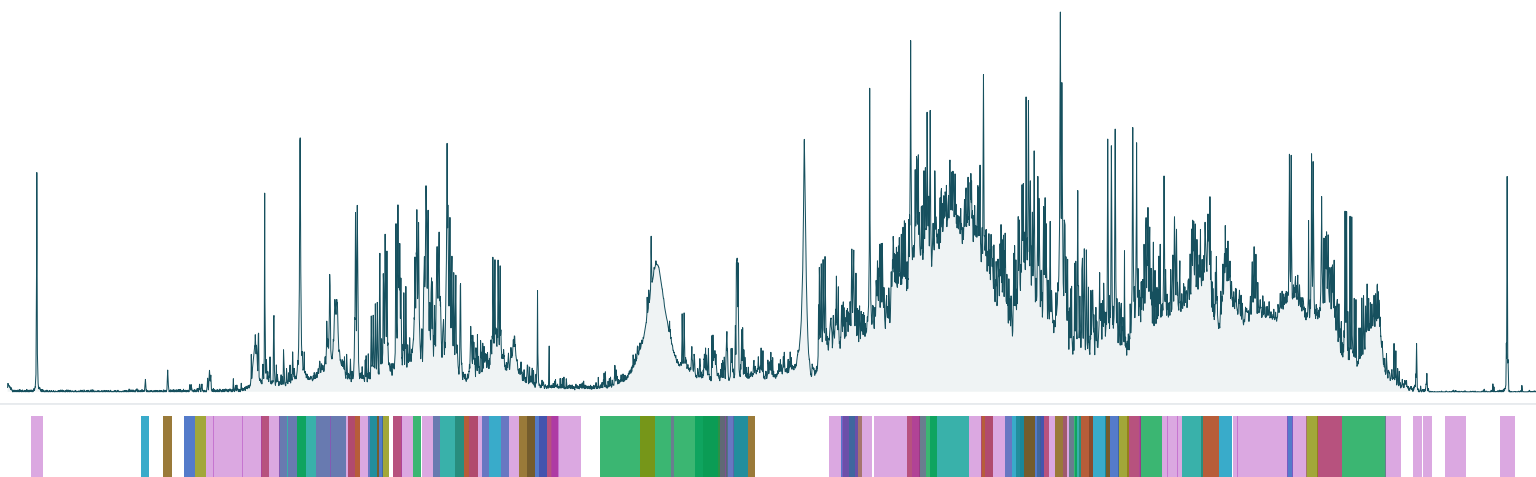




Fundamento químico y matemático

Qué hace TIAGO con un espectro, paso a paso: de la señal medida al área atribuida a cada metabolito



— Espectro de ^1H -RMN de orina de la cohorte de referencia. Debajo, las 103 ventanas de integración del catálogo; el magenta señala las regiones sin metabolitos asignados.



Índice

1. El problema	2
1.1. Por qué el área importa	2
1.2. Por qué en orina no basta con integrar	2
2. Vocabulario y notación	2
3. Paso 1 · Del archivo a la escala química	3
4. Paso 2 · Estimar el ruido	3
5. Paso 3 · La línea base	3
5.1. Qué es y qué no	3
5.2. Dos regímenes distintos	4
5.3. Anclaje y familias	4
6. Paso 4 · El modelo de una señal	4
6.1. Forma de línea	4
6.2. Multiplicidad	5
6.3. Restricciones de ventana	5
7. Paso 5 · Ajuste conjunto y deconvolución	6
8. Paso 6 · La partición exclusiva	6
8.1. El problema	7
8.2. Proyección del agregado dentro de la señal	7
8.3. División del agregado entre componentes	8
8.4. El residual	8
9. Paso 7 · Curva de integración y curva dibujada	9
10. Paso 8 · Estados de calidad	9
11. Determinismo	10
12. Del área a la concentración	11
13. Resumen del flujo	11

1 El problema

— 1.1 Por qué el área importa

En ^1H -RMN cuantitativa, la integral de una señal es proporcional al número de núcleos que la producen. Para una señal s de un metabolito m :

$$A_s = k n_s c_m, \quad (1)$$

donde c_m es la concentración molar del metabolito, n_s el número de protones equivalentes que originan esa señal y k una constante instrumental común a todo el espectro bajo condiciones cuantitativas. Esa proporcionalidad, y no la altura del pico, es lo que convierte un espectro en una medida: la altura depende del ancho de línea, que a su vez depende del pH, de la viscosidad y del shimming; el área no.

— 1.2 Por qué en orina no basta con integrar

La orina es una matriz con centenares de metabolitos. Su espectro de protón, comprimido en unos 10 ppm, presenta tres dificultades que hacen que “integrar el pico” no sea una operación bien definida:

Solapamiento Dos o más metabolitos comparten el mismo intervalo de desplazamiento químico. La integral del intervalo no pertenece a ninguno de los dos.

Deriva de la línea base Bajo las señales hay una contribución ancha —macromoléculas, imperfecciones de fase, colas de señales intensas vecinas— que no es cero y no es constante.

Desplazamiento dependiente de la matriz El centro de una señal se mueve con el pH, la fuerza iónica, la temperatura y la complejación. Un centro tabulado es un valor a priori, no una posición garantizada.

La pregunta que TIAGO responde

Dada una ventana del espectro y una lista de metabolitos que se sabe que la ocupan, **¿qué parte del área observada corresponde a cada uno?** La respuesta debe ser un reparto *exclusivo* —nada se cuenta dos veces— y *completo* —nada se pierde—.

2 Vocabulario y notación

Sea una región de interés (ROI) definida por un intervalo $[\delta_{\min}, \delta_{\max}]$ de desplazamiento químico. Sobre la rejilla discreta $\{\delta_i\}_{i=1}^N$ de ese intervalo:

Símbolo	Nombre	Definición
y_i	señal medida	Intensidad del espectro en δ_i
b_i	línea base	Contribución no atribuible a las componentes
y_i	señal corregida	$y_i = \max(y_i - b_i, 0)$
$M_{k,i}$	componente k	Perfil ajustado del metabolito k
r_i	residual	Área no explicada por ninguna componente
σ	ruido	Desviación robusta del ruido local
K	número de metabolitos declarados activos en la región	

El área se calcula siempre por la regla del trapecio sobre la rejilla real, no por suma de alturas:

$$A[f] = \left| \int_{\delta_{\min}}^{\delta_{\max}} f(\delta) d\delta \right| \approx \left| \sum_{i=1}^{N-1} \frac{f_i + f_{i+1}}{2} (\delta_{i+1} - \delta_i) \right|. \quad (2)$$

3 Paso 1 · Del archivo a la escala química

El espectro llega con el eje en hercios. El desplazamiento químico se obtiene dividiendo por la frecuencia del espectrómetro:

$$\delta [\text{ppm}] = \frac{\nu [\text{Hz}]}{\nu_0 [\text{MHz}]}, \quad \nu_0 = 600,132\,817\,8 \text{ MHz}. \quad (3)$$

Esta división es lo que hace comparables espectros medidos en equipos de campo distinto: en ppm, la posición de una señal es una propiedad del compuesto y de su entorno químico, no del imán.

La constante de acoplamiento, en cambio, **no** se escala con el campo: es una propiedad de la molécula. Por eso, al construir un multiplete, una separación J en hercios se convierte a ppm con la misma división:

$$\Delta\delta = \frac{J [\text{Hz}]}{\nu_0}. \quad (4)$$

4 Paso 2 · Estimar el ruido

Antes de decidir qué es señal hace falta saber cuánto vale “nada”. TIAGO estima el ruido por la desviación absoluta mediana de las *diferencias* entre puntos consecutivos:

$$\sigma = \frac{1,4826 \cdot \text{MAD}(\Delta y)}{\sqrt{2}}, \quad \Delta y_i = y_{i+1} - y_i. \quad (5)$$

Tres decisiones están condensadas ahí:

- Trabajar sobre **diferencias** elimina cualquier tendencia suave: una línea base inclinada no infla la estimación.
- La **mediana absoluta** en lugar de la desviación típica hace el estimador robusto: un pico intenso dentro de la ventana no lo contamina.
- El factor 1,4826 convierte la MAD en un equivalente de desviación típica bajo normalidad, y el $\sqrt{2}$ corrige que la diferencia de dos variables independientes tiene el doble de varianza.

5 Paso 3 · La línea base

— 5.1 Qué es y qué no

La línea base es la contribución de la ventana que no se atribuye a ninguna componente positiva. Debe ser lo bastante flexible para seguir la deriva real y lo bastante rígida para **no absorber señal**: una base con demasiados grados de libertad puede pasar por debajo de un pico y hacerlo desaparecer.

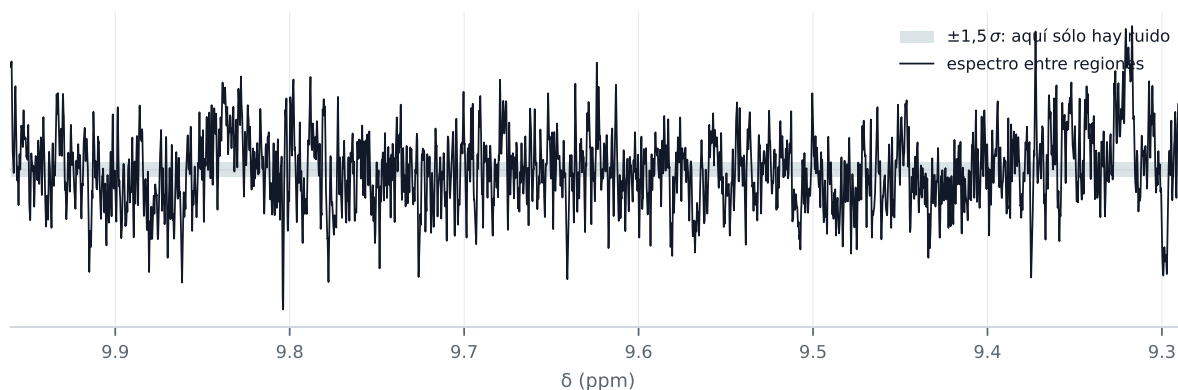


Figura 1: El estimador de ruido, medido donde de verdad sólo hay ruido: un hueco del espectro entre dos regiones del catálogo. La banda marca $\pm 1,5\sigma$, que es el umbral con el que se decide qué punto puede servir de anclaje para la línea base. Dentro de una ventana SET no queda ningún punto por debajo de ese nivel, porque las regiones están dibujadas ceñidas a la señal.

— 5.2 Dos regímenes distintos

Regiones SET Se estima una línea base local *común a todas las componentes de la región*, apoyada en los tramos libres de señal.

Regiones UNK, DSS y ERETIC La línea base es **fija en cero**. Se integra toda la señal positiva de la ventana sin modelo paramétrico. Aquí no se pretende identificar nada: se mide cuánta área hay.

— 5.3 Anclaje y familias

Los puntos de anclaje son aquellos donde la señal corregida no supera $1,5\sigma$, es decir, donde estadísticamente sólo hay ruido. Sobre esos anclajes se ajusta una de varias familias —horizontal en cero, lineal, lineal segmentada, dos mínimos horizontales, polinómica o interpolación monótona— y se elige por auditoría, no por menor error cuadrático.

Tras ajustar, se aplica una referencia a cero: se resta la mediana de la señal corregida en la máscara de base, de modo que el nivel de “nada” quede exactamente en cero.

$$y_i = \max(y_i - b_i - \text{med}(y - b)|_{\text{base}}, 0). \quad (6)$$

El recorte a cero es deliberado: el área negativa no tiene sentido físico como contribución de un metabolito.

6 Paso 4 · El modelo de una señal

— 6.1 Forma de línea

Una señal de RMN en un líquido, con relajación exponencial, produce una lorentziana:

$$L(\delta; a, \delta_0, w) = a \frac{\gamma^2}{(\delta - \delta_0)^2 + \gamma^2}, \quad \gamma = \frac{w}{2}, \quad (7)$$

con a la altura, δ_0 el centro y w la anchura a media altura (FWHM). En la práctica, la inhomogeneidad del campo añade un ensanchamiento de carácter gaussiano:

$$G(\delta; a, \delta_0, w) = a \exp\left[-\frac{1}{2}\left(\frac{\delta - \delta_0}{\sigma_G}\right)^2\right], \quad \sigma_G = \frac{w}{2\sqrt{2\ln 2}}. \quad (8)$$

TIAGO admite ambas y su mezcla, la pseudo-Voigt, que interpola con un solo parámetro $\eta \in [0, 1]$:

$$V_p(\delta) = \eta L(\delta) + (1 - \eta) G(\delta). \quad (9)$$

Cuando hace falta el perfil exacto, se usa la Voigt verdadera —la convolución de ambas— evaluada mediante la función de error complementaria escalada $w(z)$ de Faddeeva:

$$V(\delta) \propto \Re \left[w \left(\frac{(\delta - \delta_0) + i\gamma}{\sigma_G \sqrt{2}} \right) \right]. \quad (10)$$

Por qué importa la elección

Las colas de una lorentziana decaen como $(\delta - \delta_0)^{-2}$; las de una gaussiana, exponencialmente. Sobre un intervalo finito eso cambia el área de forma apreciable. Elegir la familia por el menor error cuadrático no basta: hay que considerar también la identidad, la multiplicidad, la anchura, la estabilidad del AUC y que el residuo no conserve estructura de pico.

— 6.2 Multiplicidad

El acoplamiento escalar desdobra una señal en un patrón cuyas posiciones relativas e intensidades están fijadas por el número de vecinos. TIAGO compone el multiplete como suma de líneas con centros derivados del centro global y de J según (4):

$$M(\delta) = \sum_p c_p L(\delta; a, \delta_0 + o_p \Delta\delta, w), \quad (11)$$

con los desplazamientos o_p y los pesos c_p del triángulo de Pascal:

Multiplicidad	Líneas	Desplazamientos o_p	Pesos c_p
Singlete	1	0	1
Doblete	2	$-\frac{1}{2}, +\frac{1}{2}$	1:1
Triplete	3	$-1, 0, +1$	1:2:1
Cuartete	4	$-\frac{3}{2}, -\frac{1}{2}, +\frac{1}{2}, +\frac{3}{2}$	1:3:3:1

Cuadro 1: Patrones de desdoblamiento. Todas las líneas de un multiplete comparten amplitud base, anchura y J : son la misma señal, no señales independientes.

Esta restricción es la clave de la robustez frente al solapamiento. Un doblete no se ajusta como dos picos libres, sino como un objeto con cuatro parámetros —amplitud, centro, anchura y J — cuya forma está determinada por la química. Un vecino que caiga entre sus dos ramas no puede ser absorbido sin romper el patrón.

— 6.3 Restricciones de ventana

El centro ajustado debe permanecer dentro de una ventana química validada alrededor del valor tabulado, y J dentro de un intervalo también validado. La anchura se acota al rango instrumental esperado, típicamente 0,8 Hz a 1,5 Hz para esta cohorte.

La regla que evita el error más frecuente

Un modelo **no debe saltar al pico más intenso de la región** si ese pico no corresponde al metabolito. Optimizar el residuo sin restringir el centro lleva justo a ese fallo: el ajuste mejora y la identidad se pierde.

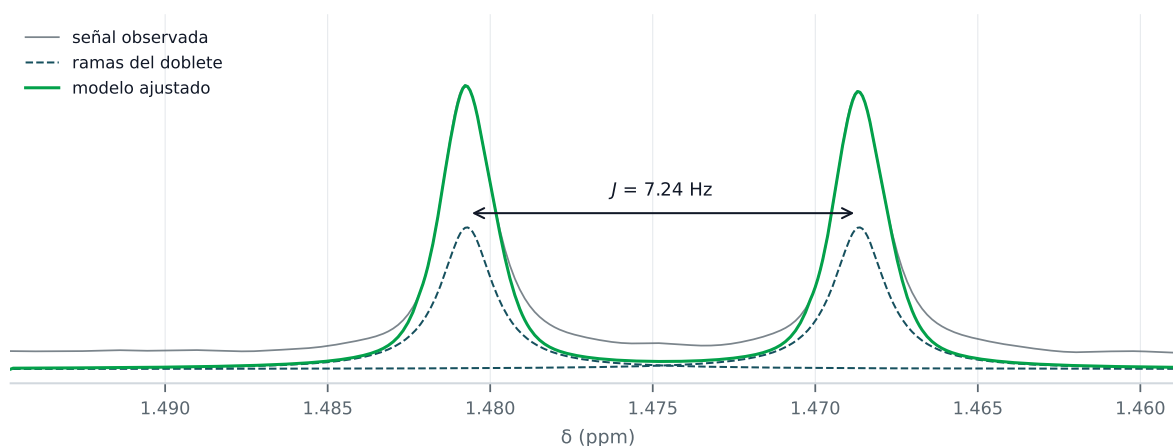


Figura 2: El doblete de la alanina en SET_4, ajustado sobre un espectro real. Las dos ramas discontinuas no son picos independientes: comparten amplitud, anchura y separación, y su suma es el modelo. Esa ligadura es lo que impide que una señal vecina sea absorbida sin romper el patrón.

7 Paso 5 · Ajuste conjunto y deconvolución

Cuando K metabolitos comparten la ventana, no se ajusta uno y se resta. Se ajustan todos a la vez, minimizando el residuo del conjunto:

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \left[y_i - \sum_{k=1}^K M_k(\delta_i; \theta_k) \right]^2, \quad (12)$$

resuelto por mínimos cuadrados no lineales con restricciones de caja sobre cada parámetro. El área individual de cada metabolito es entonces la integral de su propia componente:

$$A_k = A[M_k]. \quad (13)$$

Lo que está prohibido, y por qué

- **Repartir por valles o intersecciones.** Un mínimo entre dos señales no marca dónde acaba una y empieza otra: ambas contribuyen a ambos lados.
- **Restar áreas geométricas sin modelo.** Equivale a suponer que la señal se corta verticalmente, lo que ninguna forma de línea hace.
- **Copiar la señal observada como AUC.** Incorpora ruido, hombros ajenos y solapamiento al número que se va a publicar.

Si el reparto no es identificable, la respuesta correcta es marcarlo como tal y reportar sólo el área combinada.

8 Paso 6 · La partición exclusiva

Aquí está el núcleo del método, y es donde TIAGO se aparta de un ajuste convencional.

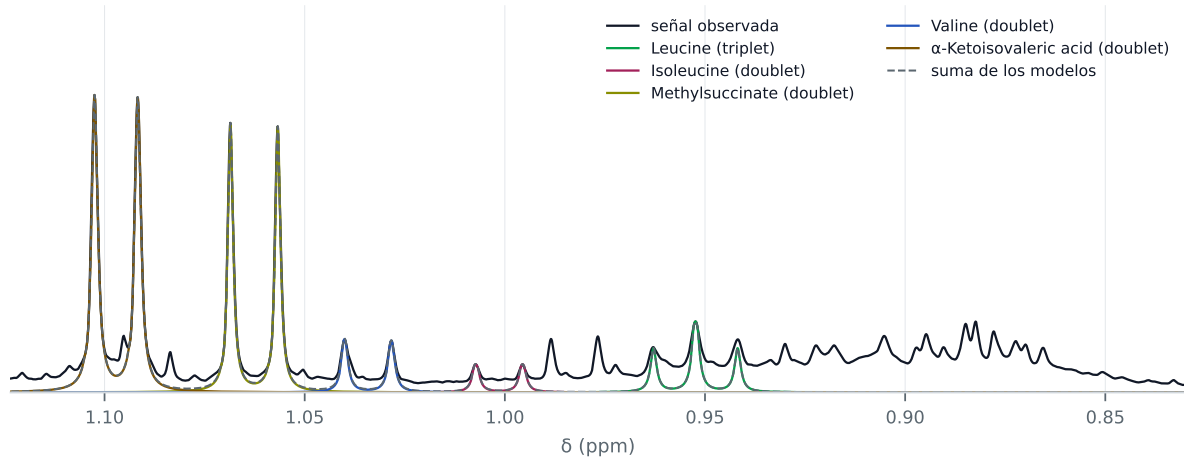


Figura 3: Los cinco metabolitos de SET_1 ajustados a la vez sobre la señal observada. Cada color es una componente; la línea discontinua es su suma. Se ajustan conjuntamente y no uno detrás de otro: ajustar y restar haría que el error del primero se propagara al resto.

— 8.1 El problema

Los modelos ajustados son perfiles matemáticos de soporte infinito. Nada garantiza que su suma quede por debajo de la señal observada en todos los puntos: por acumulación de colas, la suma puede superarla localmente. Si eso ocurre, dibujar las áreas apiladas produciría una figura que miente, y sumarlas daría más área de la que hay.

Se busca por tanto una partición $\{P_k\}$ del área de la región que cumpla tres condiciones simultáneamente:

$$(i) \text{ Exclusividad y capacidad: } \sum_{k=1}^K P_{k,i} \leq y_i \quad \forall i, \quad (14)$$

$$(ii) \text{ Conservación del AUC: } A[P_k] = A_k \quad \forall k, \quad (15)$$

$$(iii) \text{ Cierre: } A[y] = \sum_k A[P_k] + A[r]. \quad (16)$$

— 8.2 Proyección del agregado dentro de la señal

Primero se lleva la *suma* de los modelos dentro de la capacidad física de la región. Sea $S_i = \sum_k M_{k,i}$ la suma original y $T = \sum_k A_k$ el área total a repartir.

Se define el recorte punto a punto y la capacidad libre que queda:

$$C_i = \min(S_i, y_i), \quad Q_i = y_i - C_i \geq 0. \quad (17)$$

Si $A[C] < T$, falta área por colocar: el déficit $D = T - A[C]$ se distribuye sobre la capacidad libre Q usando la forma original como peso. Con $\omega_i = \max(S_i / \max_j S_j, \varepsilon)$, se busca el escalar λ tal que

$$A[\min(Q, \lambda \omega)] = D, \quad (18)$$

resuelto por bisección. La función del lado izquierdo es monótona no decreciente y continua en λ , de modo que la bisección converge; el código realiza 90 iteraciones tras duplicar el extremo superior hasta acotar la raíz, lo que deja el error muy por debajo de la tolerancia numérica. La proyección resultante es

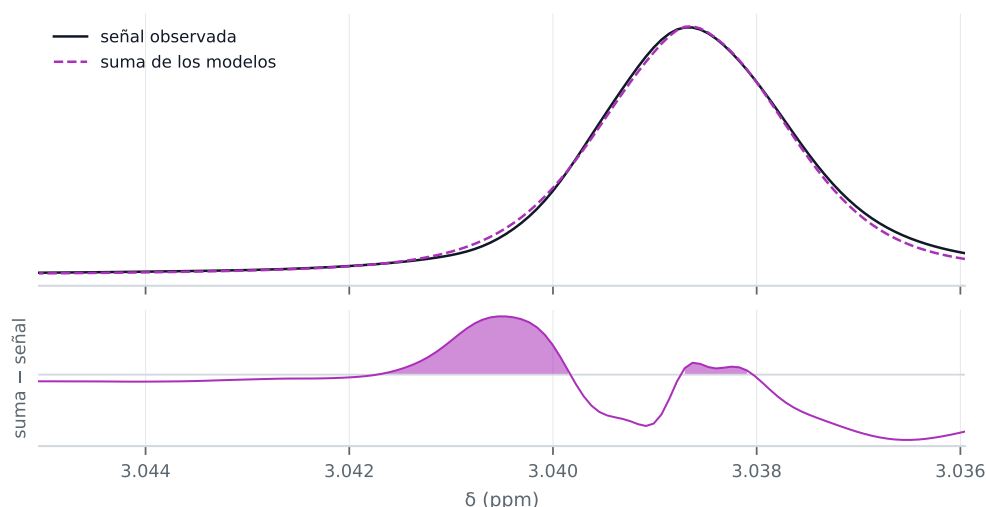


Figura 4: Por qué hace falta proyectar. Arriba, un detalle de SET_5 donde la suma de los modelos (discontinua) va por encima de la señal observada. Abajo, la diferencia: a la escala del pico es un pelo, pero apilar áreas que suman más de lo que hay produciría una figura falsa.

$$\Pi_i = \min(C_i + \min(Q_i, \lambda\omega_i), y_i). \quad (19)$$

El uso de ω como peso significa que el área que hubo que recolocar se deposita donde el modelo original ya era intenso, no repartida uniformemente: así la forma se conserva en lugar de aplanarse.

— 8.3 División del agregado entre componentes

Con la suma ya dentro de la señal, queda repartirla de modo que cada componente recupere *exactamente* su AUC. Se normaliza cada modelo por su máximo, $\beta_{k,i} = M_{k,i} / \max_j M_{k,j}$, y se buscan multiplicadores μ_k tales que las cuotas

$$P_{k,i} = \frac{\mu_k \beta_{k,i}}{\sum_l \mu_l \beta_{l,i}} \Pi_i \quad (20)$$

satisfagan $A[P_k] = A_k$. Nótese que (20) garantiza $\sum_k P_{k,i} = \Pi_i \leq y_i$ en *cada* punto por construcción: las cuotas son una partición de la unidad. Los multiplicadores se obtienen por iteración de punto fijo,

$$\mu_k \leftarrow \mu_k \cdot \frac{A_k}{A[P_k]}, \quad (21)$$

hasta que el error relativo máximo cae por debajo de 2×10^{-13} . Es el mismo esquema de escalado proporcional iterativo que se usa para ajustar una matriz a márgenes dados; converge porque cada actualización corrige monótonamente la discrepancia de su propia fila mientras la normalización mantiene la restricción de columna.

Por último, la componente activa de mayor índice se fija por diferencia, $P_{K,i} = \Pi_i - \sum_{k < K} P_{k,i}$, con lo que la suma es exacta hasta el último bit y no acumula el error de redondeo de K sumas independientes.

— 8.4 El residual

Lo que la partición no ocupa es el residual:

$$r_i = \max\left(y_i - \sum_k P_{k,i}, 0\right). \quad (22)$$

Con eso, (16) se cumple por construcción. El error de cierre observado en la práctica es del orden de 10^{-11} , es decir, precisión de coma flotante y no aproximación del método.

Qué significa un residual alto

No es un defecto del cálculo: es información. Un residual elevado en una región SET dice que los metabolitos declarados *no explican* lo que hay ahí. Puede deberse a un compuesto no incluido en el catálogo, a una impureza, o a una señal mal asignada. Por eso el residual se reporta, se dibuja y se exporta en lugar de esconderse.

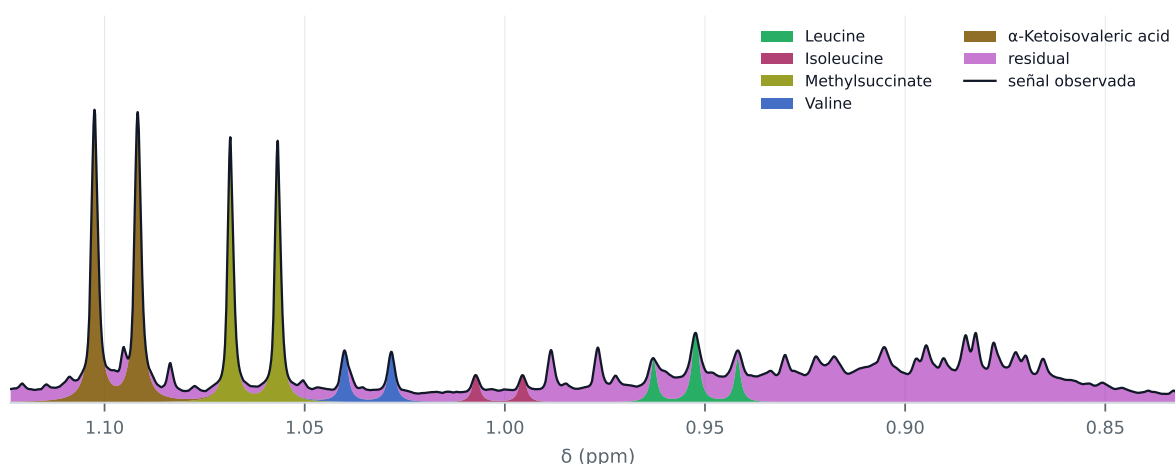


Figura 5: El resultado en SET_1: las áreas de los cinco metabolitos, apiladas y mutuamente excluyentes, con el residual en magenta encima. La pila nunca supera la línea de la señal observada, y su suma es exactamente el área de la región. El área rellena es la misma que se integró para reportar cada AUC.

9 Paso 7 · Curva de integración y curva dibujada

Una regla del proyecto, con consecuencias prácticas:

$$\text{curva dibujada} \equiv \text{curva integrada}. \quad (23)$$

Toda región coloreada en una figura pertenece exactamente a la curva cuyo área se reportó. No existe la categoría de “ala visual”: si una prolongación de la señal se acepta como físicamente real, entra en el AUC; si no se acepta, no se dibuja. Pueden existir curvas auxiliares de diagnóstico —candidatos rechazados, residuos, límites de soporte— pero se muestran con estilo distinto, no se rellenan y no alimentan el área reportada.

Esto convierte la figura en un instrumento de verificación: mirarla es comprobar el número.

10 Paso 8 · Estados de calidad

El resultado no es sólo un número: lleva un juicio sobre si ese número es defendible.

El orden de prioridades cuando estos criterios entran en conflicto es explícito y no negociable:

Estado	Criterio
ACCEPT_FOR_QNMR	Identidad química correcta, área identificable y estable frente a perturbaciones del ajuste.
LOW_CONFIDENCE	Señal plausible con incertidumbre material. No puede promoverse a cuantitativa.
REJECT_FOR_QNMR	La identidad o el área individual no son defendibles.
UNREVIEWED	Evidencia insuficiente para juzgar.
Combinado	El área conjunta es defendible pero su reparto entre metabolitos no es identificable; se reporta como envoltente.

1. identidad química correcta;
2. no atribuir área ajena ni contarla dos veces;
3. área identificable y compatible con el alcance declarado;
4. centro, multiplicidad y J coherentes con los datos;
5. cuerpo, ápice y paredes bien representados;
6. alas físicamente justificadas;
7. residuo, estética y proporción de casos aceptados.

Nunca se mejora un criterio inferior sacrificando uno superior. En particular, **reducir el residuo no justifica mover un centro**.

11 Determinismo

Que dos ejecuciones del mismo espectro den el mismo número no es automático. Un optimizador no lineal es sensible a diferencias del último bit, y esas diferencias aparecen solas si las reducciones internas de las bibliotecas de álgebra lineal se reparten entre hilos de forma no determinista.

TIAGO fija, antes de importar NumPy:

```
PYTHONHASHSEED=0 OMP_NUM_THREADS=1 MKL_NUM_THREADS=1
MKL_CBWR=COMPATIBLE OPENBLAS_NUM_THREADS=1
KMP_DETERMINISTIC_REDUCTION=TRUE
```

y analiza cada espectro en un intérprete nuevo, de modo que el resultado no depende de la cohorte, del orden ni de qué muestra se procesó antes.

El determinismo no sobrevive a un cambio de biblioteca

Al comparar la misma muestra bajo NumPy 1.26.4 / SciPy 1.16.3 frente a NumPy 2.5.1 / SciPy 1.18.0, la mayoría de las regiones se movió dentro del ruido de coma flotante ($\sim 10^{-9}$ relativo), pero **SET_41 / Histidine_1 cambió un 5,95 %** —de 25 908 a 24 367— y SET_9 / 3-Hydroxyisovalerate_2 un 4×10^{-5} . El optimizador aterriza en otro mínimo local.

Por eso las versiones están fijadas y no acotadas, y la aplicación avisa si el entorno instalado se aparta del validado.

12 Del área a la concentración

TIAGO se detiene en el área. El paso siguiente, que corresponde al usuario, requiere una referencia:

$$c_m = \frac{A_m}{n_m} \cdot \frac{n_{\text{ref}}}{A_{\text{ref}}} \cdot c_{\text{ref}} \cdot f_{\text{dil}}, \quad (24)$$

donde n es el número de protones de cada señal y f_{dil} recoge la dilución. El proyecto contempla dos referencias:

DSS Referencia de desplazamiento químico en 0 ppm; su singlete de 9 protones puede servir además como patrón interno de concentración.

ERETIC Señal patrón sintética, generada electrónicamente, que evita añadir un compuesto a la muestra.

En orina es habitual normalizar por creatinina o por cuantiles para corregir la dilución, que varía entre muestras por razones fisiológicas ajenas al metabolismo que se estudia. Esa elección es del estudio, no del programa.

13 Resumen del flujo

