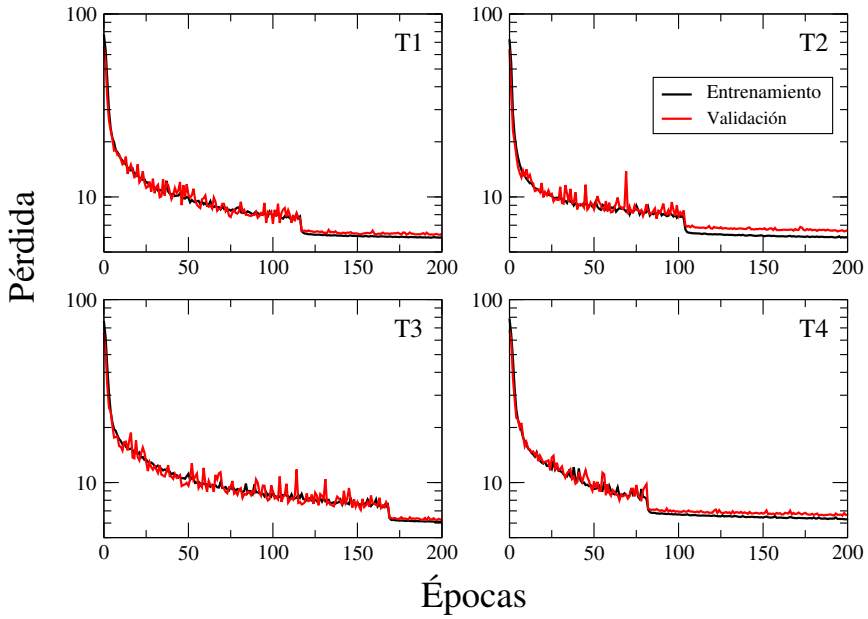



Figura 4.1: Curva de aprendizaje para los conjuntos de entrenamiento y validación, para cuatro pliegues diferentes.

## 4.3. Selección de las características de entrada

Al diseñar el modelo de red neuronal, es fundamental realizar un estudio detallado de las características que deben incorporarse como datos de entrada. Si bien el objetivo de la red consiste en predecir la sección eficaz de frenamiento en un blanco en función de la energía incidente del proyectil, no se conoce a priori cuáles son las propiedades relevantes del blanco y del proyectil que deben considerarse como variables de entrada. Además, es necesario tener en cuenta tanto las escalas de los valores de entrada como su comportamiento físico. A continuación, se analizan estos aspectos de forma individual.

En el problema abordado en este trabajo, las energías incidentes cubren varios órdenes de magnitud, lo que hace imprescindible aplicar una transformación de escala. Se determinó que la forma más efectiva de hacerlo es mediante la aplicación del

logaritmo a la energía incidente, en lugar de utilizar su valor original. Esta transformación facilita el entrenamiento del modelo y mejora su capacidad de generalización.

Desde el punto de vista físico, se sabe que el poder de frenado debe tender a cero asintóticamente, tanto para energías muy altas como para energías muy bajas. Estos comportamientos límites son fundamentales y deben ser incorporados en el modelo de alguna manera. Dado que se espera que la red neuronal infiera estos comportamientos exclusivamente a partir de los datos experimentales, se decidió no incluir puntos artificiales con valores de energía extremadamente bajos o altos y sección eficaz nula. Asimismo, tampoco se consideró adecuado incorporar una penalización física explícita durante el entrenamiento, ya que las escalas de energía varían significativamente entre diferentes problemas colisionales. Se optó, entonces, por una solución simple y efectiva: anular el sesgo (*bias*) de la primera capa de la red. Esta modificación fuerza indirectamente que, para un valor nulo de energía, la red prediga una sección eficaz también nula. Adicionalmente, se eliminó la función de activación en la última capa, lo que previene posibles divergencias en las predicciones.

Los efectos de estas modificaciones pueden observarse en la Figura 4.2, donde se compara el modelo final (arriba, derecha) con otros modelos que incluyen sesgo, utilizan una escala lineal para la energía o emplean una función de activación en la última capa.

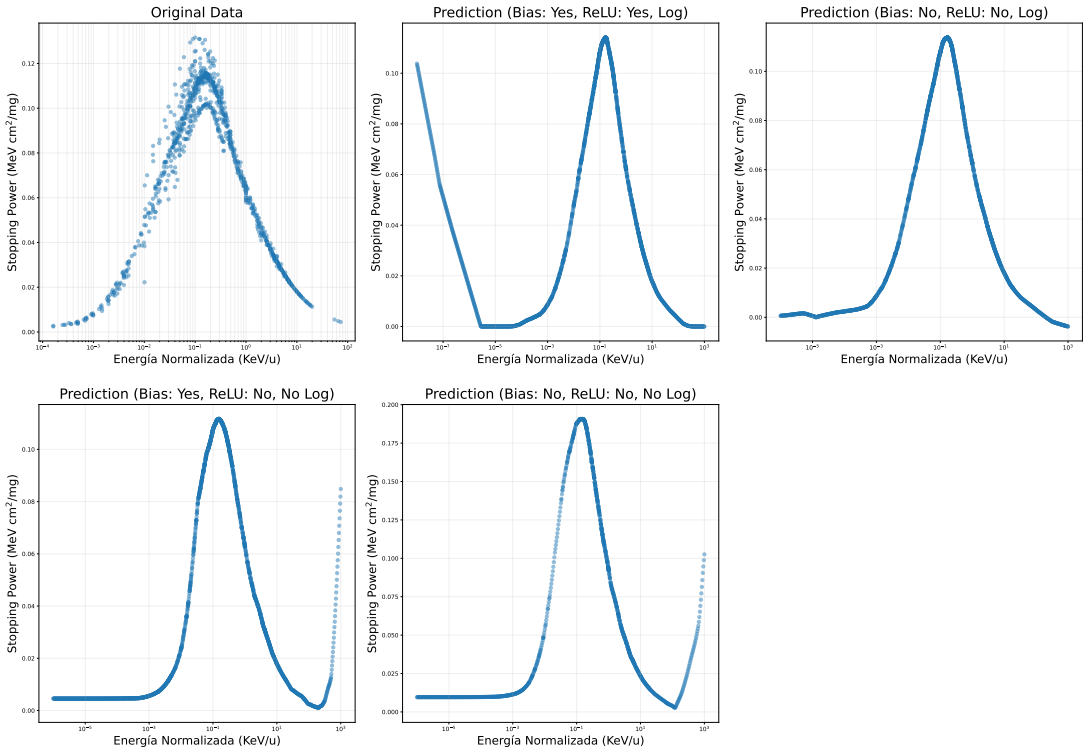



Figura 4.2: Curva de H en Au para distintos hiperparámetros del sistema.

La red neuronal fue entrenada con diversos conjuntos de características de entrada, y se concluyó que es posible determinar con buena precisión el poder de frenado electrónico de los átomos utilizando únicamente cinco variables:

1-2) Los números atómicos $Z_p$ y $Z_t$, del proyectil y del blanco, respectivamente,

3-4) Las masas atómicas $m_p$ y $m_t$, del ion incidente y del blanco,

5) La energía incidente del proyectil $E_i$ (por unidad de masa, expresada en `keV/amu`).

Con este conjunto básico de características se obtuvieron predicciones con un error porcentual absoluto medio (MAPE) cercano al 6 % en el conjunto de validación. Se observó que reemplazar la energía incidente por su logaritmo mejora significativamente el rendimiento del modelo, por lo que esta transformación fue adoptada en

todas las combinaciones posteriores. Posteriormente, se exploraron otras combinaciones de características, incluyendo la energía de ionización del blanco, lo que condujo a una mejora adicional en el valor de MAPE. Motivados por este resultado, se evaluó la incorporación de otras variables, como la segunda energía de ionización del blanco, la energía de ionización del proyectil y las electronegatividades de ambos elementos. Sin embargo, la inclusión de todas estas características resultó en un desempeño inferior, con errores más elevados en todos los casos. La influencia de las diferentes combinaciones de características sobre el desempeño del modelo se resume en la Tabla 4.2.

| **Características** | **MAPE (%)** |
|---|---|
| Por defecto: $Z_p$, $m_p$, $Z_t$, $m_t$, $E_i$ | 5.76 |
| $E_i \rightarrow \log E_i$ | 5.47 |
| + Primera energía de ionización (blanco) | **5.07** |
| + Primera + segunda energía de ionización (blanco) | 14.90 |
| + Primera ionización (blanco) + primera ionización (proyectil) | 16.10 |
| + Primera ionización + electronegatividad (blanco) | 5.11 |
| + Primera ionización (blanco) + electronegatividad (proyectil) | 23.80 |

Tabla 4.2: Errores MAPE (%) evaluados en el conjunto de validación cruzada para diferentes combinaciones de características de entrada utilizadas en el entrenamiento del modelo. En negrita se indican las características seleccionadas para el modelo final.

## 4.4. Resultados

En esta sección se presentan y analizan los resultados obtenidos con el modelo ESPNN, descrito en la Sección 4.2. Las predicciones de las secciones eficaces del poder de frenado se obtienen mediante la propagación hacia adelante (*forward propagation*) de los datos de entrada a través de la red neuronal. Las Figuras 4.3, 4.4, 4.5 y 4.6 muestran algunos resultados representativos obtenidos con el modelo ESPNN. En los paneles de la izquierda se presentan los datos experimentales originales, coloreados según el año de publicación. Los paneles centrales muestran los datos filtrados por el algoritmo descrito en la Sección 3.2, que constituyen la entrada efectiva al modelo.

Finalmente, en los paneles de la derecha se ilustran las predicciones generadas por la red neuronal, junto con los datos de entrada, con el fin de facilitar la comparación visual. Para la mayoría de las combinaciones proyectil-blanco, el modelo reproduce de manera satisfactoria las tendencias generales observadas en los datos experimentales. Cabe destacar que, aunque las predicciones fueron calculadas punto por punto para cada valor de energía incidente, los resultados del modelo se presentan como curvas suaves y continuas, lo que evidencia su capacidad de generalización. En particular, la Figura 4.3 muestra secciones eficaces del poder de frenado para distintos sistemas colisionales con oro (Au) como blanco. En estos casos, las predicciones del modelo tienden a seguir la tendencia establecida por las mediciones más recientes (indicadas en colores oscuros). Por otro lado, la Figura 4.4, correspondiente a sistemas con hidrógeno (H) como proyectil, evidencia que el modelo es capaz de reproducir adecuadamente la forma de los datos estadísticamente más representativos. La suavidad de las curvas predichas por ESPNN puede interpretarse como una consecuencia de utilizar un conjunto reducido de características de entrada en el proceso de entrenamiento, lo cual contribuye a evitar el sobreajuste. Es notable que se obtengan predicciones suaves y coherentes a pesar de que los datos experimentales presenten dispersión significativa en ciertas regiones del dominio energético.

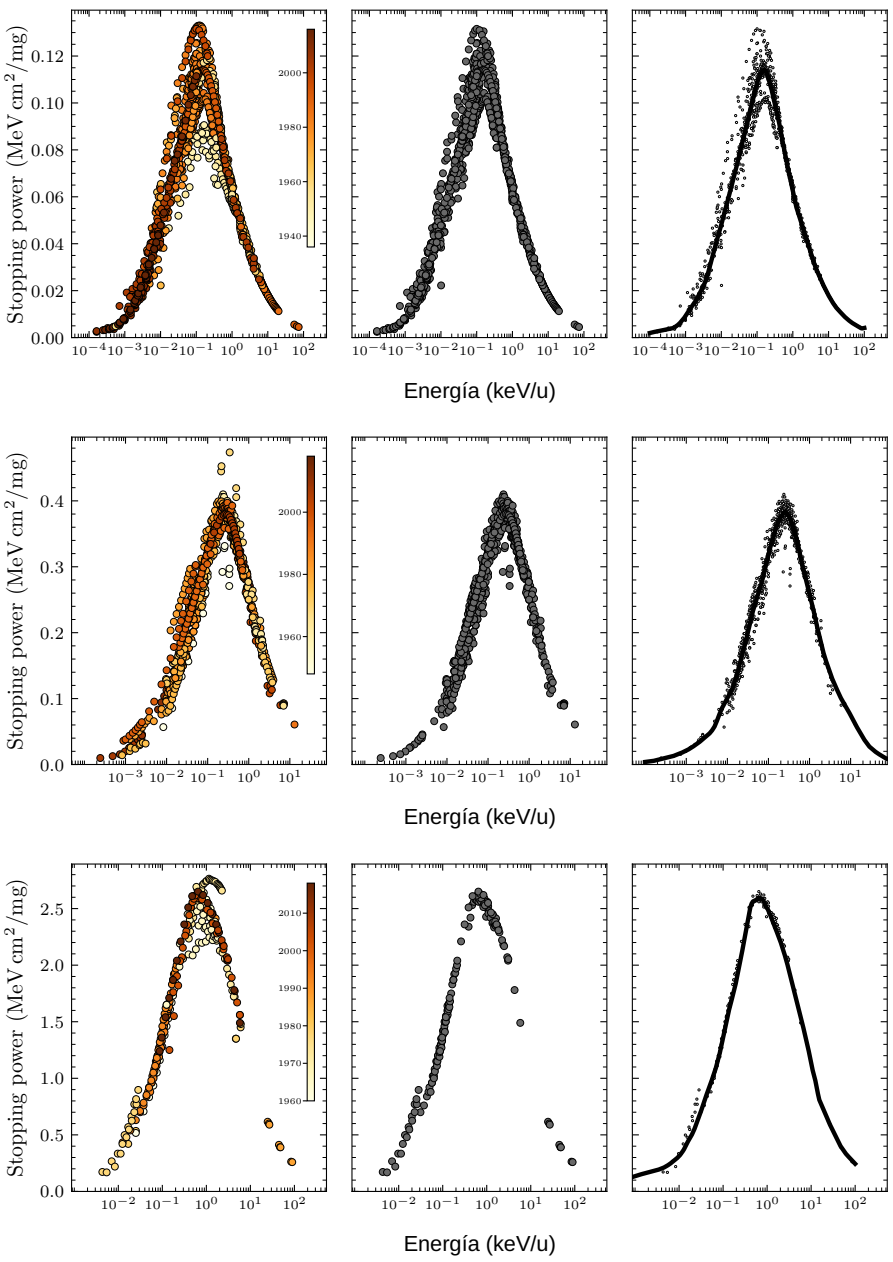



Figura 4.3: Izquierda: Resultados experimentales para secciones eficaces de poder de frenado de sistemas colisionales con Au como blancos. Los colores indican el año de publicación de los datos. Centro: Datos filtrados, resultantes del procedimiento de limpieza explicado en el texto. Derecha: Datos predichos por la red neuronal. Arriba: H en Au. Medio: He en Au. Abajo: O en Au.

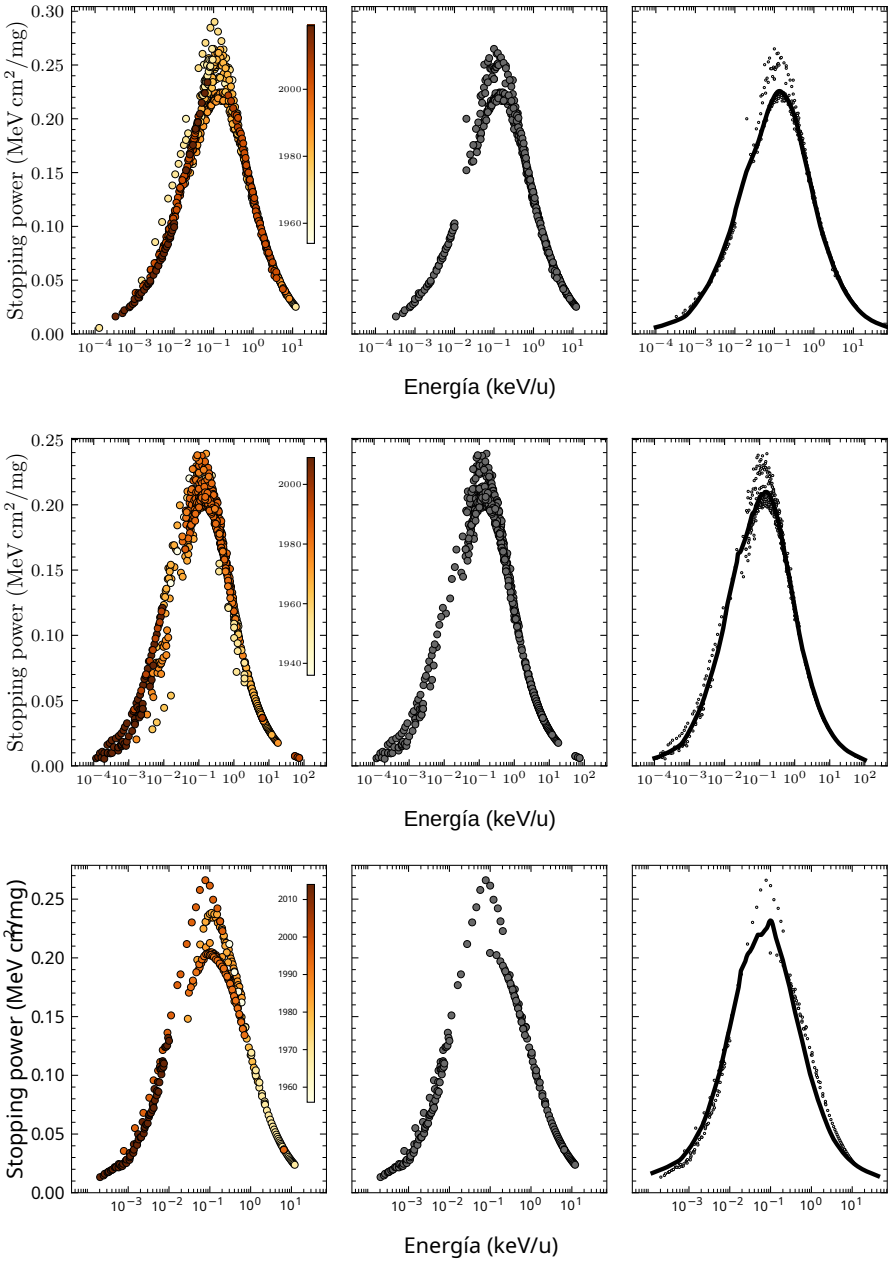



Figura 4.4: Izquierda: Resultados experimentales para secciones eficaces de poder de frenado, para proyectiles de H. Los colores indican el año de publicación de los datos. Centro: Datos filtrados. Derecha: Datos predichos por la red neuronal. Arriba: H en Ni. Medio: H en Cu. Abajo: H en Zn.

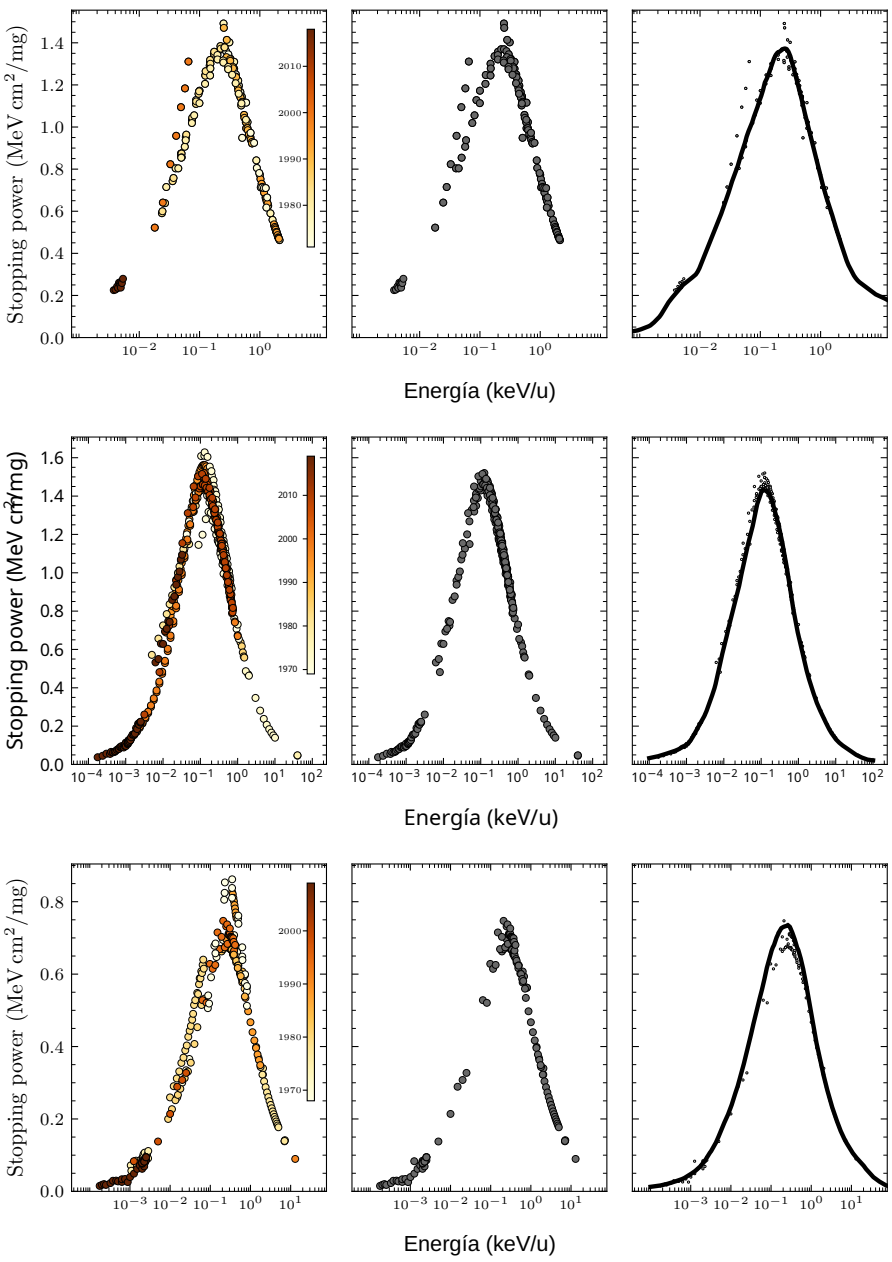



Figura 4.5: Izquierda: Resultados experimentales para secciones eficaces de poder de frenado, para proyectiles de He. Los colores indican el año de publicación de los datos. Centro: Datos filtrados. Derecha: Datos predichos por la red neuronal. Arriba: He en Ne. Medio: He en Si. Abajo: He en Cu.

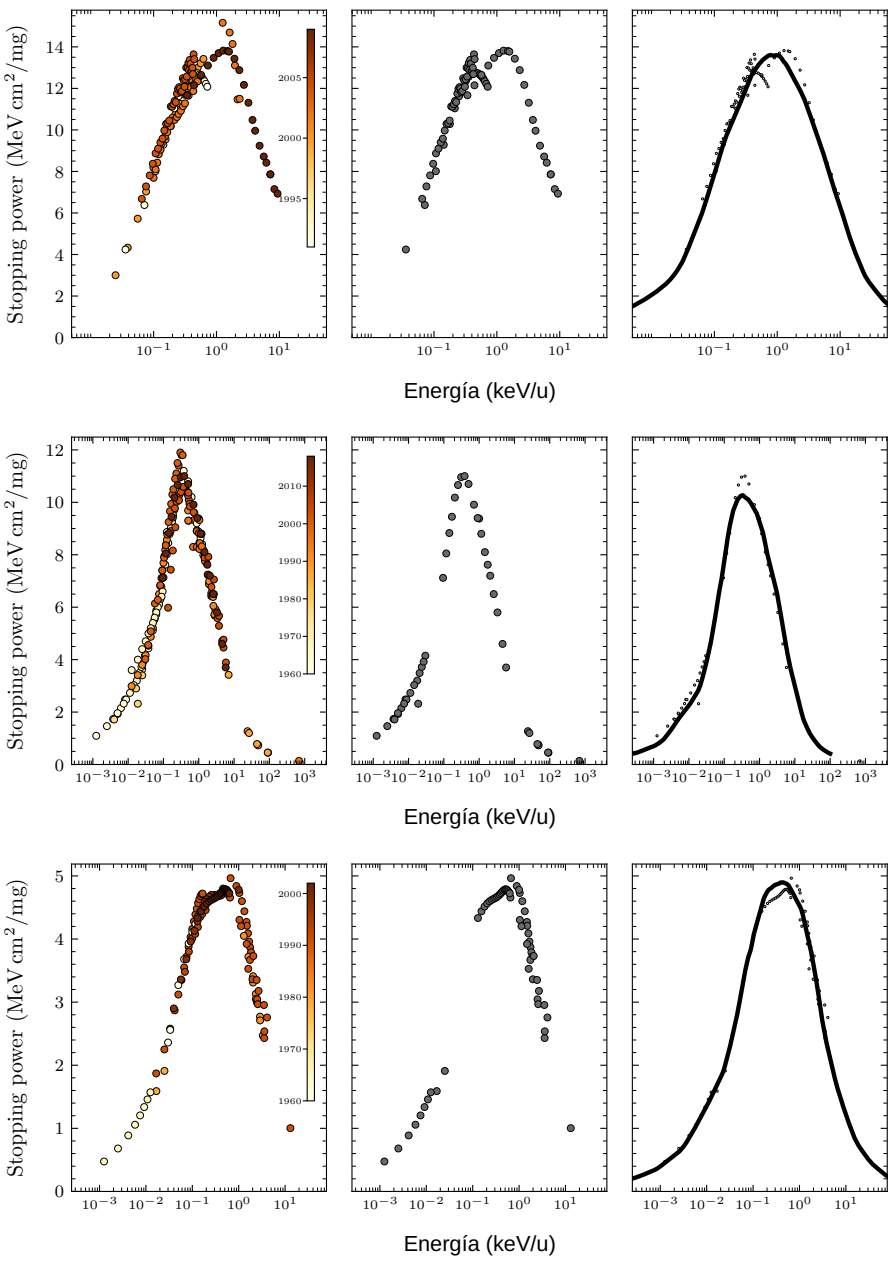



Figura 4.6: Izquierda: Resultados experimentales para secciones eficaces de poder de frenado. Los colores indican el año de publicación de los datos. Centro: Datos filtrados. Derecha: Datos predichos por la red neuronal. Arriba: Si en Si. Medio: O en C. Abajo: C en Al.

Para verificar la capacidad de generalización del modelo, calculamos las secciones eficaces del poder de frenado para los puntos de datos experimentales inicialmente excluidos del procedimiento de entrenamiento, o sea del `conjunto de prueba`. Aunque estos valores no fueron incluidos en los conjuntos de entrenamiento ni de validación, nuestro modelo puede reproducir estos resultados con alta precisión, como se demuestra en la Fig. 4.7.

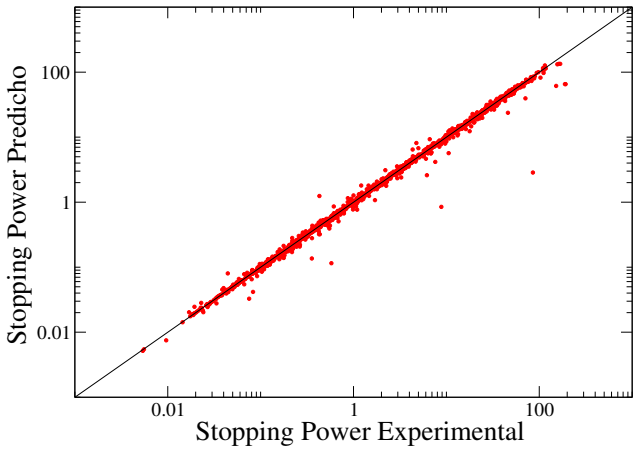



Figura 4.7: Valores predichos por ESPNN para las secciones eficaces del poder de frenado, comparados con los resultados experimentales. Todos los resultados corresponden al conjunto de prueba, datos nunca antes visto por el modelo.

Podemos examinar más profundamente el desempeño del ESPNN en el conjunto de prueba analizando los residuos dados por

$$R \equiv \frac{y_{\text{true}} - y_{\text{pred}}}{y_{\text{true}}} \ .$$

En la Fig. 4.8, mostramos los residuos del conjunto de prueba sobre el rango de energía experimental (panel izquierdo) y la distribución de frecuencia de estos valores (panel derecho). Podemos estimar un valor de error medio del 5 % a partir de esta figura. Debido a que los datos experimentales están dispersos en un amplio rango de incertidumbres y discrepancias, la precisión del presente modelo puede considerarse excepcional.

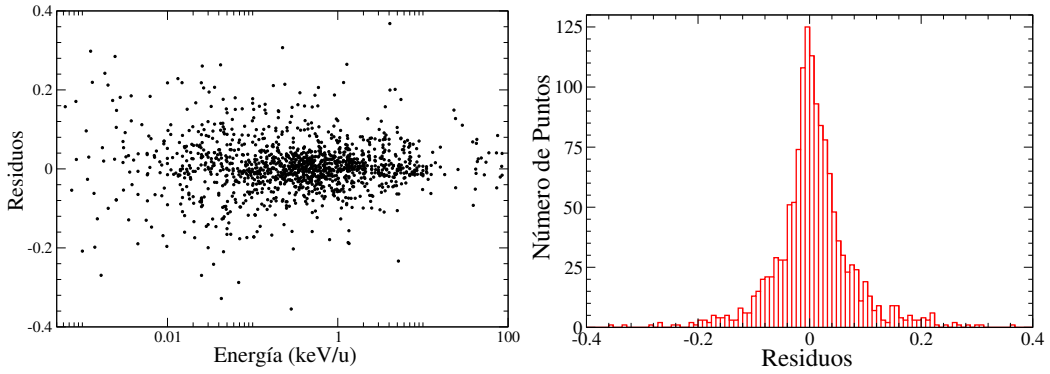



Figura 4.8: Error de los valores predichos por ESPNN para el conjunto de datos de prueba. Izquierda: valores residuales. Derecha: Histograma de los errores de predicción.

### 4.4.1. Comparación con otros métodos de ML

Para una comparación cuantitativa de nuestro modelo con otros enfoques, mostramos en la Tabla 4.3 los valores MAPE reportados por SRIM [39], el algoritmo de bosques aleatorios (RF - *random forest*) de Parfitt y Jackman [37], el algoritmo de aprendizaje profundo (DL - *deep learning*) de Guo *et al.* [34], y los cálculos actuales de ESPNN. La tabla no proporciona una comparación rigurosa entre estos métodos porque muestra los valores según lo informado por los respectivos autores, y se consideraron diferentes números de puntos de datos en la evaluación final de MAPE. Por esa razón, separamos la tabla en dos partes, mostrando a la izquierda los errores publicados para el conjunto de entrenamiento y, a la derecha, el MAPE en el conjunto de prueba. Sin embargo, solo podemos hacer una comparación adecuada con los cálculos de RF reportados por Parfitt y Jackman. Aunque el tamaño y las características de los conjuntos de entrada y prueba de este trabajo se eligieron de manera diferente a los nuestros, estos autores abordaron el entrenamiento y un conjunto de prueba no visto por separado. Los valores MAPE en la Tabla 4.3 proporcionan una imagen clara del rendimiento de varios códigos. Además, los resultados de ESPNN en el conjunto de prueba no visto son excelentes. Algunos detalles que vale la pena mencionar: SRIM es un código semiempírico que considera todos los datos experimentales

publicados hasta 2013. Por lo tanto, los respectivos errores no corresponden necesariamente a su capacidad predictiva en mediciones posteriores. Por otra parte, RF describe muy bien los datos de entrenamiento pero los sobreajusta; los altos valores de MAPE para el conjunto de prueba pueden ser una indicación de esto.

| Iones | Entrenamiento | | | Prueba | |
|---|---|---|---|---|---|
| | SRIM | RF | DL | RF | ESPNN |
| H | 4.0 | 2.4 | 5.6 | 12 | **4.5** |
| He | 3.9 | 2.1 | 6.0 | 9.1 | **4.1** |
| Li | 4.8 | 2.1 | 4.2 | 25 | **7.7** |
| Be-U | 5.8 | 2.0 | 7.0 | 12 | **7.1** |
| All | 4.6 | 2.1 | 5.7 | 23 | **5.7** |

Tabla 4.3: Comparaciones entre los valores MAPE reportados por SRIM [39], el algoritmo de bosques aleatorios (RF) de Parfitt y Jackman [37], el algoritmo de aprendizaje profundo (DL) de Guo *et al.* [34], y los cálculos actuales (ESPNN). Columnas izquierdas: valores de error en el conjunto de entrenamiento. Columnas derechas: valores de error en el conjunto de prueba.

Debido a la importancia y el uso generalizado de SRIM, diseñamos un experimento detallado para comparar los cálculos de nuestro modelo. Intentando hacer una comparación justa entre nuestras predicciones y SRIM, realizamos el siguiente procedimiento: entrenamos la red neuronal solo con resultados publicados antes de 2013. Esto significa que aplicamos nuestra heurística basada en DBSCAN a todos los datos publicados hasta 2013. Los valores experimentales restantes (publicados después de 2013) se dejaron aparte, y quedaron sin limpiar para evitar la contaminación de las métricas a extraer. El conjunto excluido (el nuevo conjunto de prueba), se separó además en diferentes grupos de forma iterativa. Primero, consideramos todos los resultados experimentales publicados después de 2013, y luego, todas las publicaciones que aparecieron después de 2015, y sucesivamente redujimos el conjunto de prueba para incluir solo las publicaciones más recientes. Es destacable que esta red neuronal se entrenó solo una vez, con los datos recopilados en la base de datos de la IAEA hasta 2013. De esta manera, podemos comparar el poder de predicción de ambos métodos en igualdad de condiciones. Los resultados se resumen en la Tabla 4.4. Esta

tabla muestra que el poder de predicción de ESPNN es consistentemente mejor que el de SRIM, en todos los casos. Además, comparando estos resultados con los valores MAPE obtenidos con el entrenamiento completo de datos, observamos que el poder de predicción mejora, como era de esperar, a medida que se incluyen nuevos valores en el conjunto de datos.

| | >2013 | | >2015 | | >2017 | | >2019 | |
|---|---|---|---|---|---|---|---|---|
| Iones | ESPNN | SRIM | ESPNN | SRIM | ESPNN | SRIM | ESPNN | SRIM |
| H | 7.0 | 19.2 | 4.6 | 15.3 | 6.6 | 13.3 | 3.5 | 7.1 |
| He | 8.4 | 10.6 | 9.3 | 10.1 | 9.3 | 10.1 | 5.5 | 8.3 |
| Be-U | 6.2 | 6.6 | 5.4 | 6.5 | 5.3 | 6.3 | | |
| all | 7.0 | 11.4 | 6.0 | 9.8 | 6.8 | 8.9 | 4.1 | 7.4 |

Tabla 4.4: MAPE predictivo del código semiempírico SRIM [39] y el modelo de red neuronal ESPNN entrenado solo con datos recopilados antes de 2013. Las primeras columnas muestran las predicciones de error para todos los valores experimentales reportados después de 2013. La segunda incluye todos los valores reportados después de 2015. Las últimas columnas muestran el error de las predicciones para los resultados experimentales producidos después de 2017 y 2019.

Por completitud, agregamos los resultados de la Tabla 4.5. Donde se muestra que usar el filtrado por DBSCAN es beneficioso, en el conjunto de prueba sin intervenir, para los distintos cortes temporales. De esta manera queda saldada la cuestión de la efectividad del filtrado planteada en el capítulo anterior.

| | >2013 | | >2015 | | >2017 | | >2019 | |
|---|---|---|---|---|---|---|---|---|
| Iones | ESPNN | ESPNN sin filtrar | ESPNN | ESPNN sin filtrar | ESPNN | ESPNN sin filtrar | ESPNN | ESPNN sin filtrar |
| H | 7.0 | 6.8 | 4.6 | 4.3 | 6.6 | 5.6 | 3.5 | 4.5 |
| He | 8.4 | 10.5 | 9.3 | 11.5 | 9.3 | 11.5 | 5.5 | 6.3 |
| Be-U | 6.2 | 6.4 | 5.4 | 5.7 | 5.3 | 5.8 | | |
| all | 7.0 | 7.6 | 6.0 | 7.5 | 6.8 | 7.5 | 4.1 | 4.9 |

Tabla 4.5: MAPE predictivo del m Desde un punto de vista estad´ıstico, la base de datos empleada contiene un total de 36 000 puntos experimen- talesodelo de red neuronal ESPNN entrenado solo con datos recopilados antes de 2013. En un caso los datos de entrenamiento son filtrados con la heuristica basada en DBSCAN, en ambos casos los datos de validación son tal cual se encuentran en la base de datos de IAEA sin ningún tipo de filtrado. Las primeras columnas muestran las predicciones de error para todos los valores experimentales reportados después de 2013. La segunda incluye todos los valores reportados después de 2015. Las últimas columnas muestran el error de las predicciones para los resultados experimentales producidos después de 2017 y 2019.

# Capítulo 5

# Modelos para blancos pluri-elementales

## 5.1. Introducción

Cuando el material del blanco está compuesto por elementos diversos, el cálculo del poder de frenado se vuelve aún más complejo. Esto se debe a que los compuestos multi-elementales presentan composiciones químicas heterogéneas y estructuras geométricas variadas, y pueden existir en distintas fases y conformaciones. Estas características introducen un alto grado de variabilidad estructural y electrónica que dificulta la modelización precisa del fenómeno. La base de datos de la IAEA, si bien es una fuente valiosa de datos experimentales, no contiene suficiente información sobre compuestos complejos como para que los modelos de aprendizaje automático puedan aprender representaciones robustas que capturen adecuadamente la física subyacente de estos sistemas. Para abordar este desafío, proponemos reutilizar representaciones previamente aprendidas por modelos con fundamentos físicos, en particular el modelo SchNet [6], el cual ha sido entrenado sobre conjuntos de datos de gran escala, como MaterialsProject [40]. Estas representaciones, que codifican información estructural y electrónica de materiales, se utilizarán como entrada para una nueva red neuronal que será entrenada específicamente para predecir el poder de frenado

en compuestos complejos. Este enfoque de transferencia de conocimiento permite aprovechar la riqueza estructural y química contenida en grandes bases de datos computacionales, facilitando la generalización del modelo a dominios con escasez de datos experimentales.

## 5.2. SchNet

### 5.2.1. Estructura de SchNet

Al revisar la literatura especializada, se observa la existencia de diversos modelos de redes neuronales diseñados para predecir propiedades de materiales a partir de su estructura atómica o molecular. Uno de los modelos más representativos en esta área es SchNet [6], una red neuronal basada en aprendizaje profundo que opera directamente sobre representaciones estructurales. SchNet toma como entrada las coordenadas espaciales de los átomos $r_i$ y un vector de *embedding* inicial $x_i^0$, que codifica características específicas de cada átomo (por ejemplo, su número atómico). A través del proceso de propagación hacia adelante, la red actualiza estos vectores en cada capa, refinándolos y agregando información proveniente del entorno local de cada átomo. De este modo, el vector $x_i^l$ en la capa $l$ representa una versión enriquecida del embedding del átomo $i$, incorporando progresivamente las contribuciones de sus vecinos a medida que se avanza en la red. Este enfoque permite capturar interacciones atómicas de corto y largo alcance, lo que resulta particularmente útil para modelar propiedades materiales complejas de forma eficiente y físicamente informada.

La arquitectura de la red neuronal incluye tanto capas de actualización local, que operan átomo por átomo, como una capa de tipo convolucional continuo (`FCConv`). Las capas átomo-a-átomo actualizan únicamente las representaciones individuales de cada átomo, sin considerar su entorno inmediato. En cambio, la capa `FCConv` generaliza el concepto de convolución tradicional, permitiendo capturar interacciones entre átomos a distintas distancias espaciales, incluso si estos se encuentran ubicados en posiciones arbitrarias dentro del sistema. A diferencia de las redes convolucionales tradicionales, donde los filtros actúan sobre una grilla regular (como píxeles en una

imagen), en este caso los filtros dependen explícitamente de la distancia relativa entre átomos. Específicamente, dada una representación a nivel atómico $x^l$ en posiciones $\mathbf{r}$, la actualización de la representación del átomo $i$ en la capa $l+1$ se define mediante una convolución sobre los vecinos circundantes:

$$x_i^{l+1} = (X^l * W^l)_i = \sum_{j=0}^{n_{\text{atoms}}} x_j^l \circ W^l(\mathbf{r}_j - \mathbf{r}_i)\,, \tag{5.1}$$

donde "$\circ$" denota la multiplicación elemento a elemento.

Esta operación se realiza de forma separada para cada característica, lo que permite una implementación eficiente desde el punto de vista computacional. La interacción entre diferentes características se lleva a cabo posteriormente mediante capas adicionales de tipo átomo-a-átomo. En lugar de definir un banco fijo de filtros convolucionales como en redes tradicionales, se utiliza una red generadora de filtros $W^l : \mathbb{R}^3 \to \mathbb{R}^F$, la cual mapea las diferencias de posición entre átomos al espacio de filtros. Esto permite construir filtros continuos adaptados a la geometría tridimensional del sistema atómico.

### 5.2.2. SchNet para el cálculo de la energía de formación

Siguiendo el enfoque propuesto en SchNet [6], reentrenamos la red utilizando la versión actualizada de la base de datos MaterialsProject, con el objetivo de predecir las energías de formación de compuestos cristalinos. Esta base de datos contiene una amplia variedad de estructuras cristalinas tridimensionales, incluyendo materiales con elementos que abarcan toda la tabla periódica hasta el número atómico $Z = 94$. La arquitectura empleada incluye $T = 6$ bloques de interacción y representaciones atómicas de dimensión $F = 64$. Se establece un radio de corte $r_{\text{cut}} = 5$ Å para limitar las interacciones entre átomos. Con esta configuración, se descartan dos ejemplos de la base de datos que presentan átomos aislados, al no tener vecinos dentro del radio de corte. Posteriormente, se realiza una partición aleatoria de los datos, conformando un conjunto de entrenamiento con 60 000 estructuras, un conjunto de validación con 4 500 ejemplos y el resto de la base como conjunto de prueba. El error absoluto medio (MAE) obtenido en este reentrenamiento resultó

comparable al reportado en [6], donde se informa un valor de 0.127 `eV/átomo` para la predicción de energía de formación.

### 5.2.3. Arquitectura de la red para blancos pluri-elementales

Para el caso de blancos multi-elementales, se mantiene la estructura general de la red neuronal descripta en la Sección 4.2, con modificaciones específicas en la capa de entrada. En particular, esta capa se amplía a una dimensión de 43 para incorporar de forma adecuada las distintas fuentes de información relevantes para el problema. Los vectores de entrada incluyen la representación atómica extraída de la última capa oculta del modelo SchNet, previamente entrenado sobre la base de datos MaterialsProject. También incluyen un vector adicional de 4 componentes que codifica la fase del material del blanco. También incluye las mejores características utilizadas en el caso mono-elemental, identificadas previamente: número atómico del proyectil ($Z_p$), masa atómica del proyectil ($m_p$), número atómico efectivo del blanco ($Z_t$), masa atómica efectiva del blanco ($m_t$), la energía incidente del proyectil en escala logarítmica ($\log(E_i)$), y la energía de ionización del blanco.

Estas entradas combinan información estructural, fisicoquímica y electrónica, permitiendo que la red neuronal aprenda a modelar de forma más generalizada el poder de frenado en materiales compuestos.

### 5.2.4. Características y párametros de aprendizaje

La forma de los lotes (*batches*) utilizados durante el entrenamiento queda dada por:

```
[Longitud del batch, máximo número de átomos, dimensión del embedding
        + codificación de fase + características manuales]
```

Dado que los distintos compuestos pueden tener un número variable de átomos, se aplica un *relleno* (*padding*) a aquellos con menor cantidad de átomos, de modo

que todos los elementos del lote tengan una estructura homogénea. Estas representaciones se introducen al modelo, y en el último paso se calcula un promedio sobre las salidas correspondientes a cada átomo del compuesto. Es importante señalar que este promedio no equivale a promediar la contribución de átomos individuales tratados como entidades aisladas, ya que las representaciones embebidas (*embeddings*) de cada átomo contienen información contextual que depende del entorno químico particular dentro del material. Un aspecto importante a considerar es que un mismo material puede presentarse en distintas conformaciones cristalinas, es decir, diferentes geometrías atómicas que comparten la misma composición química. Estas conformaciones pueden exhibir propiedades macroscópicas notablemente distintas, tales como la densidad, la energía de formación, el módulo elástico, entre otras. Esta diversidad estructural introduce una complejidad adicional en la predicción del poder de frenado, ya que es necesario que el modelo neuronal sea capaz de generalizar adecuadamente frente a estas variaciones geométricas y físicas.

Para el entrenamiento de la red neuronal se utilizó el optimizador de estimación de momento adaptativo (Adam) como algoritmo de minimización. Se emplearon tasas de *dropout* de 0.2 en las capas de entrada y salida, y de 0.5 en cada capa oculta, con el objetivo de prevenir el sobreajuste. La tasa de aprendizaje fue fijada en $\alpha = 0{,}001$, el tamaño del lote en $b = 64$, y se aplicó una penalización por norma de los pesos (degradación de peso) con parámetro $\lambda = 10^{-10}$. El entrenamiento se llevó a cabo durante un máximo de 300 épocas, con un criterio de parada temprana activado si no se observaba mejora en el conjunto de validación durante 50 épocas consecutivas. Además, se utilizó el truco de reparametrización propuesto en [38], lo que permitió acelerar la convergencia y mejorar la estabilidad numérica del proceso de optimización.

Se evaluaron tres heurísticas distintas para abordar el problema de múltiples conformaciones cristalinas de un mismo material: utilizar la conformación con mínima energía de formación, aquella con mínima densidad, y finalmente el promedio de las representaciones correspondientes a todas las conformaciones disponibles (opción que modela un material amorfo o con dominios cristalinos variados). La estrategia que arrojó mejores resultados fue la del promedio sobre todas las conformaciones, como

se muestra en la Tabla 5.1.

Para esta comparación, cada modelo fue entrenado siguiendo un protocolo común. Se separó aleatoriamente un 10 % del conjunto total como subconjunto de prueba. Posteriormente, el conjunto de entrenamiento fue purificado mediante nuestro método basado en DBSCAN. Luego, se entrenaron cinco modelos independientes mediante validación cruzada *5-fold*, y los resultados presentados corresponden al promedio de la función de costo obtenida en los subconjuntos de validación.

Tabla 5.1: Función de costo promedio en validación para diferentes estrategias de tratamiento de conformaciones cristalinas.

| Estrategia | Costo en validación |
|---|---|
| Mínima densidad | 4.41 |
| Mínima energía de formación | 4.29 |
| Promedio de conformaciones | **3.03** |
| SRIM | 9.64 |

Tal como se realizó en la Sección 4.4, evaluamos la consistencia de nuestra estrategia de entrenamiento para distintos cortes temporales. El protocolo seguido consiste en separar como conjunto de prueba todos los datos posteriores a un cierto año; luego, se purifica el conjunto de entrenamiento mediante nuestro método basado en DBSCAN, y se entrena un conjunto de cinco modelos independientes utilizando validación cruzada *5-fold*. Finalmente, las predicciones de cada *fold* se promedian para obtener la predicción final del modelo.

En la Tabla 5.2 se presentan los resultados obtenidos para diferentes elecciones del año de corte temporal aplicado al conjunto de prueba. Se observa que el desempeño del modelo propuesto se mantiene consistente a lo largo del tiempo y supera sistemáticamente las predicciones del modelo SRIM.

| | >2013 | | >2015 | | >2017 | | >2019 | |
|---|---|---|---|---|---|---|---|---|
| Iones | ESPNN | SRIM | ESPNN | SRIM | ESPNN | SRIM | ESPNN | SRIM |
| H | 6.6 | 19.2 | 5.2 | 16.2 | 8.3 | 12.9 | 3.5 | 7.1 |
| He | 6.2 | 8.5 | 6.4 | 7.9 | 6.4 | 9.4 | 5.5 | 8.3 |
| Be-U | 5.4 | 20.1 | 5.96 | 70.9 | 5.11 | 6.9 | | |
| all | 5.9 | 24.4 | 5.4 | 40.8 | 6.3 | 8.9 | 4.1 | 7.4 |

Tabla 5.2: MAPE predictivo del código semiempírico SRIM [39] y el modelo de red neuronal ESPNN entrenado solo con datos recopilados antes de 2013. Las primeras columnas muestran las predicciones de error para todos los valores experimentales reportados después de 2013. La segunda incluye todos los valores reportados después de 2015. Las últimas columnas muestran el error de las predicciones para los resultados experimentales producidos después de 2019.

# Capítulo 6

# Conclusiones

Como se enunció al comienzo de este trabajo, el objetivo principal de esta Tesis fue contribuir al desarrollo de sistemas más precisos para la predicción del poder de frenado (*Stopping Power*, SP) de iones en materiales, utilizando modelos de aprendizaje profundo a partir de los datos experimentales disponibles en la base de datos de la IAEA [16].

En el Capítulo 2, se describió el proceso de limpieza, depuración y modernización aplicado a dicha base de datos. Como resultado, se logró una versión accesible, organizada y reutilizable por parte de la comunidad científica. Asimismo, se identificaron combinaciones de proyectil-blanco con escasez de datos y se analizó la evolución temporal del interés experimental en diferentes sistemas colisionales.

En el Capítulo 3, se abordaron dos enfoques distintos para la eliminación de ruido en los datos. En particular, se implementó un sistema basado en el algoritmo DBSCAN, priorizando la conservación de datos densamente agrupados y provenientes de mediciones recientes. Esta estrategia permitió obtener conjuntos de datos más coherentes y con menor dispersión, lo que a su vez resultó en mejoras significativas en la capacidad de generalización de los modelos, especialmente sobre datos correspondientes a épocas futuras no incluidas en el entrenamiento.

A partir de este conjunto depurado, se construyó una red neuronal cuya arquitectura y resultados fueron presentados en [27]. En el Capítulo 4, se describió la red y

se evaluó su desempeño mediante las métricas MAE y MAPE, comparándolo con el modelo pseudo-empírico SRIM, y se verificó su superioridad en distintos escenarios de evaluación temporal. Además, el modelo fue publicado junto con su correspondiente estimación de incertidumbre.

En el Capítulo 5, se extendió el modelo hacia materiales compuestos o pluri-elementales. Para ello, se incorporaron representaciones estructurales extraídas mediante el modelo SCHNET, lo que permitió incluir información detallada sobre la geometría cristalina y molecular de los materiales. Los resultados obtenidos muestran que esta extensión no solo mantiene la precisión del modelo en sistemas simples, sino que amplía su aplicabilidad a una gama mucho más amplia de compuestos, superando nuevamente al enfoque pseudo-empírico tradicional.

# Bibliografía


[1] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91(4):045002, 2019.

[2] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv:2010.08895*, 2020.

[3] K. T. Butler, D. W. Davies, H. Cartwright, and O. Isayev. Machine learning for molecular and materials science. *Nature*, 559:547–555, 2018.

[4] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.*, 145:170901–170911, 2016.

[5] K.T Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller, and A. Tkatchenko. Quantum–chemical insights from deep tensor neural networks. *Nat. Commun.*, 8:13890, 2017.

[6] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. Schnetpack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.*, 15(1):448–455, 2019.

[7] J. Hermann, Z. Schätzle, and F. Noé. Deep-neural–network solution of the electronic Schrödinger equation. *Nat. Chem.*, 12(10):891–897, 2020.

[8] K. Mills, M. Spanner, and I. Tamblyn. Deep learning and the Schrödinger equation. *Phys. Rev.* **A**, 96:042113–04222, 2017.

[9] P. Sigmund. *Particle Penetration and Radiation Effects, General Aspects and Stopping of Swift Point Charges*, volume 1. Springer, Berlin, Heidelberg, 2006.

[10] A. R. Smith. Proton therapy. *Phys. Med. & Biol.*, 51(13):R491, 2006.

[11] Z. A. Alrowaili, J. S. Alzahrani, H. Arslan, N. S. Alruwaili, C. Mutuwong, and M. S. Al-Buriahi. Bragg curve, dose distribution, and target fragmentation for thyroid proton therapy. *Radiat. Phys. Chem.*, 212:111118, 2023.

[12] List of Paul's publications on stopping power. https://www-nds.iaea.org/stopping/stoppingdocu.html.

[13] H. Paul. Stopping power for light ions. Original webpage by H. Paul at http://www.exphys.jku.at/stopping/. Available online at: https://www-nds.iaea.org/stopping-legacy/stopping_201510/, 2015.

[14] S. Rosenblum. Recherches expérimentales sur le passage des rayons travers la matière. *Annales de Physique*, 10(10):408–471, 1928.

[15] https://www.iaea.org, .

[16] https://www-nds.iaea.org/stopping/, .

[17] C. C. Montanari and P. Dimitriou. The IAEA stopping power database, following the trends in stopping power of ions in matter. *Nucl. Instrum. Methods Phys. Res.* **B**, 408:50–55, 2017.

[18] C.C. Montanari, P. Dimitriou, L. Marian, A.M.P. Mendez, J.P. Peralta, and F. Bivort-Haiek. The IAEA electronic stopping power database: Modernization, review, and analysis of the existing experimental data. *Nucl. Instrum. Methods Phys. Res.* **B**, 551:165336, 2024.

[19] N. R. Arista and A. F. Lifschitz. Non-linear approach to the energy loss of ions in solids. *Adv. Quantum Chem.*, 45:47–77, 2004.

[20] H. Paul. New results about stopping power for positive ions: Experiment and theory. In *AIP Conference Proceedings*, volume 1525, pages 295–299. American Institute of Physics, 2013.

[21] J. P. Peralta, A. M. P. Mendez, D. M. Mitnik, and C. C. Montanari. Stopping power of iron for protons: Theoretical calculations from very low to high energies. *Atoms*, 13(3), 2025.

[22] H. Paul. New developments in stopping power for fast ions. *Nucl. Instrum. Methods Phys. Res.* **B**, 261(1-2):1176–1179, 2007.

[23] International Commission on Radiation Units. *Stopping of ions heavier than helium*. Number 73. Oxford University Press, 2005.

[24] J.F. Ziegler, J.P. Biersack, and M.D. Ziegler. *SRIM, The Stopping and Range of Ions in Matter*. SRIM Co. Maryland, 2008. http://www.srim.org/.

[25] J. F. Ziegler, J. P. Biersack, and U. Littmark. *The Stopping and Range of Ions in Solids*. Pergamon Press, 1985.

[26] https://www-nds.iaea.org/stopping-legacy/, .

[27] F. Bivort Haiek, A. M. P. Mendez, D. M. Mitnik, and C. C. Montanari. ESPNN: A novel electronic stopping power neural-network code built on the IAEA stopping power database. I. Atomic targets. *J. Appl. Phys.*, 132(24):245103, 2022.

[28] X. Guo, H. Wang, S. Zhao, K. Jin, and J. Xue. A high accuracy electrical stopping power prediction model based on deep learning algorithm and its applications. *arXiv:2010.09943*, 2020.

[29] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine Learning*, pages 1096–1103, 2008.

[30] S. Kiranyaz, T. Ince, R. Hamila, and M. Gabbouj. Convolutional neural networks for patient-specific ecg classification. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2608–2611, 2015.

[31] H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[33] Y. Ozaki, Y. Tanigaki, S. Watanabe, M. Nomura, and M. Onishi. Multiobjective tree-structured Parzen estimator. *Journal of Artificial Intelligence Research*, 73: 1209–1250, 2022.

[34] Xun Guo, Hao Wang, Changkai Li, Shijun Zhao, Ke Jin, and Jianming Xue. Development of an electronic stopping power model based on deep learning and its application in ion range prediction. *Chinese Phys.* **B**, 31:073402, 2022.

[35] M. Ester, H.-P. Kriegel, Jörg Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery in Databases (KDD)*, volume 96, pages 226–231, 1996.

[36] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, 42 (3):1–21, 2017.

[37] W. A. Parfitt and R. B. Jackman. Machine learning for the prediction of stopping powers. *Nucl. Instrum. Methods Phys. Res.* **B**, 478:21, 2020.

[38] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[39] J. F. Ziegler, M. D. Ziegler, and J. P. Biersack. Srim–the stopping and range of ions in matter. *Nucl. Instrum. Methods Phys. Res.* **B**, 268:1818, 2010.

[40] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

[41] M. V. Moro, T. F. Silva, A. Mangiarotti, Z. O. Guimaraes-Filho, M. A. Rizzutto, N. Added, and M. H. Tabacniks. Traceable stopping cross sections of al and mo elemental targets for 0.9–3.6-mev protons. *Phys. Rev.* **A**, 93(2):022704, 2016.

[42] https://www.kaggle.com/c/champs-scalar-coupling.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[44] C.C. Montanari and P. Dimitrou. The IAEA stopping power database, following the trends in stopping power of ions in matter. *Nucl. Instrum. Methods* **B**, 408: 50, 2017.

[45] H. A. Bethe and J. Ashkin. Experimental nuclear physics. *Wiley, New York*, 1953.

[46] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12), 2010.

[47] C. C. Montanari, C. D. Archubi, D. M. Mitnik, and J. E. Miraglia. Energy loss of protons in Au, Pb, and Bi using relativistic wave functions. *Phys. Rev.* **A**, 79(3):032903, 2009.

[48] C.C. Montanari, P.A. Miranda, E. Alves, A.M.P. Mendez, D.M. Mitnik, J.E. Miraglia, R. Correa, J. Wachter, M. Aguilera, N. Catarino, et al. Stopping power of hydrogen in hafnium and the importance of relativistic 4f electrons. *Phys. Rev.* **A**, 101(6):062701, 2020.

[49] W. A. Parfitt and R. B. Jackman. Machine learning for the prediction of stopping powers. *Nucl. Instrum. Methods Phys. Res.* **B**, 478:21–33, 2020.

[50] P. Sigmund. Kinetic theory of particle stopping in a medium with internal motion. *Phys. Rev.* **A**, 26:24970, 1982.

[51] Stopping Power of Matter for Ions. Computer Programs. https://www-nds.iaea.org/stopping/stoppingprog.html.

[52] N. P. Barradas, A. Bergmaier, K. Mizohata, M. Msimanga, J. Räisänen, T. Sajavaara, and A. Simon. Determination of molecular stopping cross section of 12C, 16O, 28Si, 35Cl, 58Ni, 79Br, and 127I in silicon nitride. *Nucl. Instrum. Methods Phys. Res.* **B**, 360:90, 2015.

[53] T. Materna, E. Berthoumieux, Q. Deshayes, D. Doré, M. Kebbiri, A. Letourneau, L. Thulliez, Y. H. Kim, U. Köster, and X. Ledoux. Stopping power of fission fragments in thin mylar and nickel foils. *Nucl. Instrum. Methods Phys. Res.* **B**, 505:1, 2021.

[54] M. V. Moro, P. Bauer, and D. Primetzhofer. Experimental electronic stopping cross section of transition metals for light ions: Systematics around the stopping maximum. *Phys. Rev.* **A**, 102:022808, 2020.

[55] F. F. Selau, H. Trombini, G. G. Marmitt, A. M. H. De Andrade, J. Morais, P. L. Grande, I. Alencar, M. Vos, and R. Heller. Stopping and straggling of 60–250-kev backscattered protons on nanometric Pt films. *Phys. Rev.* **A**, 102: 032812, 2020.

[56] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, 42(3):Article 19, 2017.

[57] H. Paul and A. Schinner. An empirical approach to the stopping power of solids and gases for ions from 3Li to 18Ar. *Nucl. Instrum. Methods Phys. Res. B*, 179 (3):299–315, 2001.