

IRONHACK

BACTERIA ID

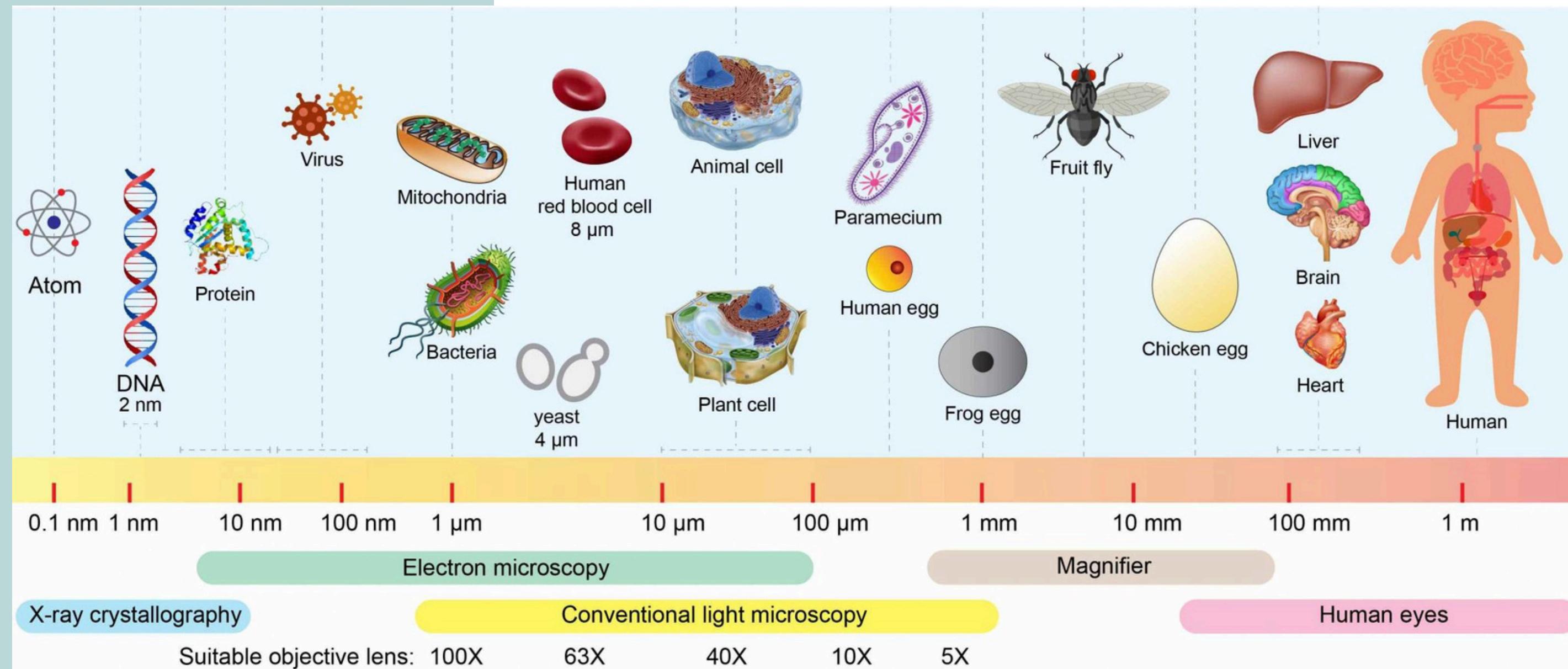
Presented by
Anastasia Tkacheva

Proyecto final del curso
Científico de datos

Introduction

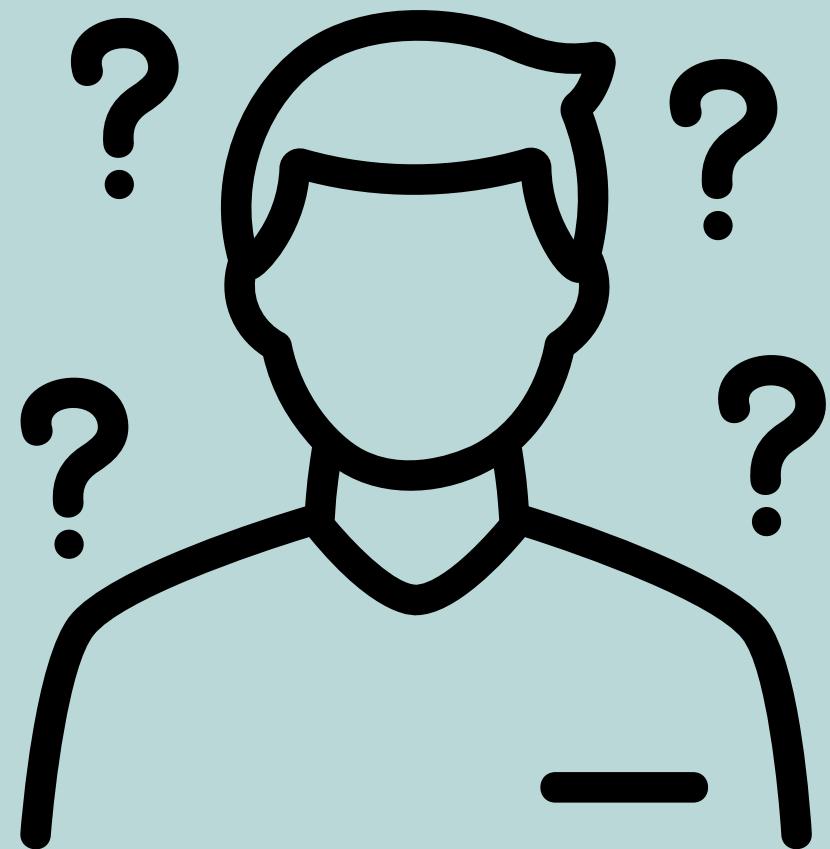
66

Las infecciones bacterianas son una de las principales causas de muerte tanto en los países desarrollados como en los países en desarrollo, cobrando más de 6,7 millones de vidas cada año.*



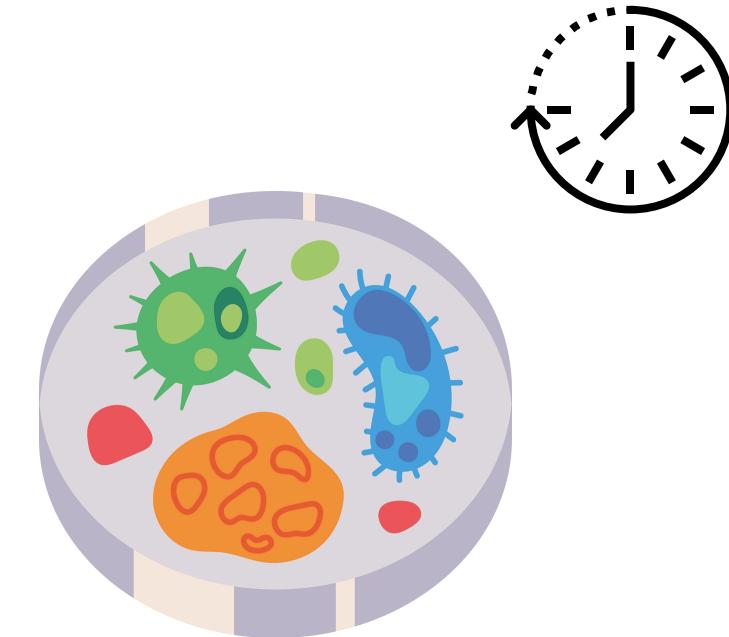
*Am. J. Respir. Crit. Care Med. 193, 259–272 (2016)

PLAN DE CONTENCION



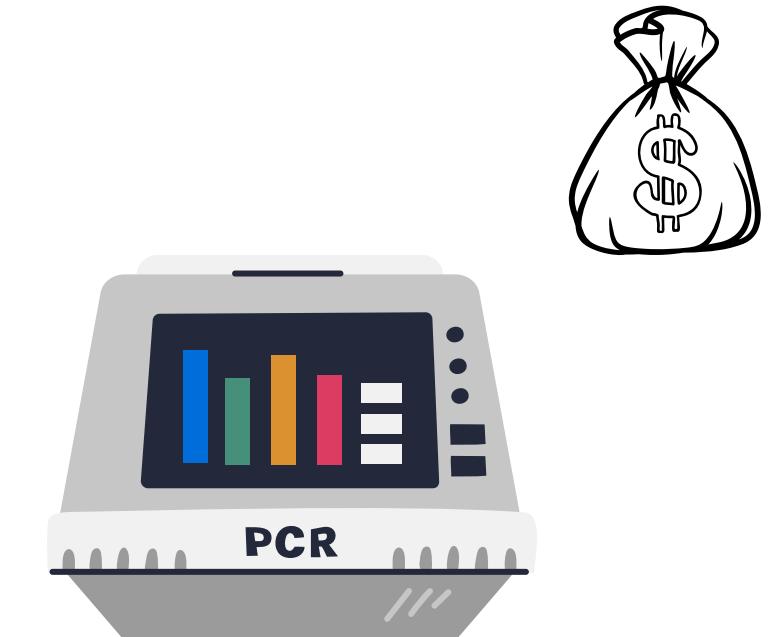
Medidas Preventivas

- Diseño Higiénico de Instalaciones
- Control de Calidad de Materias Primas
- Buenas Prácticas de Manufactura (BPM)
- ...



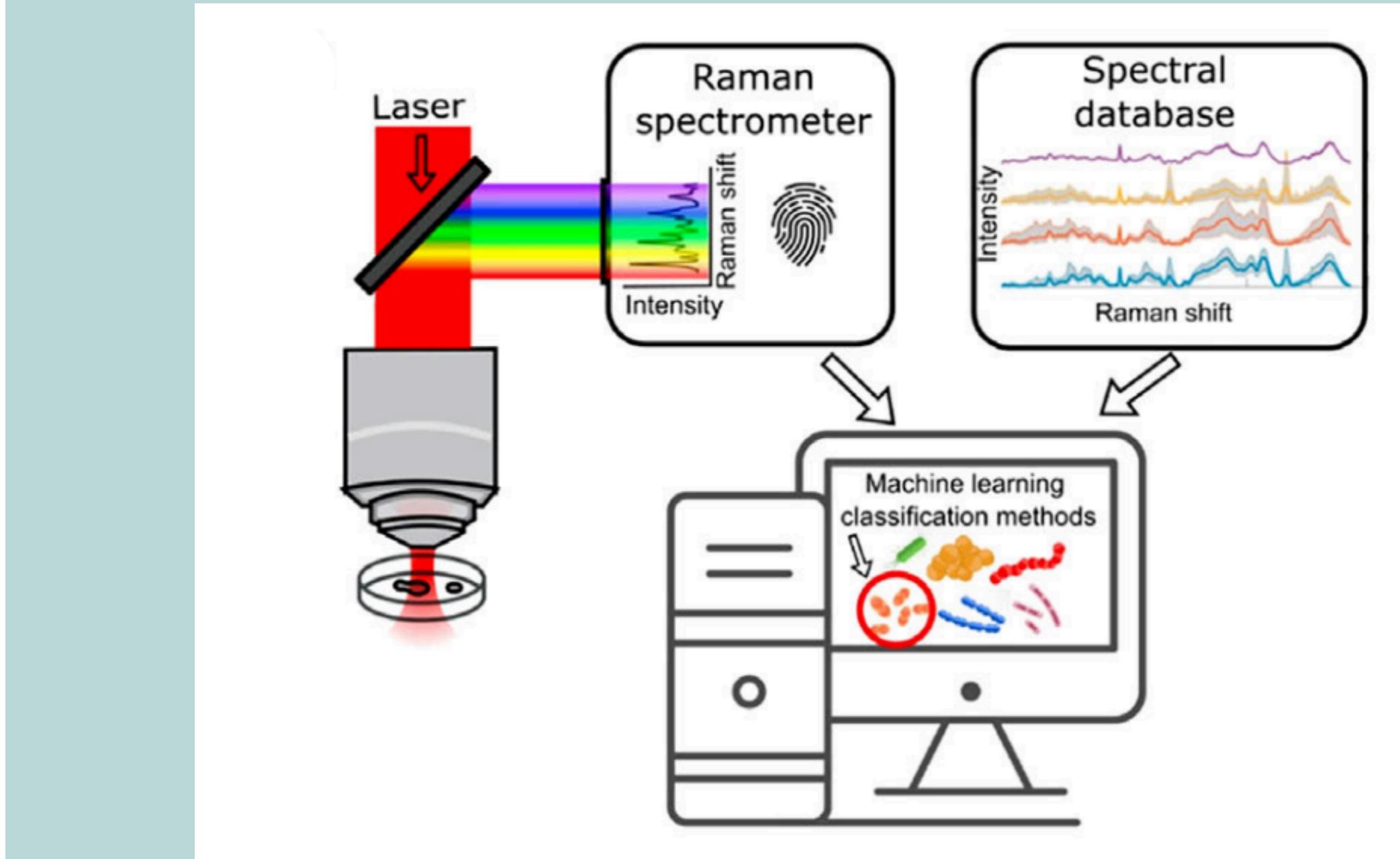
Medidas Correctivas:

- Antibióticos, biocida
- Drenaje de abscesos, cuidado de heridas
- Limpieza y Desinfección Rigurosa
- Control de Calidad del Agua
- Capacitación del Personal
- Pruebas Microbiológicas Regulares
- ...



Espectroscopia Raman

La espectroscopía Raman es una técnica que se utiliza cada vez más para la identificación y caracterización de bacterias debido a sus capacidades no invasivas y rápidas.



No Destructiva

La espectroscopía Raman no requiere la destrucción de la muestra, lo que permite análisis repetidos en la misma muestra



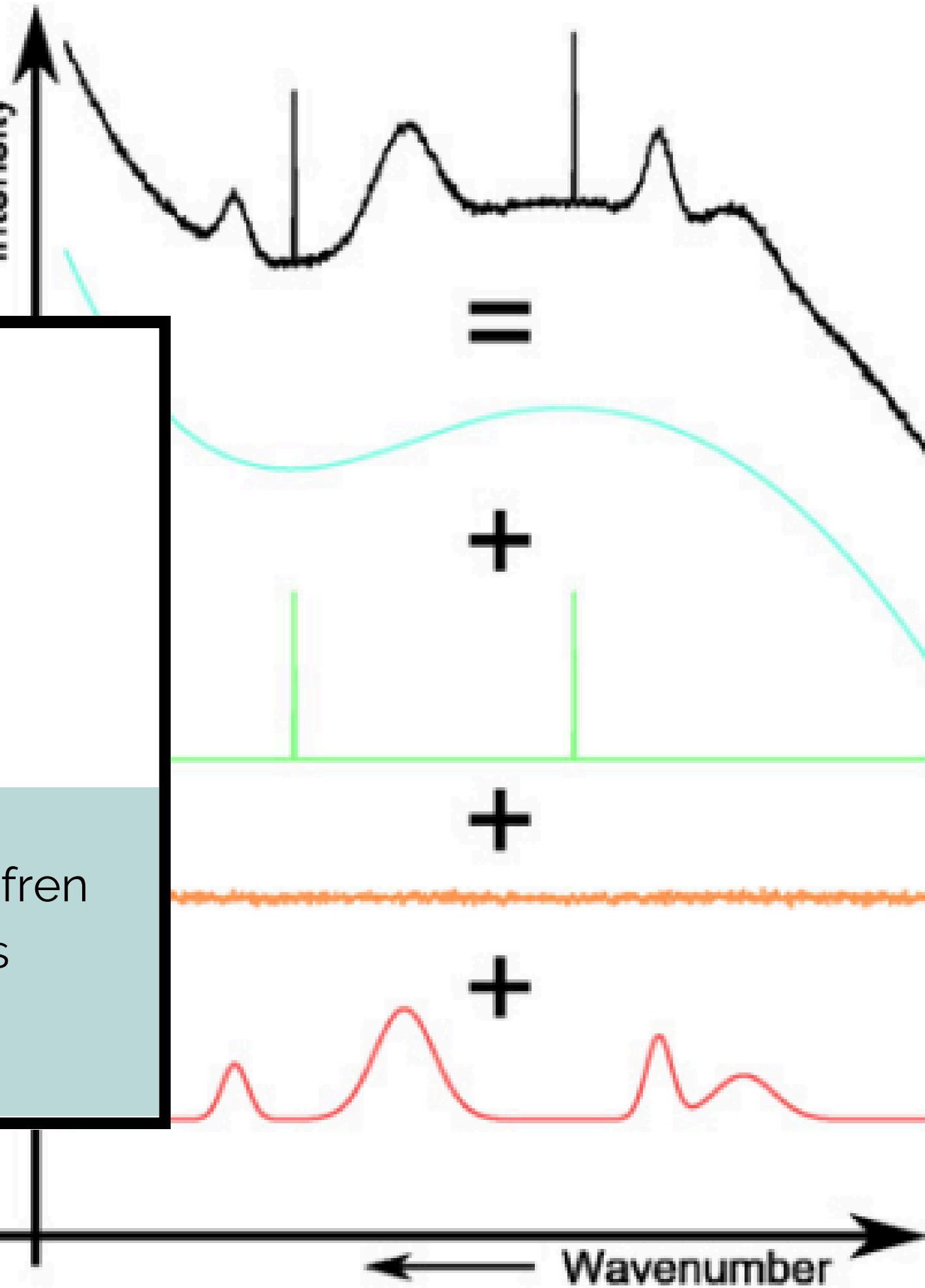
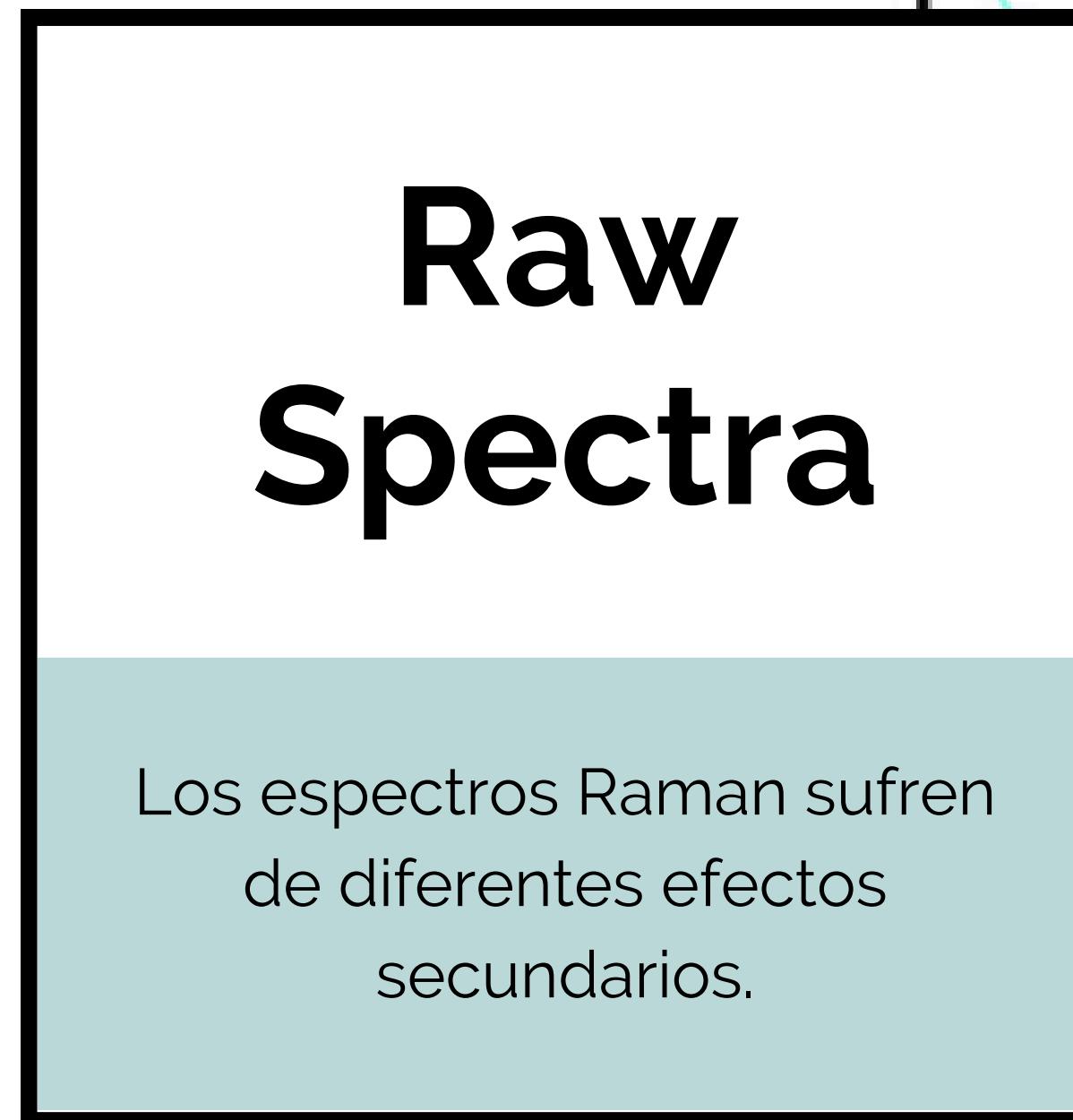
Rapida, analysis in situ

Ofrece resultados casi instantáneos, lo que es ventajoso para aplicaciones que requieren monitoreo en tiempo real o análisis rápido.



Machine Learning

Las técnicas de aprendizaje automático pueden extraer información valiosa de los datos espectrales de manera más eficiente, creando oportunidades sin precedentes para la ciencia analítica



Composition

Fluorescence Backg
+ CCD Baseline

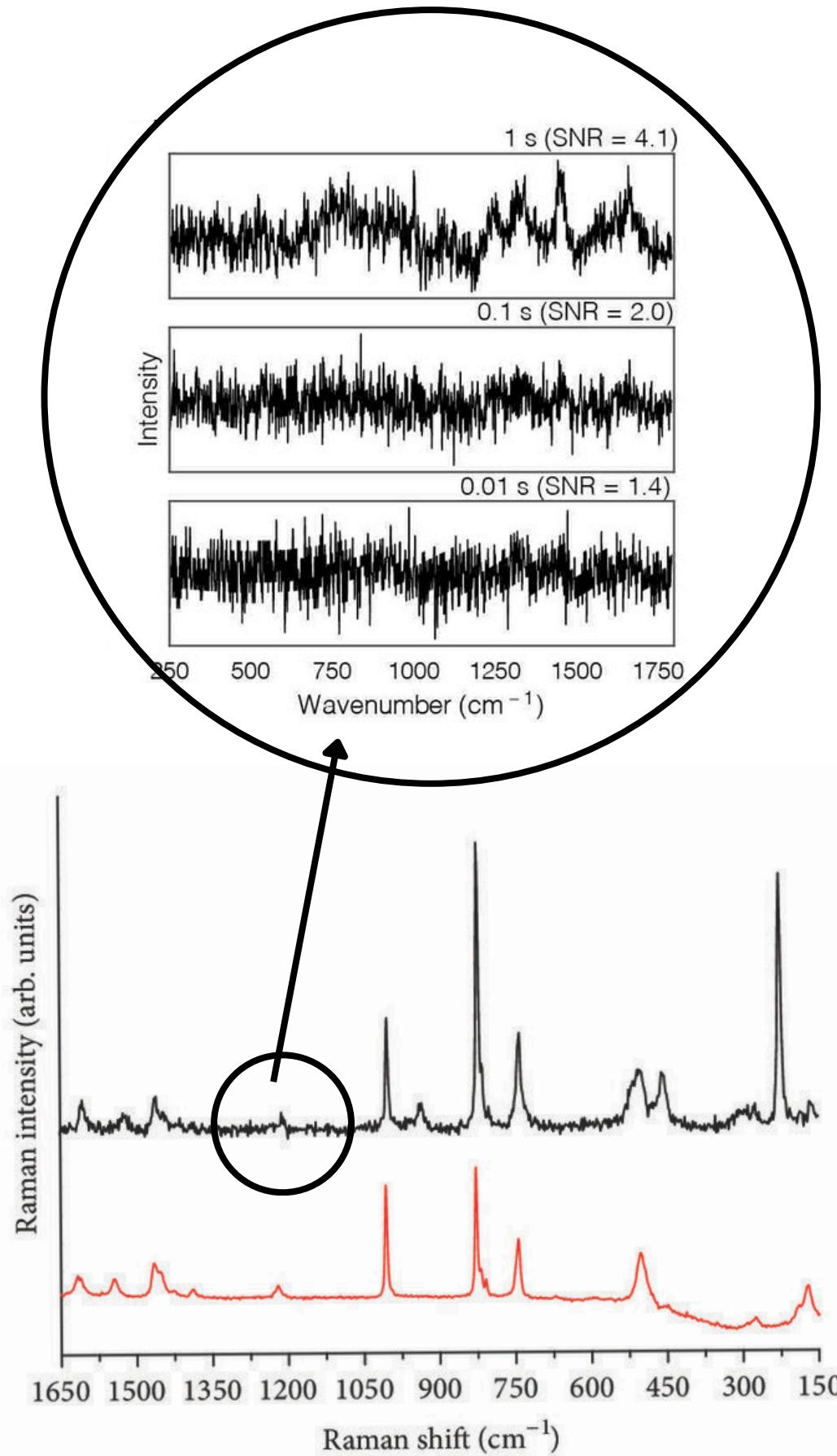
Cosmic Spikes

White Noise

Raman Spectrum

66

1 de cada 10 millones de fotones (o 0.00001%) que interactúan con una molécula producen un efecto Raman



Data

Resolucion

SNR = 4.1

Preprocesamiento

Los espectros Raman en bruto se limpiaron inicialmente para eliminar picos cosméticos. Posteriormente, se identificó y restó la función lineal entre los valores inicial y final de cada espectro. Como paso final de preprocesamiento, los espectros se normalizaron individualmente al rango entre cero y uno.

Nat Commun 10, 4927 (2019)

1 de cada 10 millones de fotones (o 0.00001%) que interactúan con una molécula producen un efecto Raman

MRSA 1 (isogenic)
MRSA 2
MSSA 1
MSSA 2
MSSA 3
<i>S. epidermidis</i>
<i>S. lugdunensis</i>
<i>S. pneumoniae</i> 1
<i>S. pneumoniae</i> 2
Group A Strep.
Group B Strep.
Group C Strep.
Group G Strep.
<i>S. sanguinis</i>
<i>E. faecalis</i> 1
<i>E. faecalis</i> 2
<i>E. faecium</i>
<i>E. coli</i> 1
<i>E. coli</i> 2
<i>K. pneumoniae</i> 1
<i>K. pneumoniae</i> 2
<i>K. aerogenes</i>
<i>E. cloacae</i>
<i>P. mirabilis</i>
<i>S. marcescens</i>
<i>S. enterica</i>
<i>P. aeruginosa</i> 1
<i>P. aeruginosa</i> 2
<i>C. albicans</i>
<i>C. glabrata</i>

Data

Dataset

30 especies *2000 spetcra = 60000 spectra

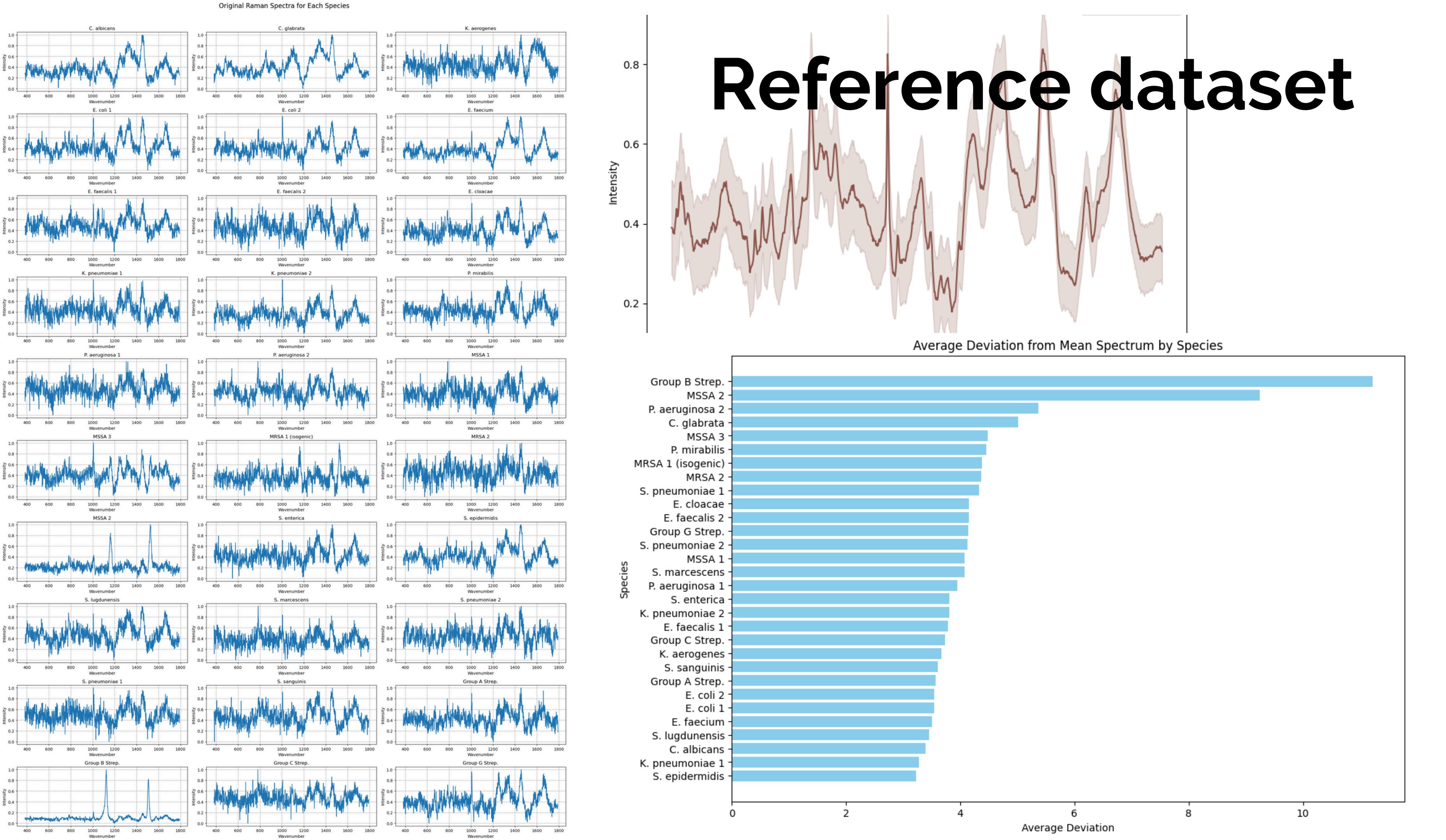
Fine-tuning dataset

30 especies *100 spetcra = 60000 spectra

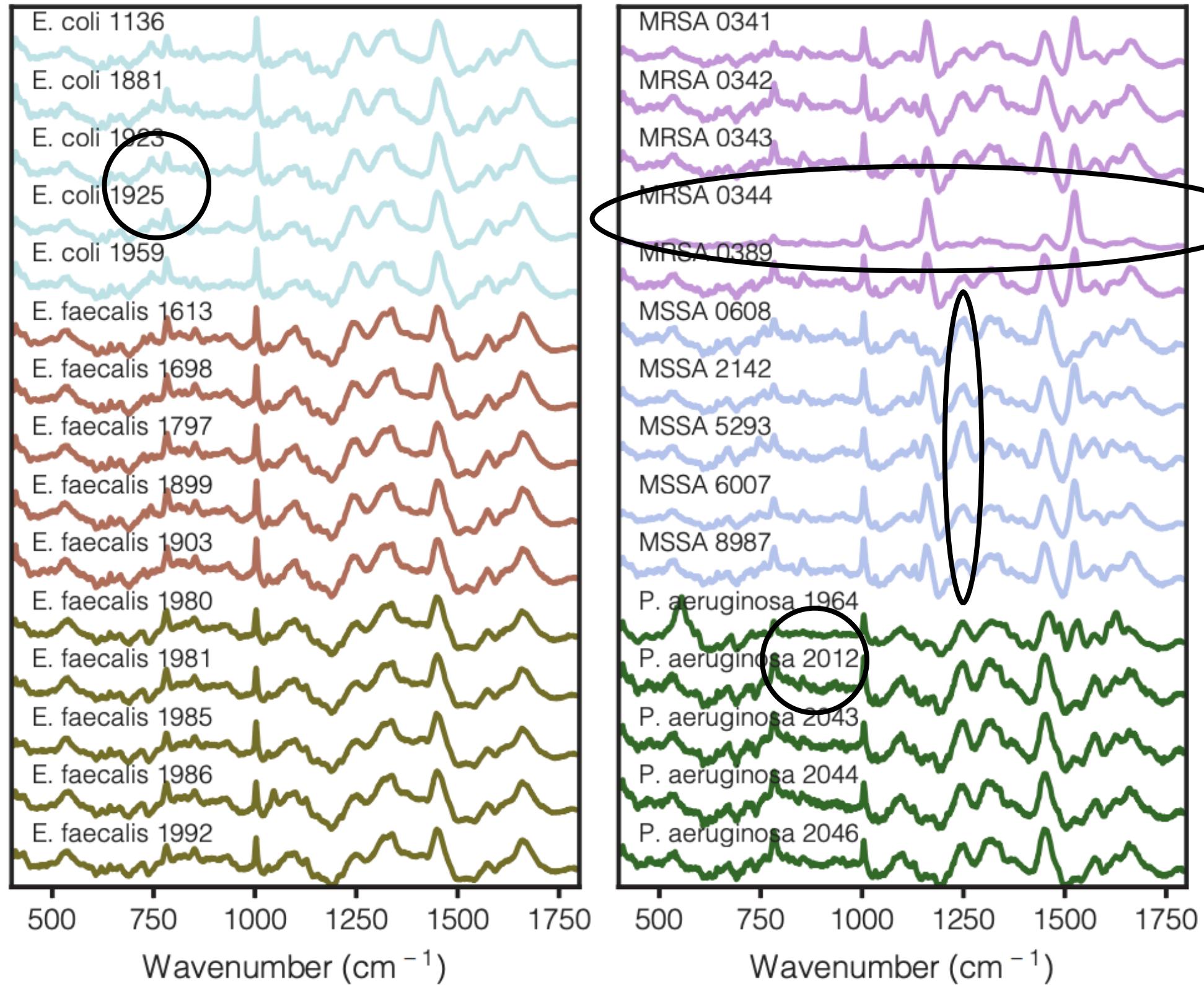
Clinical datasets

5 species*400 spectra, taken from 30 patients

Nat Commun 10, 4927 (2019)



Espectros de aislamientos individuales de pacientes, promediados en todo el conjunto de 400 espectros para cada paciente.



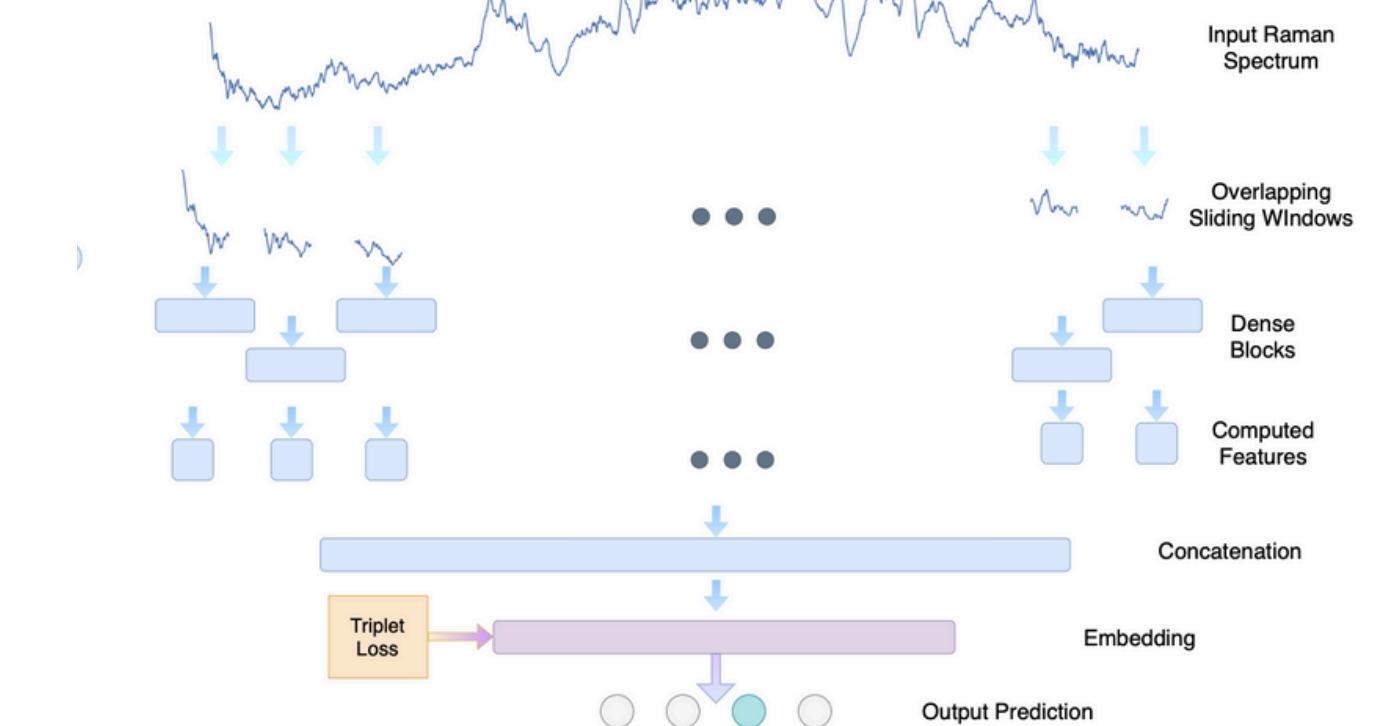
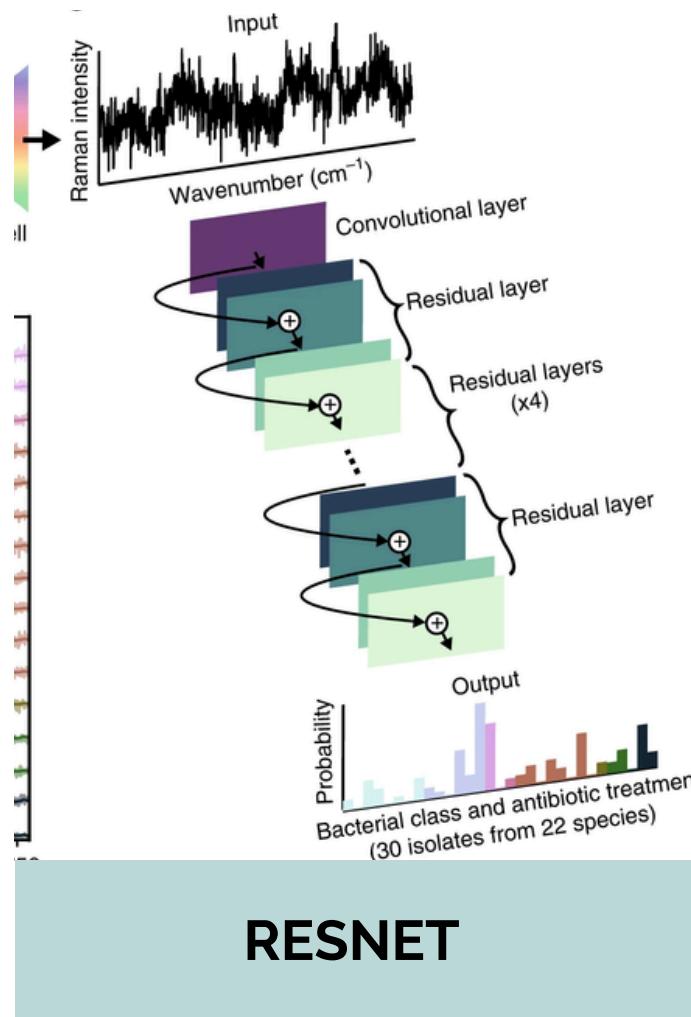
Intensidad

Algunos de los espectros claramente tienen mejor relación señal-ruido (SNR)

Inconsistencia en los picos

La comparación entre aislamientos de diferentes pacientes muestra una alta variabilidad en el patrón de intensidad Raman

RESNET VS CNN+MLP



RESNET

CNN

RESNET VS CNN+MLP

Librería utilizada

PyTorch TensorFlow/Keras

Bloques básicos

Bloques Residuales con capas convolucionales 1D y

BatchNorm

Capas densas con BatchNorm, LeakyReLU y Dropout, con

normalización L2

Mecanismo de regularización

N/A

Dropout en capas intermedias

Normalización

Batch Normalization en varias capas

Batch Normalization y Lambda para normalización L2

RESNET

CNN

RESNET

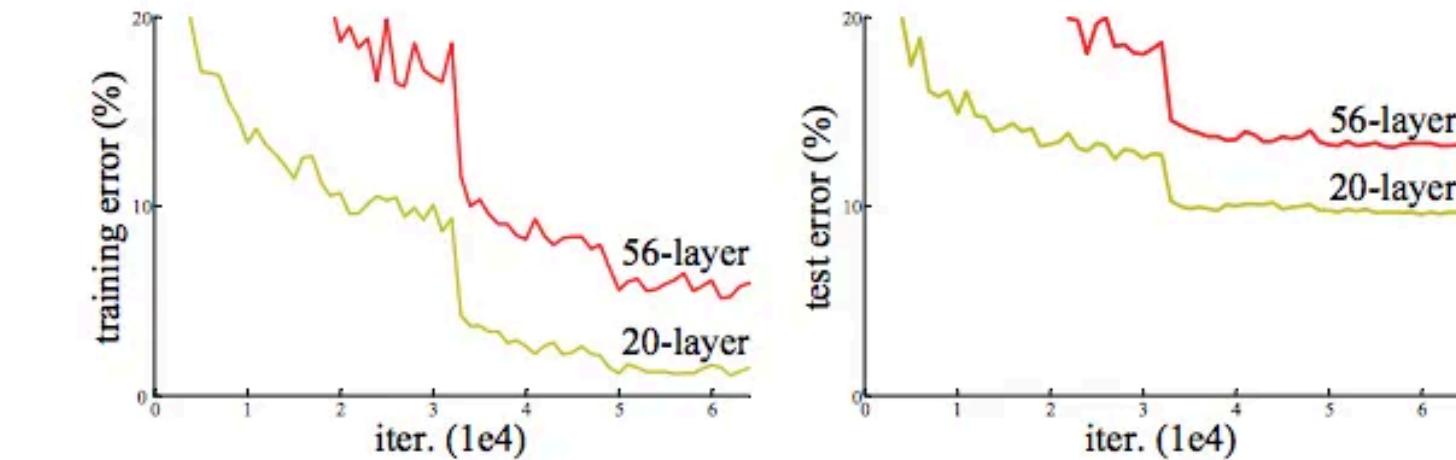
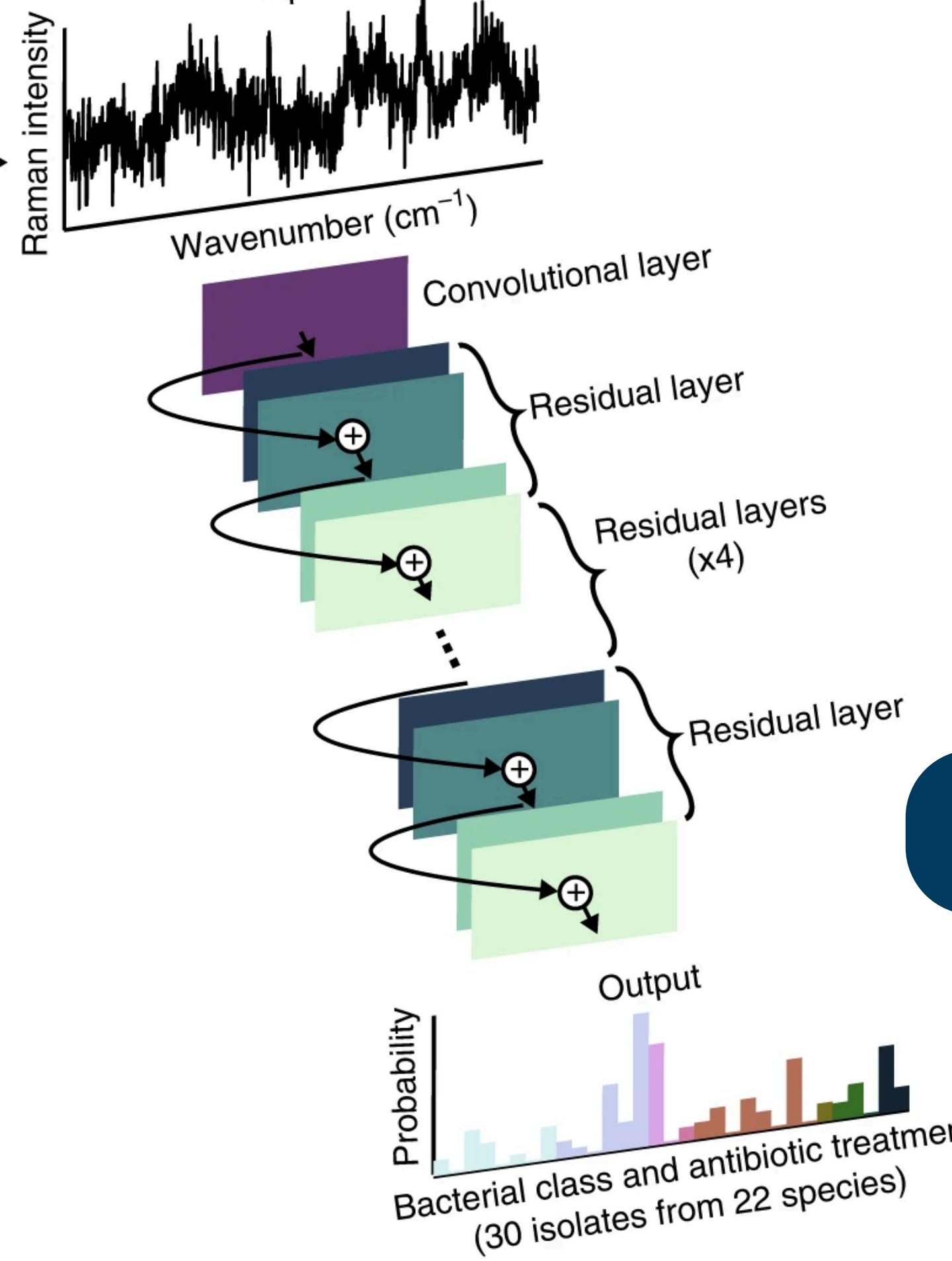
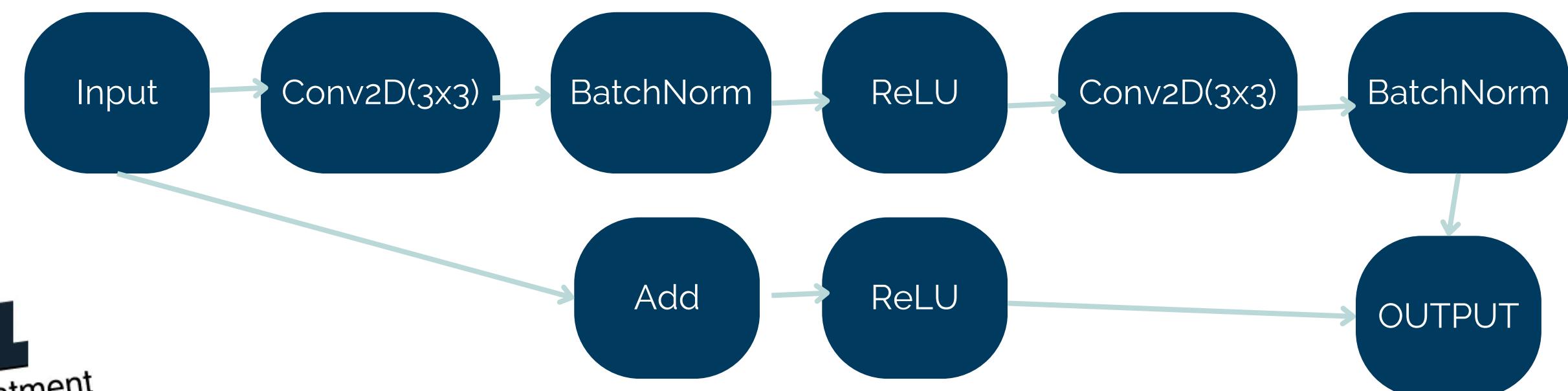
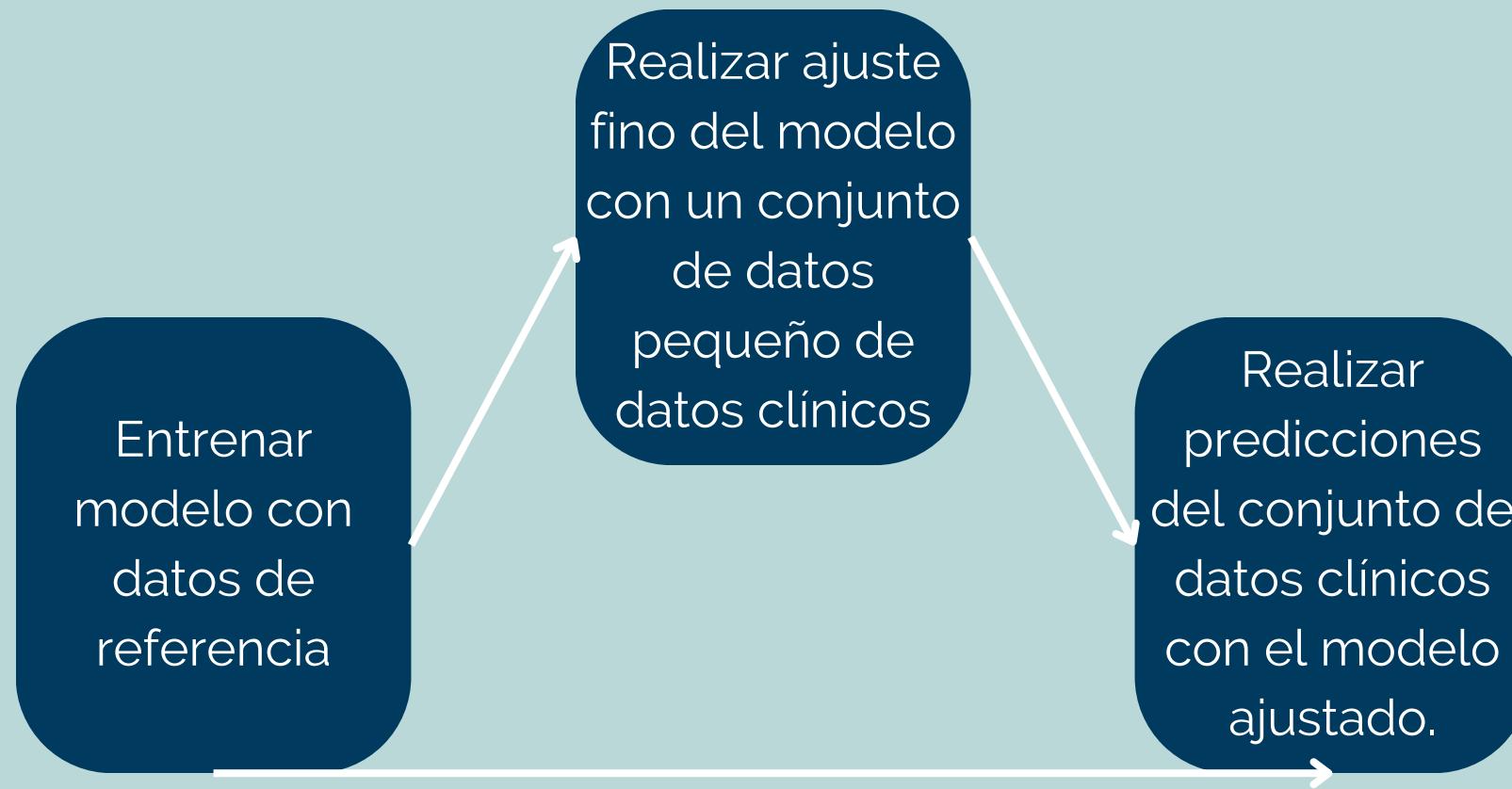


Figure 4. The graph shows that the error rate in a 20-layer is lesser than that of 56-layered network, and theoretically, it should be the opposite. (Image Source: (Original Citation) Deep Residual Learning for Image Recognition)



La estrategia original



The trained CNN classifies 30 bacterial and yeast isolates with isolate-level accuracy of $81.6 \pm 0.6\%$

a)

True	Predicted				
	MRSA 1	MRSA 2	MSSA 1	MSSA 2	MSSA 3
MRSA 1	82			17	
MRSA 2		86		4	4
MSSA 1		2	82		3 1
MSSA 2	1			98 1	
MSSA 3	22			1 75	
<i>S. epidermidis</i>				100	
<i>S. lugdunensis</i>	10	1			86
<i>S. pneumoniae</i> 1	4				
<i>S. pneumoniae</i> 2				70 18	2
Group A Strep.				6 91	2
Group B Strep.					2 95
Group C Strep.					100
Group G Strep.					98
<i>S. sanguinis</i>			1 1		2
<i>E. faecalis</i> 1			5	75 1	16
<i>E. faecalis</i> 2				2 82	12
<i>E. faecium</i>				6 7	79
<i>E. coli</i> 1					100
<i>E. coli</i> 2					
<i>K. pneumoniae</i> 1	1				
<i>K. pneumoniae</i> 2					2
<i>K. aerogenes</i>	3				
<i>E. cloacae</i>	1				
<i>P. mirabilis</i>				1 6	
<i>S. marcescens</i>			1		4
<i>S. enterica</i>					2
<i>P. aeruginosa</i> 1	2				
<i>P. aeruginosa</i> 2	5			3 1	
<i>C. albicans</i>					
<i>C. glabrata</i>					

b)

True	Predicted				
	Vancomycin	Ceftriaxone	Penicillin	Daptomycin	Meropenem
Vancomycin	97		1		
Ceftriaxone	2	93	3		2
Penicillin			97		1
Daptomycin				100	
Meropenem	1		2	95	
Ciprofloxacin			2	1	96
TZP	4	3		7	86
Caspofungin					100
<i>E. coli</i> 1					
<i>E. coli</i> 2					
<i>K. pneumoniae</i> 1					
<i>K. pneumoniae</i> 2					
<i>K. aerogenes</i>					
<i>E. cloacae</i>					
<i>P. mirabilis</i>					
<i>S. marcescens</i>					
<i>S. enterica</i>					
<i>P. aeruginosa</i> 1					
<i>P. aeruginosa</i> 2					
<i>C. albicans</i>					
<i>C. glabrata</i>					

data from Nat Commun 10, 4927 (2019)

Época	Tiempo (s)	Precisión de Entrenamiento (%)	Pérdida de Entrenamiento	Precisión de Validación (%)	Pérdida de Validación	Comentarios
1	0.02	83.66	0.0500	82.35	0.0748	Buen comienzo, alta precisión desde la primera época.
2	1757.28	91.59	0.0242	90.18	0.0485	Mejoras significativas tanto en precisión como en pérdida.
3	3480.82	92.94	0.0199	89.63	0.0616	Precisión estable, ligera fluctuación en la pérdida de validación.
4	5256.64	93.82	0.0171	93.10	0.0195	Notable mejora en la precisión de validación.
5	7027.02	94.49	0.0153	90.63	0.0342	Ligera caída en la precisión de validación, pero la pérdida sigue siendo baja.
6	8968.10	95.05	0.0139	90.55	0.1389	Aumento en la pérdida de validación, posible sobreajuste.
7	10893.23	95.36	0.0127	91.03	0.0564	Recuperación en la precisión de validación.
8	12754.06	95.80	0.0115	89.78	0.1342	Nueva caída en la precisión de validación y alta pérdida.
9	14844.24	96.07	0.0108	92.15	0.0298	Buen cierre, mejora en la precisión de validación, pero con fluctuaciones previas.

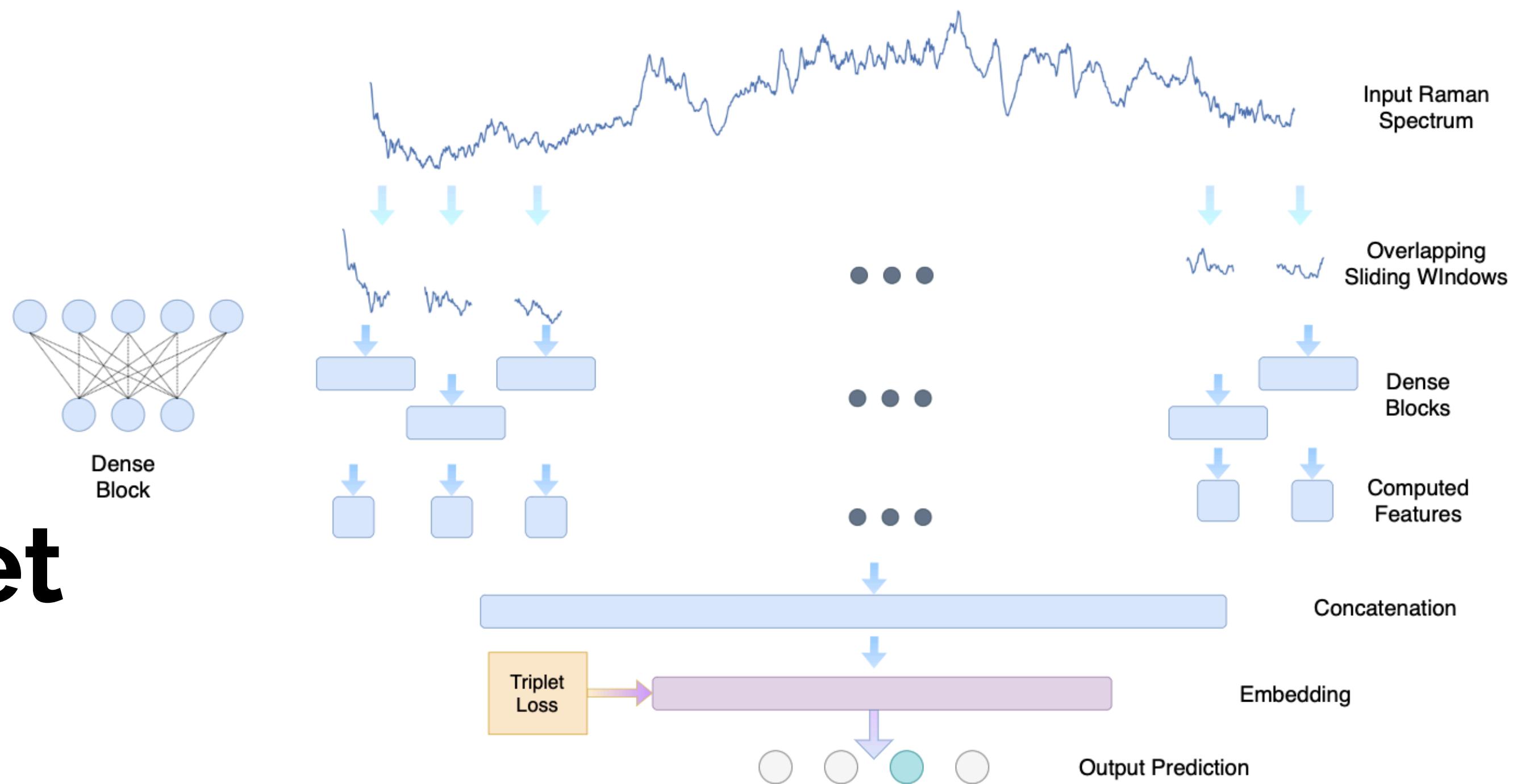
El entrenamiento se detiene si la precisión de validación no mejora en un número consecutivo de épocas (Entrenamiento approx 4 horas)

RESNET

Accuracy on the test dataset: 54.90%

Confusion Matrix

CNN RamanNet



Arquitectura Densa Inspirada en Convoluciones

capas densas en lugar de convoluciones tradicionales para procesar el espectro. Las ventanas deslizantes del espectro se pasan a bloques densos, lo que emula la operación de convolución y minimiza el riesgo de sobreajuste

Incorporación de Pérdida de Triplete

Para mejorar la separación entre clases en un espacio de características, especialmente debido al bajo ratio señal-ruido en los espectros Raman, RamanNet introduce la pérdida de triplete como una pérdida auxiliar.

Regularización

El modelo aplica técnicas de regularización como dropout y normalización por lotes en múltiples capas, utilizando la función de activación LeakyReLU, lo que mejora la capacidad del modelo para generalizar y evitar el sobreajuste

CNN

El modelo mejora significativamente en las primeras 50 épocas, alcanzando una precisión de validación cercana al 95%. (Tiempo total 1.96h)

Épocas	Tiempo (s)	Train Loss	Val Loss	Train Acc	Val Acc	Comentarios
1-10	713	0.5124 - 1.0122	0.4522 - 0.7809	67.24% - 92.30%	77.62% - 93.28%	Mejoras significativas en la precisión y la pérdida. El modelo se ajusta rápidamente.
11-20	737	0.4677 - 0.5072	0.4169 - 0.4423	91.88% - 93.88%	91.28% - 94.52%	El modelo sigue mejorando, aunque la tasa de mejora se ralentiza.
21-30	765	0.4463 - 0.4587	0.3912 - 0.4290	93.28% - 94.88%	93.17% - 95.23%	Se observa una pequeña fluctuación en las métricas, pero el rendimiento sigue mejorando.
31-40	738	0.4375 - 0.4426	0.3793 - 0.4011	93.70% - 95.20%	93.83% - 95.60%	La mejora del rendimiento es más lenta. Algunos aumentos menores en la precisión.
41-50	740	0.4262 - 0.4374	0.3919 - 0.4030	93.70% - 95.60%	94.40% - 95.23%	Comienza la estabilización de las métricas, con pequeños ajustes en la pérdida.
51-60	748	0.4215 - 0.4274	0.3854 - 0.3990	94.40% - 95.70%	94.47% - 95.25%	Se alcanzan las mejores métricas hasta ahora, pero hay poca mejora adicional.
61-70	740	0.4139 - 0.4217	0.3806 - 0.3930	94.55% - 95.80%	94.55% - 95.65%	El modelo se estabiliza cerca de su rendimiento máximo.
71-80	751	0.4099 - 0.4139	0.3761 - 0.3888	94.50% - 95.85%	94.70% - 95.95%	La mejora es marginal y se aplican ajustes menores al modelo.
81-100	751	0.4029 - 0.4103	0.3748 - 0.3900	94.53% - 95.89%	94.53% - 95.97%	Se observa una pequeña mejora en la pérdida, pero las métricas de precisión se estabilizan.

RESNET

Confusion Matrix

		Predicted Labels																				
		Predicted Labels																				
True Labels		Predicted Labels																				
		C. albicans -	C. glabrata -	K. aerogenes -	E. coli 1 -	E. coli 2 -	E. faecium -	E. faecalis 1 -	E. faecalis 2 -	E. cloacae -	K. pneumoniae 1 -	K. pneumoniae 2 -	P. mirabilis -	P. aeruginosa 1 -	P. aeruginosa 2 -	MSSA 1 -	MSSA 3 -	MRSA 1 (isogenic) -	MRSA 2 -	MSSA 2 -	S. enterica -	
C. albicans -	83	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	3
C. glabrata -	39	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
K. aerogenes -	0	0	8	3	2	0	14	12	1	0	1	0	1	0	38	0	1	13	0	0	3	0
E. coli 1 -	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E. coli 2 -	0	0	0	4	83	0	0	1	9	0	0	0	0	0	0	0	0	0	1	0	0	2
E. faecium -	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E. faecalis 1 -	0	0	0	0	0	19	60	10	0	0	0	0	0	0	2	0	0	3	0	0	0	5
E. faecalis 2 -	0	0	0	0	0	1	20	58	0	0	0	0	0	0	1	0	1	12	0	1	0	0
E. cloacae -	0	0	0	0	9	0	12	32	3	0	1	0	0	0	21	0	0	16	0	3	0	0
K. pneumoniae 1 -	0	0	1	3	16	0	1	7	34	0	0	0	1	0	6	0	0	3	0	20	0	4
K. pneumoniae 2 -	0	0	0	2	2	0	2	4	0	0	73	0	0	0	10	0	0	3	0	2	0	0
P. mirabilis -	0	0	0	0	3	0	6	2	2	0	0	4	0	0	41	0	0	35	0	2	0	2
P. aeruginosa 1 -	0	0	0	0	0	0	0	0	0	0	0	50	0	0	12	0	0	28	0	0	0	4
P. aeruginosa 2 -	0	0	0	0	0	0	0	0	0	0	0	67	0	1	0	0	26	0	0	0	0	0
MSSA 1 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
MSSA 3 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	4	0	1	0	0	0	0
MRSA 1 (isogenic) -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	12	0	22	0	0	0	0
MRSA 2 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	97	0	0	0	0	0
MSSA 2 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	91	0	0	0	0
S. enterica -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
S. epidermidis -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	1	0	0	0
S. lugdunensis -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	83	0	0	0	9	0	0
S. marcescens -	0	0	0	0	1	0	9	26	12	0	0	0	1	0	6	0	0	37	0	3	0	2
S. pneumoniae 2 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S. pneumoniae 1 -	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	5	0	0	0	0
S. sanguinis -	0	0	0	0	0	0	4	38	1	0	0	0	0	0	1	0	0	0	0	7	0	1
Group A Strep. -	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	17	0	0	1	0	0	0
Group B Strep. -	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	97	0
Group C Strep. -	0	0	0	0	0	0	7	4	0	0	0	0	0	0	9	0	0	4	0	0	20	1
Group G Strep. -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	7	0	81

CNN

Confusion Matrix

RESNET VS CNN+MLP

Duración del Entrenamiento:

9 epochs alrededor de 4.66 horas 100 epochs en aproximadamente 1.96 horas

Progresión de la Precisión:

La precisión de entrenamiento aumentó de 83.66% a más de 96%, pero la precisión de validación mostró fluctuaciones.

Comenzó con una precisión de 67% y alcanzó hasta 95%, con mejoras más consistentes en la precisión de validación.

Reducción de la Pérdida:

La pérdida de entrenamiento disminuyó de manera constante, pero la pérdida de validación fluctuó, mostrando posibles señales de sobreajuste.

La pérdida se redujo de forma continua, con un enfoque más dinámico en el entrenamiento.

Rendimiento General

Accuracy on the test dataset: 54.90%

Accuracy on the test dataset: 44.90%

RESNET

CNN



El espectro de entrada se divide en ventanas deslizantes superpuestas de longitud w con un tamaño de paso de dw .

w	dw	Accuracy on Test Dataset (%)
50	25	44.90
100	25	45.03
10	5	46.80

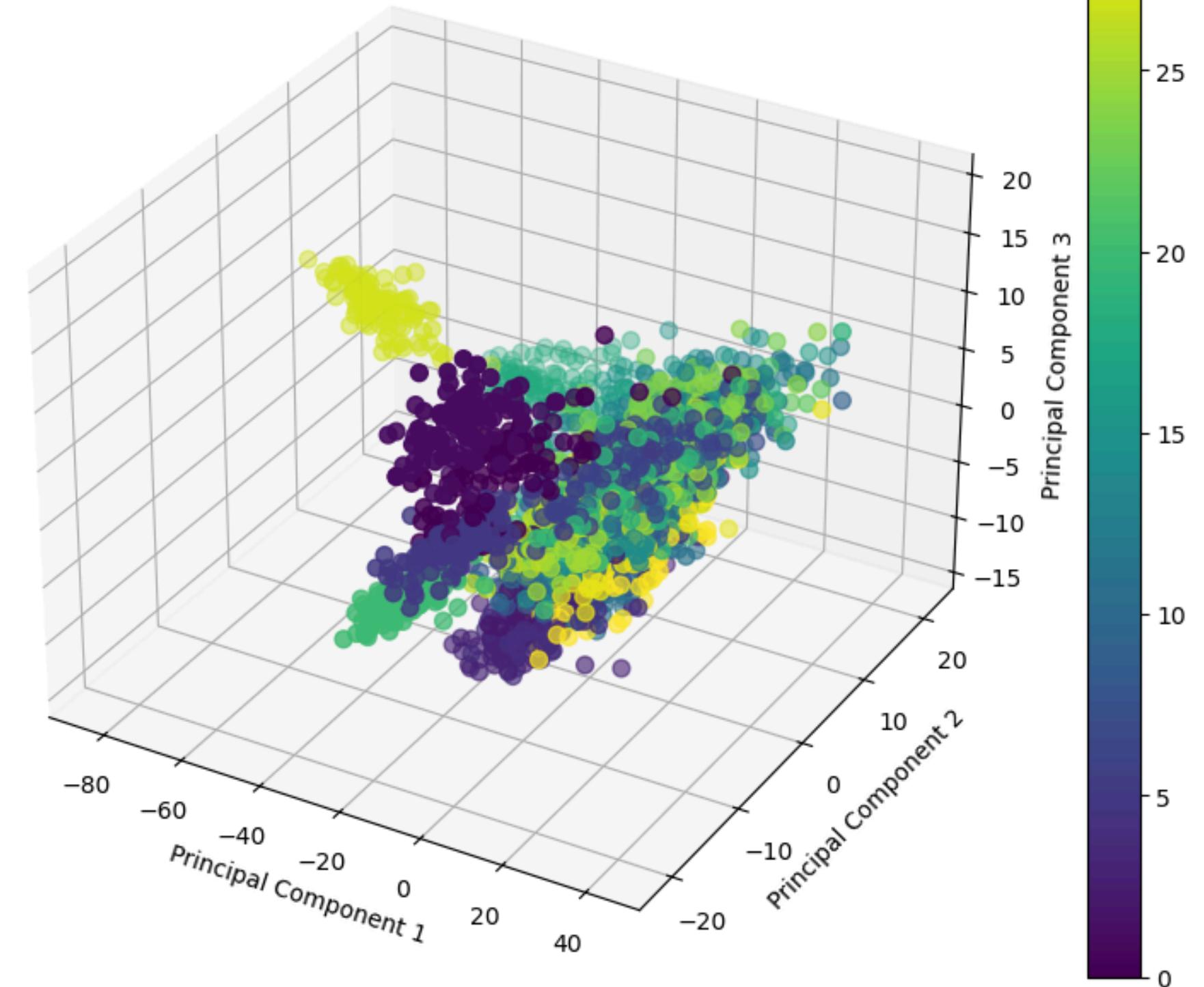
Accuracy on the test dataset ResNET model: 54.90%

Precisión de diferentes métodos en el conjunto de datos de prueba

Método	Precisión en el Conjunto de Datos de Prueba (%)
kNN	30
LDA	93
SVM	92
RandForest	65
PLSDA	58

30%
explained by first three components

PCA of Raman Spectra - First Three Components



Thank You!

For your listening...