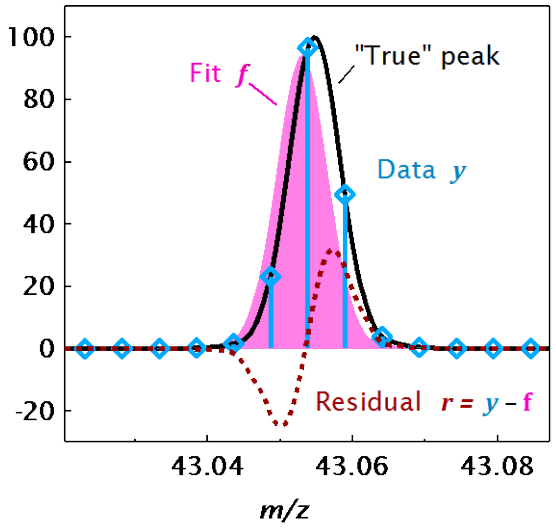
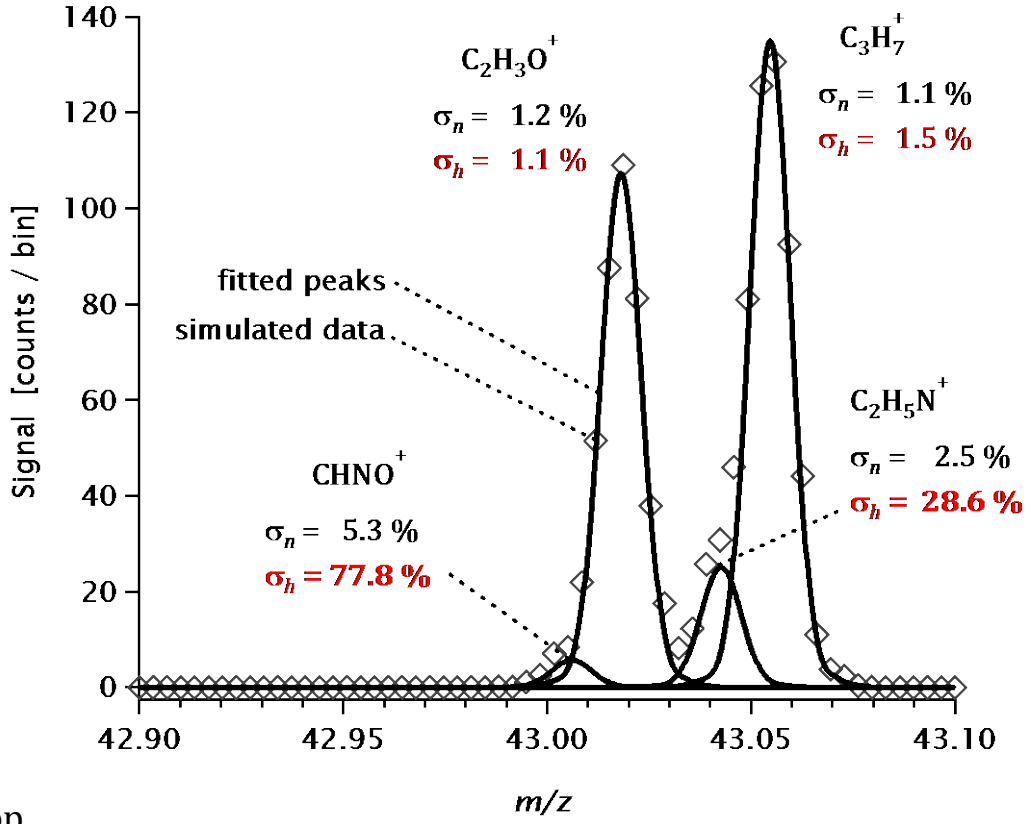
Peak-integration uncertainties for HR-AMS—PMF: overlapping peaks & importance of mass accuracy





J. C. Corbin, A. Othman, J. D. Allan, D. R. Worsnop,
J. D. Haskins, B. Sierau, U. Lohmann, A. A. Mensah, AMTD: http://dx.doi.org/10.5194/amtd-8-3471-2015

Joel C. Corbin

ETH Zurich & PSI, Switzerland

2015-09-11

When do we need uncertainties?

- PMF
- ME-2
- Linear regression (e.g. $C_2H_3O^+$ vs. CO_2^+)

Both PMF & linear regression solve

$$y = \mathbf{FG} + e$$

data = model + residuals

by minimizing

$$Q = \sum_{i,j} \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2$$

imprecision

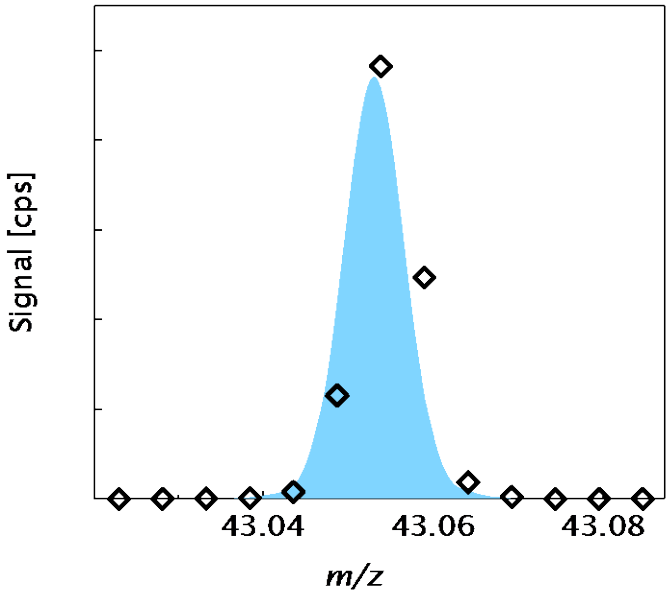
…which makes the precision (imprecision σ) as important as the measurements (y).[1]

[1] Hopke and Paatero 2003

Current HR-AMS-peak uncertainties

Same as for UMR-AMS:

"how would this signal vary if we repeated the measurement?"



n_{ions}

$\sigma_n = \sqrt{n_{\text{ions}}}$

Needs "corrections" for sampling configuration (sampling time, etc.)
see Allan et al., JGR 2003
or the AMTD link on Slide 1

Another source of uncertainty:

“How would this peak-integration vary if we repeated the analysis?”

Rest of this talk

Outline

1. PIKA peak-integration errors

2. Understanding of peak-fitting errors
   via case study for single (isolated) ions

3. Application to single & overlapping peaks

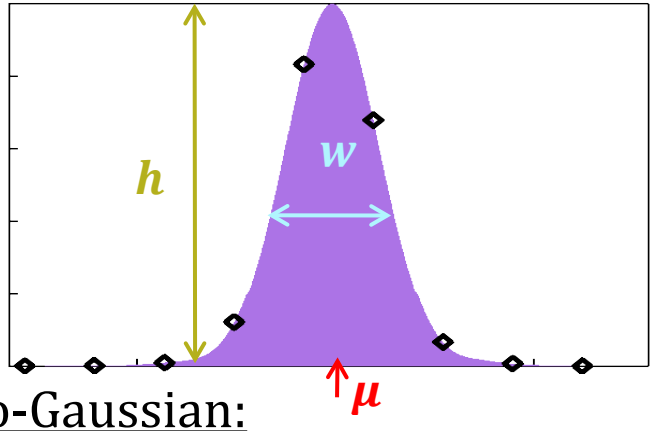
4. Practical application

NB: For simplicity I won't discuss the PIKA uncertainty changes in 1.11C. They make no difference in this context.

Definitions

- σ = imprecision
- ε = imprecision + bias (overall error)
- * All existing peaks assumed to be identified

Essential ideas in PIKA



Peak height h
from fit,
per peak

Peak location μ
from cal.,
per peak

Peak width w
from cal.,
per integer m/z

Peak shape v
from average,
per data set

1. Fit height h of pseudo-Gaussian:

$$f(x) = h \cdot v \cdot \exp\left(\frac{(x-\mu)^2}{-w^2}\right)$$

$$f(x) = h f_0$$

2. Integrate fitted peak:

$$A = hw\sqrt{\pi}\left(k_{\mathrm{DC}}\frac{A_{\mathrm{pseudoG}}}{A_{\mathrm{Gauss}}}\right)$$

3. Estimate error:

$$\sigma_n = \sqrt{n}$$

counting imprecision

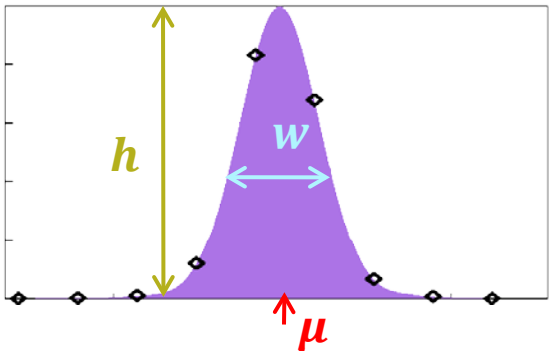
$$\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2$$

Peak-integration uncertainty

Estimating σ_A

$$\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2 \longleftarrow A = hw\sqrt{\pi}\left(k_{\mathrm{DC}}\frac{A_{\mathrm{pseudoG}}}{A_{\mathrm{Gauss}}}\right)$$

$$\sigma_{\mathrm{AMS}} = \sigma_n + \sigma_A$$



$\sigma_w \leftarrow$ direct estimate from width calibration

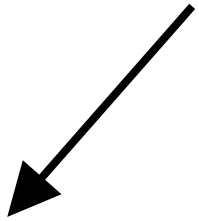
$\sigma_h \leftarrow$ Monte-Carlo estimate

from empirically-estimated σ_v, σ_w, ε_μ

Outline

1. PIKA peak-integration errors

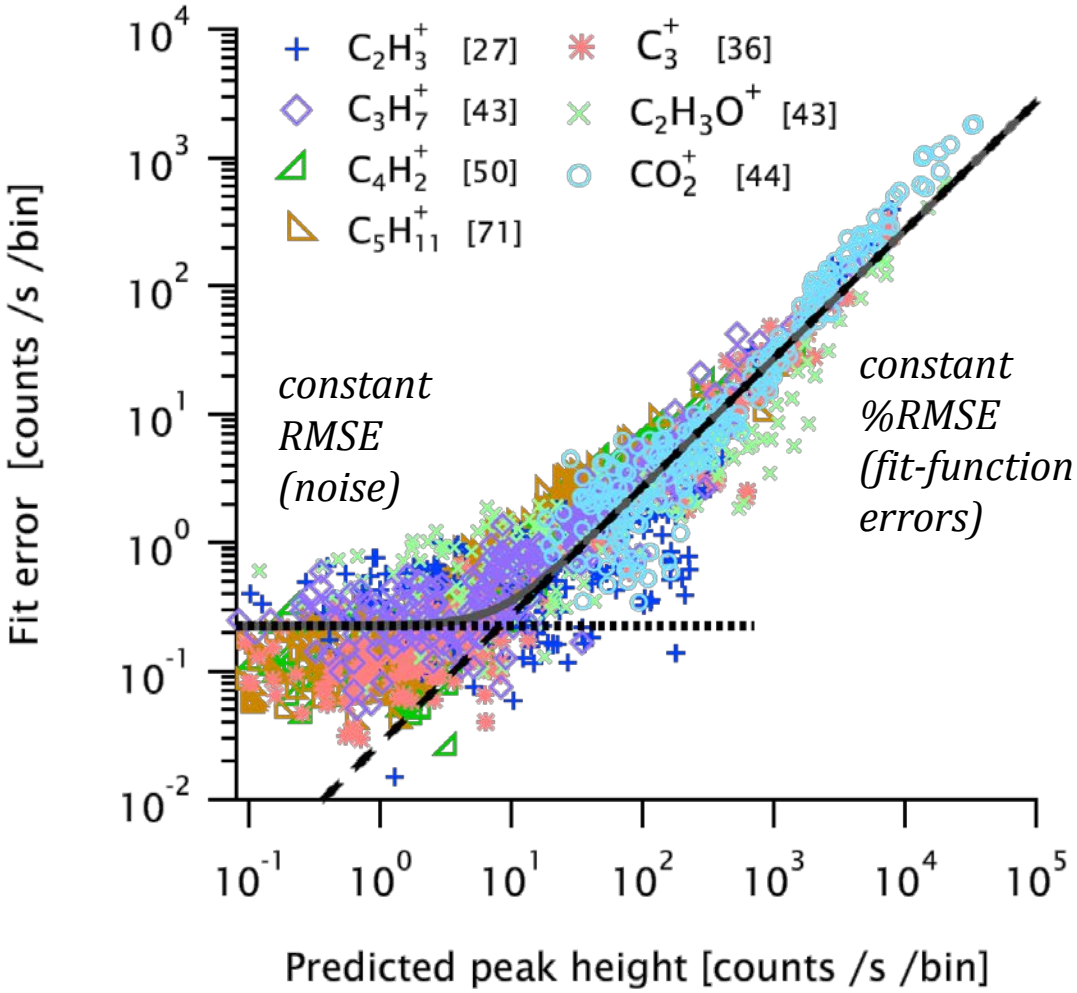
2. Understanding of peak-fitting errors
   via study of 7 single (isolated) ions in test data set



$\sigma_h \leftarrow$ Monte-Carlo estimate

from empirically-estimated σ_v, σ_w, ε_μ

Fit errors (RMSE's) for 7 isolated peaks



Each point shows a peak.

All ions show:
1. Noise regime < 10 cps
2. Constant %RMSE regime

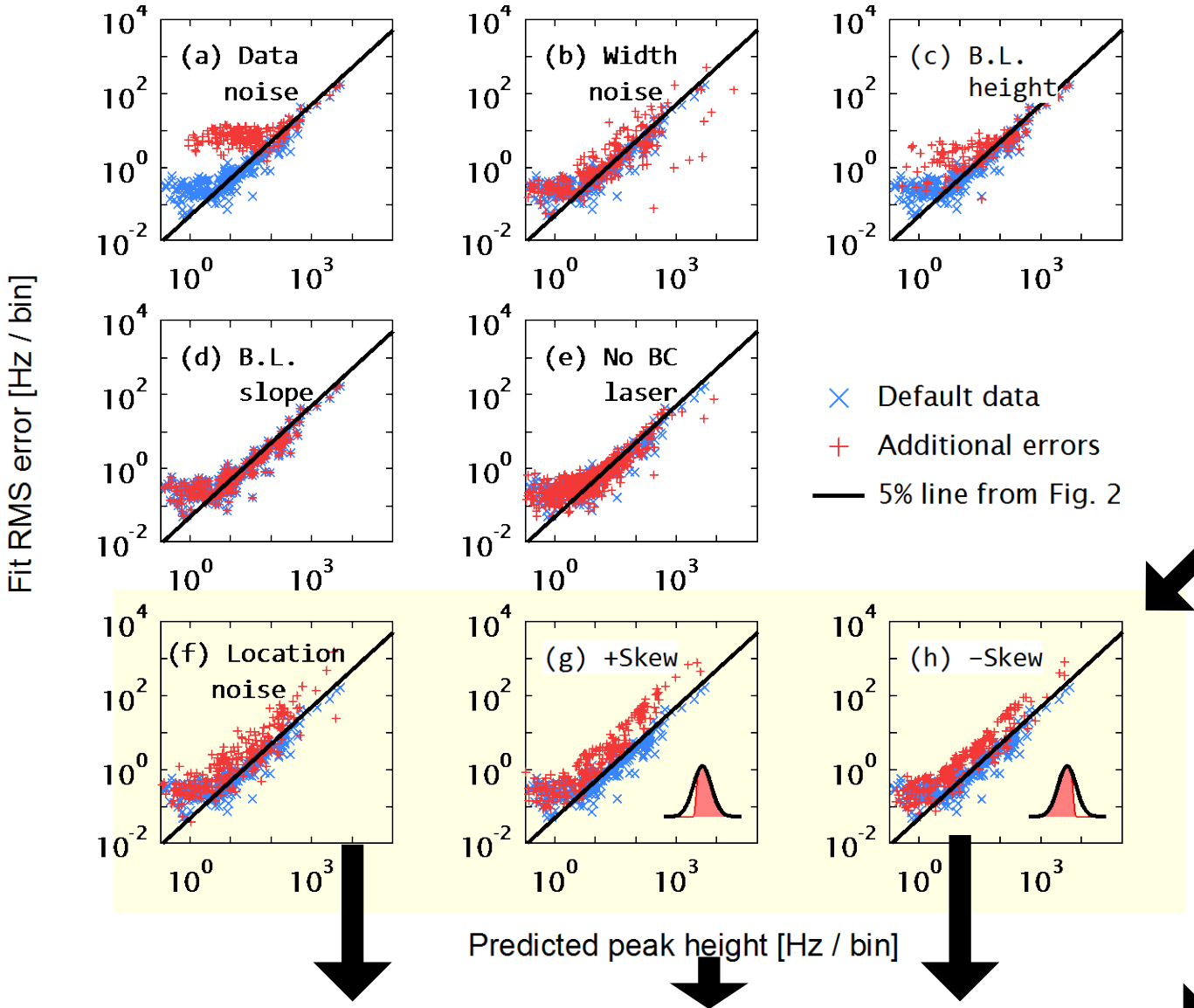
Constant ratio because h scales errors in f_0:

$$f(x) = hf_0$$

(details in AMT paper)

These are Diff data to avoid differences in backgrounds.

What controls the %RMSE?



Noise added within PIKA *(amount of noise chosen to have visible effects)*

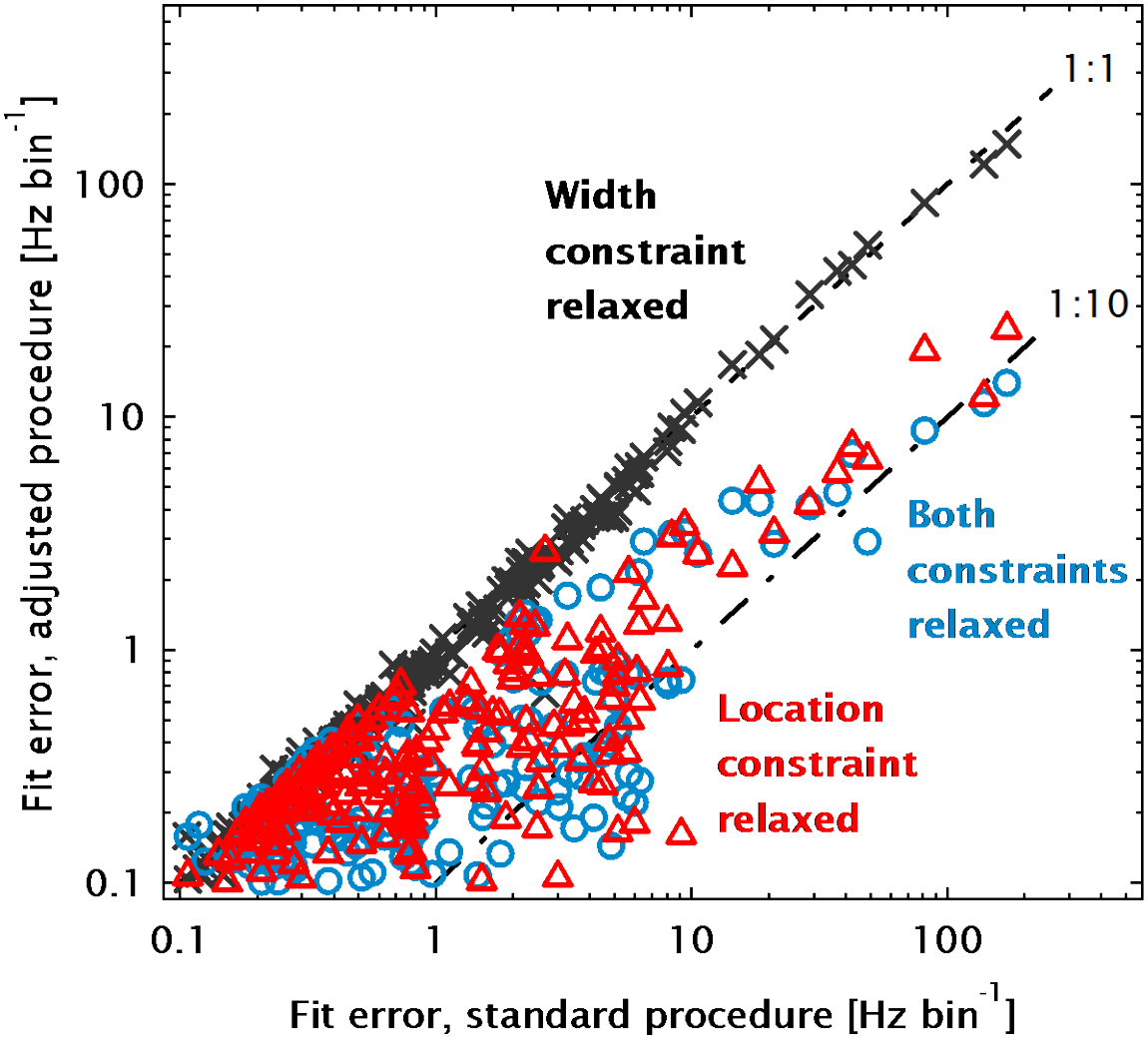
If %RMSE increases, red data become higher than blue data

errors ~10x higher

errors ~100x higher

Only noise in location μ matters for fitting errors

m/z prediction errors ε_μ are the major cause of fitting errors (RMSE).



Location-prediction (m/z calibration) errors ε_μ are ~10x more important than peak-shape errors.

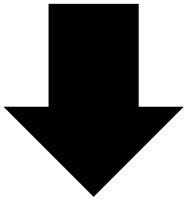
$$f(x) = h \cdot v \cdot \exp\left(\frac{(x-\mu)^2}{-w^2}\right)$$

$$f(x) = hf_0$$

Estimating σ_A (updated)

$$\left(\frac{\sigma_A}{A}\right)^2 = \left(\frac{\sigma_h}{h}\right)^2 + \left(\frac{\sigma_w}{w}\right)^2$$

$$\sigma_{\mathrm{AMS}} = \sigma_n + \sigma_A$$



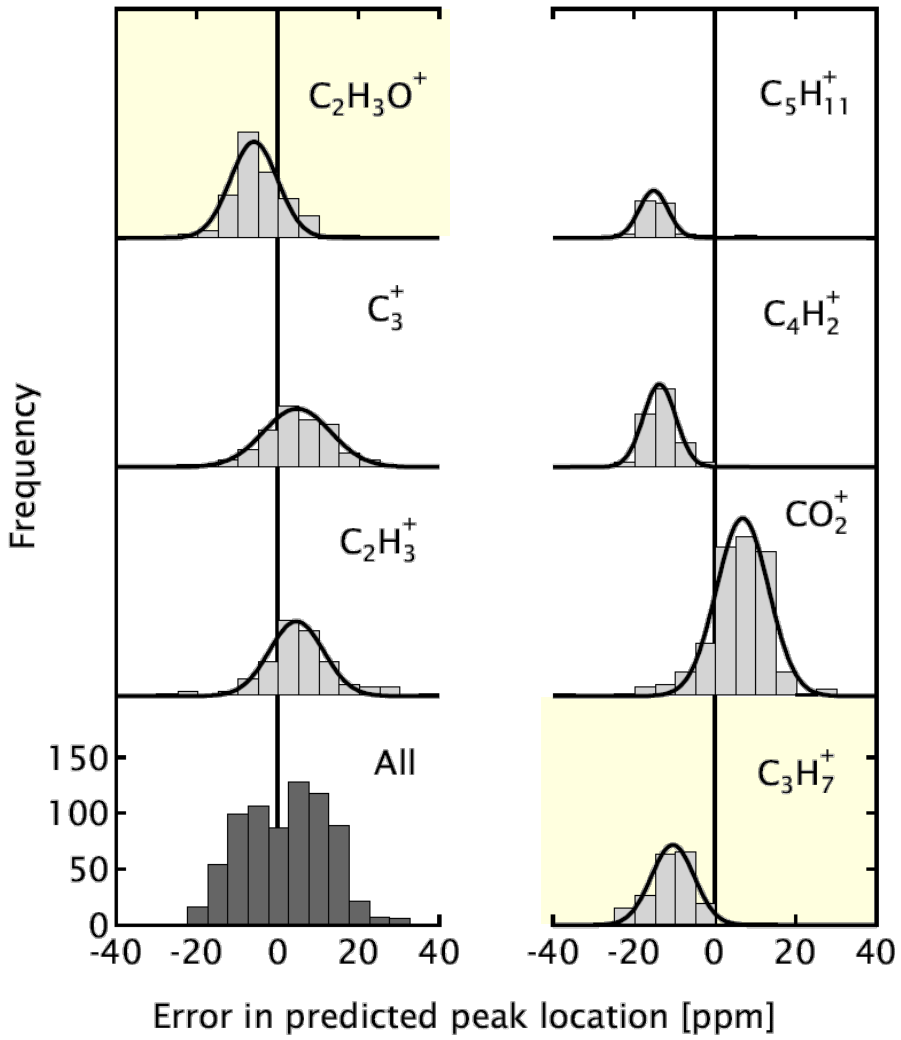
σ_w ← direct estimate from width calibration

σ_h ← Monte-Carlo estimate

from empirically-estimated ~~σ_v~~, ~~σ_w~~, ε_μ

So, we need to know ε_μ: for isolated peaks, estimate as

$$(\mu_{\text{free-fit}}) - (\mu_{m/z\ \text{cal}})$$



Different bias (mean) and imprecision (spread) at each ion,

even for $C_2H_3O^+$ and $C_3H_7^+$ (m/z 43, $\Delta m/z$ = 0.036).

Can only estimate ion-specific ε_μ for isolated peaks!

Outline

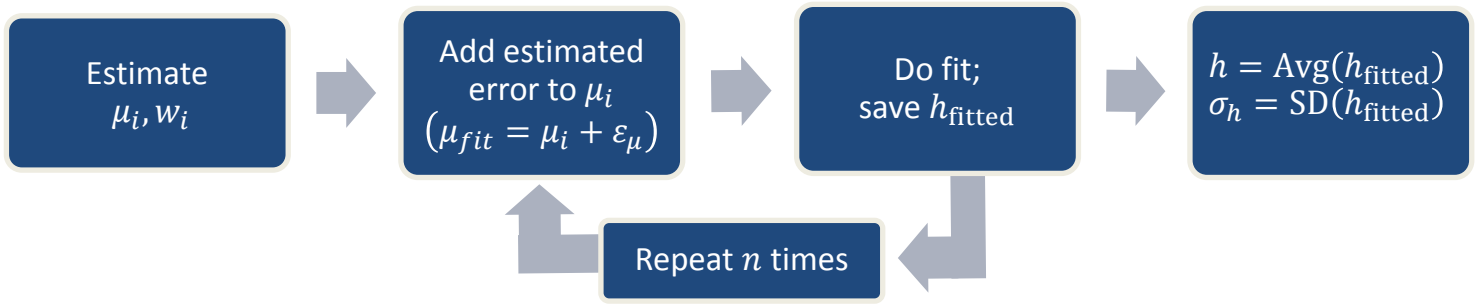
1. PIKA peak-integration errors

2. Understanding of peak-fitting errors*
   - via case study for single (isolated) ions

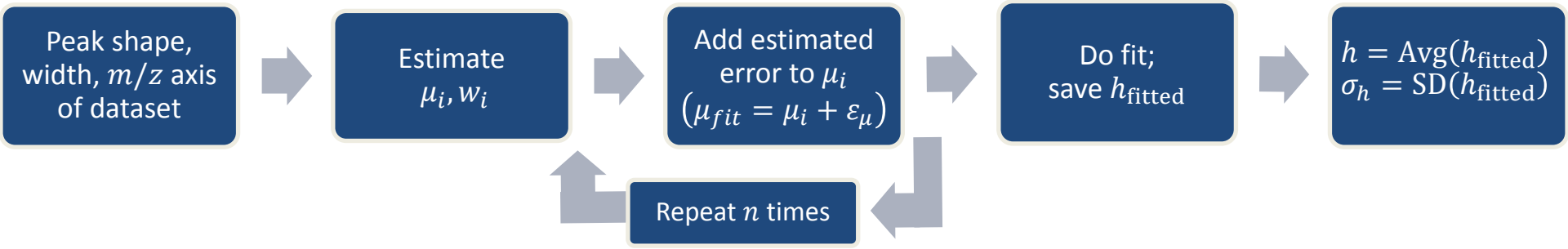
3. Application to single & overlapping peaks:
   - Monte Carlo estimation

4. Practical application

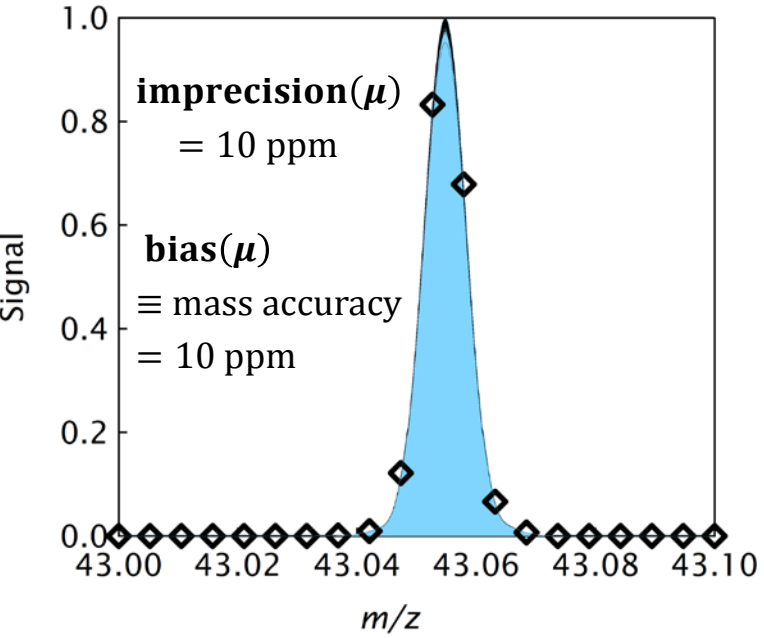
Monte Carlo σ_h estimation



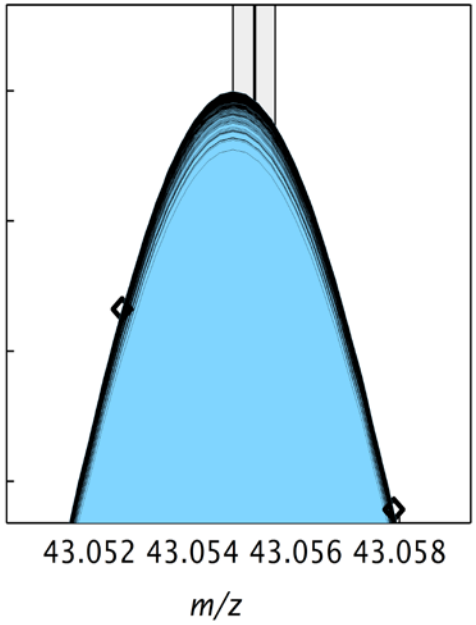
Monte Carlo σ_h estimation



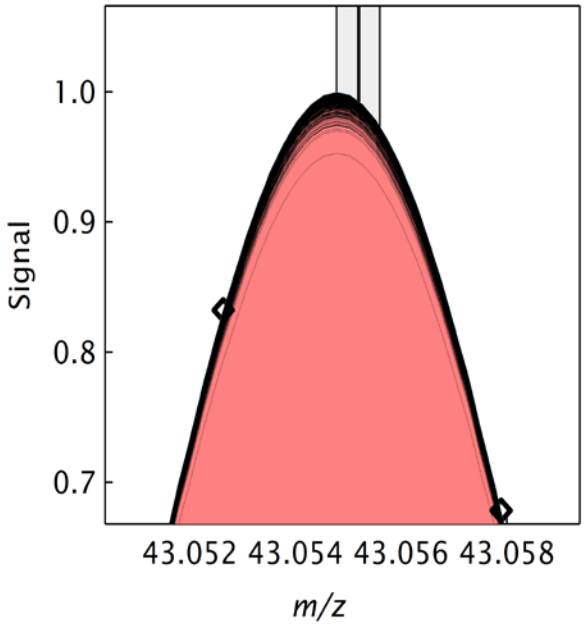
1000 fits to $C_3H_7^+$



zoom



Without bias term



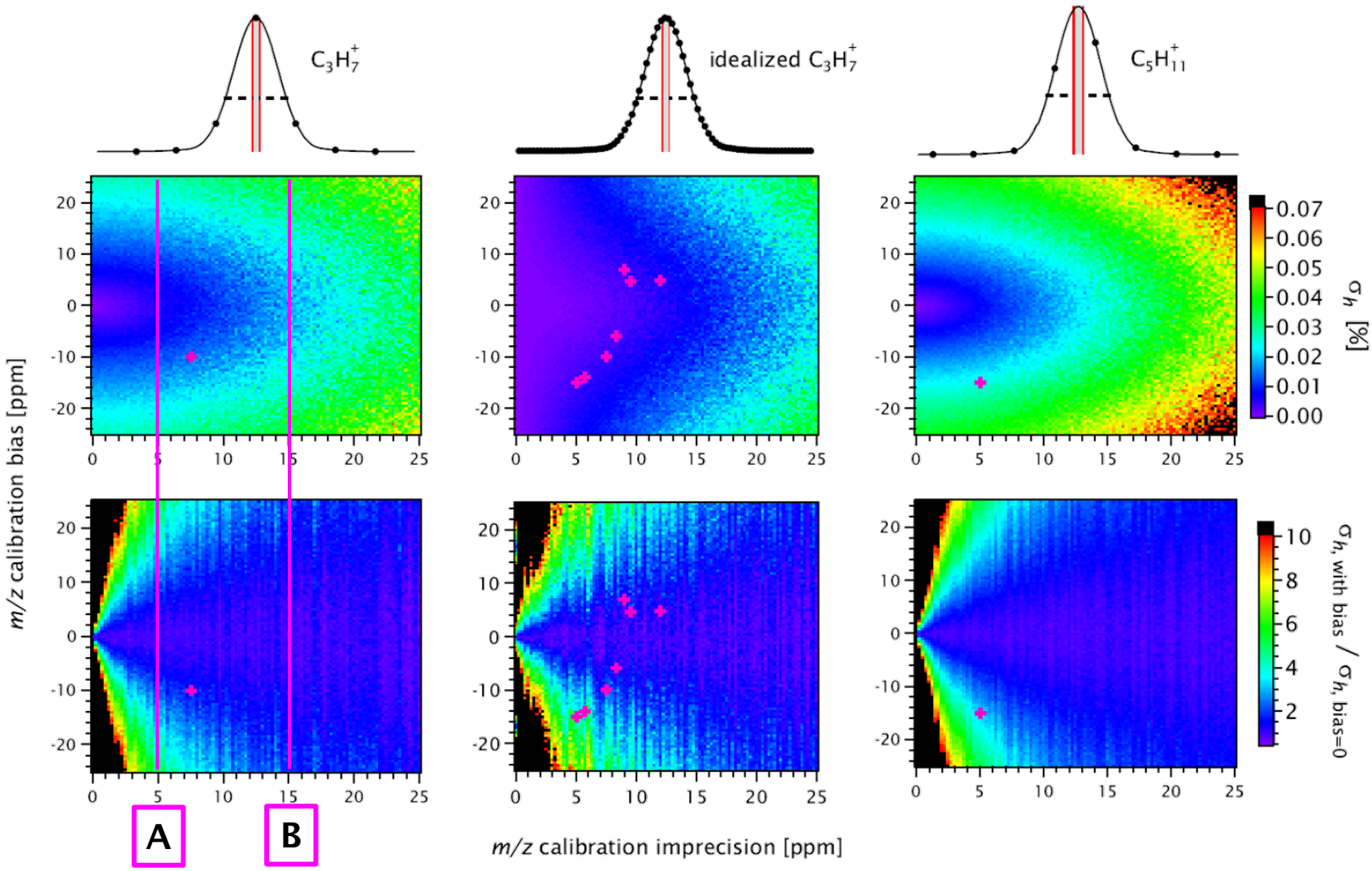
Errors in μ programmed as gaussian mean and variance.
On zoomed graphs, middle line is bias, other lines are bias $\pm 1\sigma$ imprecision.

Summary of all 7 ions' σ_h

Ion	μ-prediction error	Error in fitted h (bias [%], imprecision σ_h [%])		
		from best-estimate μ errors	with μ imprecision only	with broader peaks
$C_2H_3^+$	4.6 ± 9.5	-0.35, 1.06	-0.20, 0.93	-0.08, 0.13
$C_3H_7^+$	-10 ± 7.5	-0.65, 1.64	-0.17, 0.98	-0.19, 0.21
$C_4H_2^+$	-14 ± 5.7	-1.06, 2.10	-0.11, 0.78	-0.31, 0.23
$C_5H_{11}^+$	-15 ± 5.0	-1.03, 2.46	-0.12, 0.79	-0.49, 0.31
C_3^+	4.8 ± 12	-0.51, 1.52	-0.41, 1.43	-0.16, 0.24
$C_2H_3O^+$	-5.9 ± 8.3	-0.39, 1.36	-0.22, 1.06	-0.13, 0.17
CO_2^+	6.9 ± 9.0	-0.12, 1.41	-0.20, 1.13	-0.15, 0.21

How to deal with "unknowable" ε_μ?

$$\varepsilon_\mu = \text{bias}(\mu) + \sigma_\mu$$



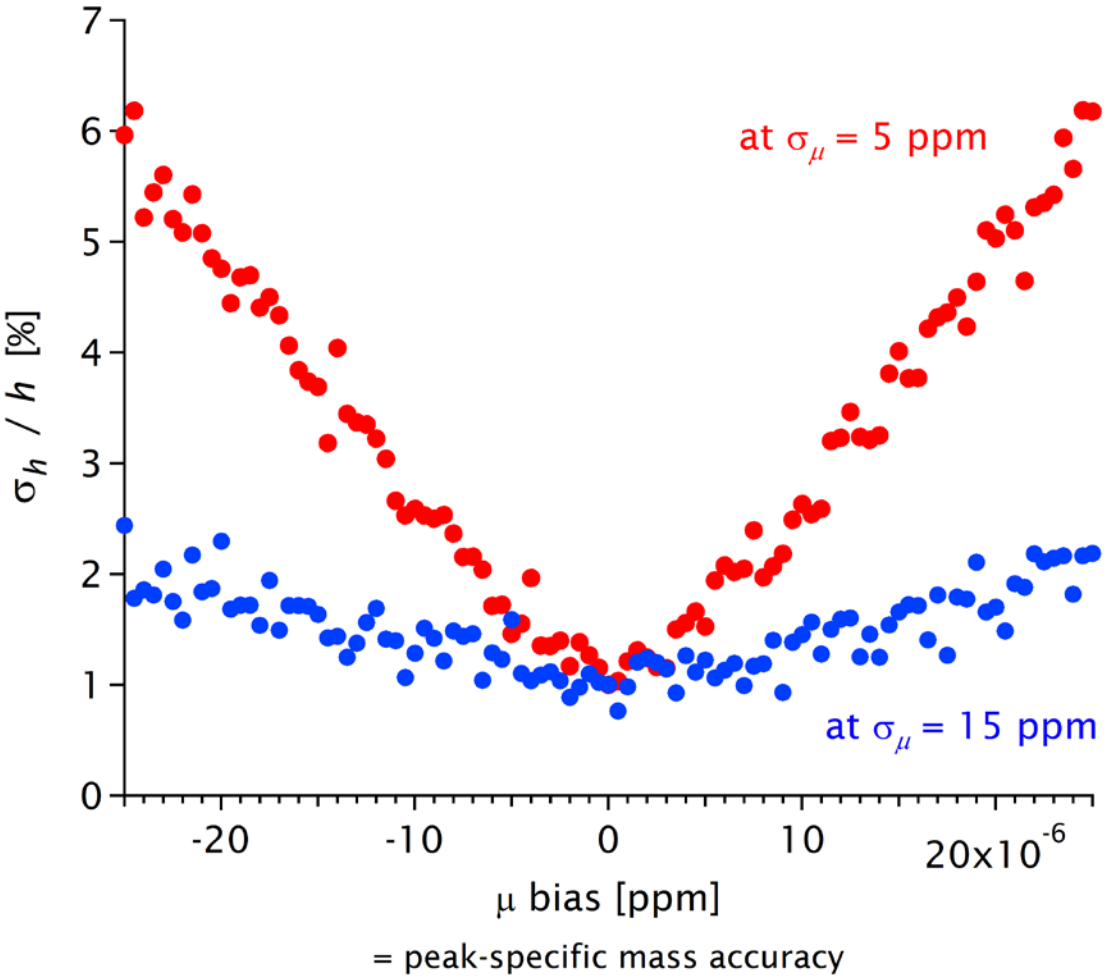
at **A**: $\sigma_\mu = 5\%$, neglecting a bias(μ) of ±10 strongly affects σ_h (1--2%)

at **B**: $\sigma_\mu = 15\%$, neglecting a bias(μ) of ±10 negligibly affects σ_h (3%)

Overestimate σ_μ !!

Cross-section of previous image (3rd row)

$$\varepsilon_\mu = \text{bias}(\mu) + \sigma_\mu$$



$\sigma_\mu = 5\%$, neglecting a $\text{bias}(\mu)$ of ±10 strongly affects σ_h (1--2%)

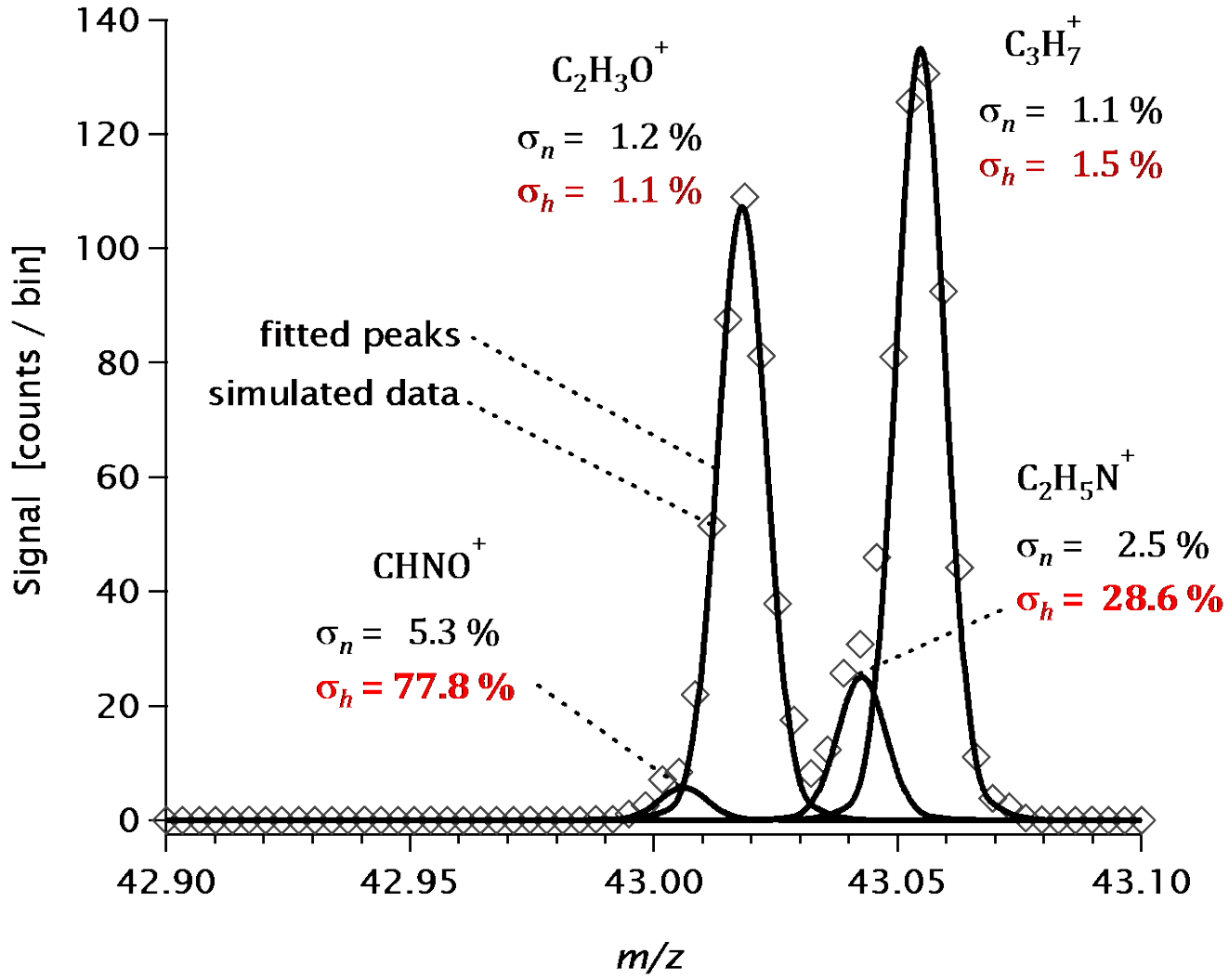
$\sigma_\mu = 15\%$, neglecting a $\text{bias}(\mu)$ of ±10 negligibly affects σ_h (3%)

} **Overestimate σ_μ !!**

Application to overlapping peaks:

MC with 100 samples, exactly the same scenario.

Inputs: a single σ_μ estimate + the usual peak parameters (w, ν, μ)



This matters for PMF

- Test data set with high signals
  - Counting imprecision
  - 5% imprecision (not overlapping errors)

- Without including 5% in PMF, high signals became residual spikes

- Lower-signal factors retrieved with r^2 0.54 without 5% included in σ, down from 0.74
  - Higher signals from 0.91 to 0.99

- Details in AMTD

Practicalities in σ_h estimation

- MC estimation with 100 fits takes 100x longer.

- For preliminary analyses, we could therefore:
  - Estimate the uncertainty in isolated peaks for only a subset of peaks, since σ_h/h is roughly constant [1]
  - Use the Cubison and Jimenez parameterization [2] to estimate the uncertainty in peaks that overlap significantly
    - For multiple overlapping peaks, we consider only the two closest peaks

- At the final stage of analysis, these initial estimates can be refined by MC estimation. The refined estimates will have the advantages of accounting for possible m/z-axis sensitivities, and for cases where >2 peaks overlap significantly.

- (Until this is available in PIKA, PMF's C3 parameter provides a rough approach to isolated-peak errors)

[1] Corbin et al., AMTD 2015; *[2]* Cubison and Jimenez, AMT 2015

Summary

1. Peak-integration uncertainties dominate ions with high signals or bad overlap.
   - For well-resolved, high-signal ions, the uncertainties can be estimated as constant %.

2. These can be quantified by Monte Carlo with m/z cal. imprecision $\boldsymbol{\sigma_\mu}$ as the only input.
   - Applies to overlapping & isolated ions (gives the constant% above).
   - The range of reasonable $\text{bias}(\boldsymbol{\mu})$ needs to be roughly known, so that $\boldsymbol{\sigma_\mu}$ can be increased to account for the effect of $\text{bias}(\boldsymbol{\mu})$ on $\boldsymbol{\sigma_h}$.

3. Will be incorporated into PIKA soon.