NUScon 2018 Evaluation Metrics

1 Introduction

The metrics defined here quantify the similarity between two peak lists. These metrics are utilized in NUScon to measure how synthetic peaks lists are recovered by candidate reconstructions. The design of the synthetic peak lists and the metrics used to make the comparison reveal specific information about the quality of the spectral reconstruction.

The following notation is used in defining the metrics

property	notation	description
peak list	$\mathcal{A}$	set
number	$\|\mathcal{A}\|$	scalar
peak	$p_i^{\mathcal{A}}$	single peak from list $\mathcal{A}$
dimensions	d	scalar
position	$\hat{p}_i = [\hat{p}_i[1], \hat{p}_i[2], \ldots, \hat{p}_i[d]]$	vector [ppm values]
intensity	$\|p_i\|$	single scalar value
linewidth	$< p_i >= [< p_i[1] >, < p_i[2] >, \ldots, < p_i[d] >]$	vector [Hz values]

2 Metrics

M1: Frequency Accuracy

In order to quantify the frequency accuracy of a list of recovered peaks ($\mathcal{R}$) relative to a list of master peaks ($\mathcal{M}$) we need a method for computing the "distance" between peak positions. While it is trivial to compute the distance between pairs of peaks, it is not trivial to assign the pairings when the two lists may have different numbers of peaks due to artifacts, inadequate resolution, or low sensitivity; this is compounded when nearest neighbors are not mutual.

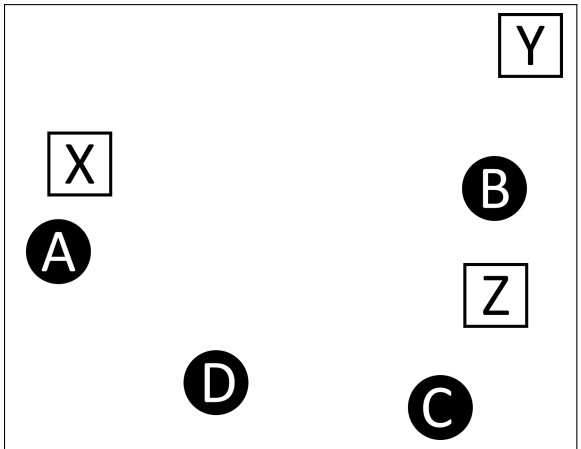

Figure 1. Correspondence. This example illustrates a case where recovered peaks (white squares) are slightly offset from the injected master peaks (black circles). One injected peak (D) is not recovered and the offset introduces challenges relating to the correspondence between the two lists.

For example, consider Figure 1, which depicts 4 master peaks (black circles):

$$\mathcal{M} = \{A, B, C, D\} \tag{1}$$

and 3 recovered peaks

$$\mathcal{R} = \{X, Y, Z\} \tag{2}$$

This illustration is meant to represent a case where 3 of the master peaks are recovered, but at slightly offset positions ($A \rightarrow X$, $B \rightarrow Y$, $C \rightarrow Z$) and a fourth master peak (D) is not recovered at all. Let's consider these relationships by finding the recovered peak that is closest to each master peak, and vice versa; this produces the following pairings:

injected → recovered	
A → X	This is as expected and provides an appropriate pairing to compute a distance.
B → Z	The slight shift in the positions of Y and Z are such that Z is the closest recovered peak for B, even though the illustration implies B should pair with Y.
C → Z	This is as expected, but Z is also claimed by B, which is even closer to Z, even though the illustration implies Z goes with C.
D → X	There is no peak recovered near D, so it gets partnered with whatever is closest by, even though it is clearly not related to the injection of D.
recovered → injected	
X → A	This is as expected and provides an appropriate pairing to compute a distance.
Y → B	This is as expected and provides an appropriate pairing to compute a distance.
Z → B	The slight shift in the positions of Y and Z are such that B is the closest injected peak for Z, even though the illustration implies Z should pair with C.

In comparing the two assignment tables, we see a variety of situations:

- A and X are mutual partners.
- B and Z are mutual partners.
- C selects Z, but Z selects B, which is also selected by Y.
- D has no appropriate partner and selects X, which is clearly paired with A.
- No injected peak selects Y.
- No recovered peak selects C.

We account for the situations depicted above by using a symmetric Hausdorff distance, which considers the correspondence problems from both directions (and takes the average distance). We also include a maximum penalty term to guard against the distance penalty on D being highly dependent on the surroundings. This metric is formally defined as

$$H(\mathcal{M}, \mathcal{R}) = \frac{1}{2}(D(\mathcal{M} \rightarrow \mathcal{R}) + D(\mathcal{R} \rightarrow \mathcal{M})) \tag{3}$$

The RMSD obtained by selecting each peak in $\mathcal{M}$ and finding its closest neighbor in $\mathcal{R}$ is computed as

$$D(\mathcal{M} \rightarrow \mathcal{R}) = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{p_m \in \mathcal{M}} \min_{p_r \in \mathcal{R}} d(p_m, p_r)^2} \tag{4}$$

The function used to compute the distance between two peaks is defined as

$$d(p_m, p_r) = \min \left(\sqrt{\sum_{i=1}^{d} (c[i] \cdot (\hat{p}_m[i] - \hat{p}_r[i]))^2} \ , \ d_{\text{max}} \right) \tag{5}$$

The first term in the "min" statement is the common L2 norm, but it also includes weighting factors (c^i). Without this adjustment, a 1 ppm difference along a proton dimension would be equivalent to a 1 ppm difference along a nitrogen dimension. In this example, it would be appropriate to use a weighting factor for the proton dimension that is roughly 15x compared to the nitrogen dimension. The second term in the "min" statement is the d_max term, which puts an upper bound on the distance and should be defined relative to the linewidth of the peaks. This term ensures that if no peak from $\mathcal{R}$ is found in the neighborhood of p_m, it does not introduce an arbitrary distance term according to whatever peak in $\mathcal{R}$ is closest.

The values of the symmetric Hausdorff in Eqn 3 fall on the interval $[0, d_{\text{max}}]$, with 0 indicating a perfect recovery of peak positions and a value of d_{max} indicating no injected peaks have a recovered peak within the cutoff distance. This is normalized onto the interval [0,1] as

$$\Delta(\mathcal{M}, \mathcal{R}) = \frac{H(\mathcal{M}, \mathcal{R})}{d_{\text{max}}} \tag{6}$$

M2: Linearity of peak intensity

The intention of this metric is to quantify the linearity of the peak intensities recovered by a reconstruction relative to the known intensities from a synthetically injected peak list. This metric assumes that the recovered peaks are neither limited in sensitivity nor resolution and that the correspondence of peaks between $\mathcal{M}$ and $\mathcal{R}$ is clearly defined. That is the peaks are detectable above the noise, not overlapping, and the same indexing can be used to refer to both sets.

Let a set of n peaks from $\mathcal{M}$ and $\mathcal{R}$ be denoted

$$\mathcal{M} = [p_1^{\mathcal{M}}, p_2^{\mathcal{M}}, \ldots, p_n^{\mathcal{M}}] \tag{7}$$
$$\mathcal{R} = [p_1^{\mathcal{R}}, p_2^{\mathcal{R}}, \ldots, p_n^{\mathcal{R}}] \tag{8}$$

where $p_i^{\mathcal{R}}$ is the recovered peak corresponding to the master peak $p_i^{\mathcal{M}}$. The intensities of the peak lists are given as the vectors

$$I^{\mathcal{M}} = [|p_1^{\mathcal{M}}|, |p_2^{\mathcal{M}}|, \ldots, |p_n^{\mathcal{M}}|] \tag{9}$$
$$I^{\mathcal{R}} = [|p_1^{\mathcal{R}}|, |p_2^{\mathcal{R}}|, \ldots, |p_n^{\mathcal{R}}|] \tag{10}$$

The Pearson correlation coefficient is computed as

$$\rho(I^{\mathcal{M}}, I^{\mathcal{R}}) = \frac{\text{cov}(I^{\mathcal{M}}, I^{\mathcal{R}})}{\text{std}(I^{\mathcal{M}}) \times \text{std}(I^{\mathcal{R}})} \tag{11}$$

The correlation coefficient is on the interval [-1,+1], where the sign indicates a positive / negative correlation and the magnitude indicates the strength of correlation. This is converted to a value on the interval [0,1] as

$$L(\mathcal{M}, \mathcal{R}) = \frac{1 + \rho(I^{\mathcal{M}}, I^{\mathcal{R}})}{2} \tag{12}$$

M3: True positive rate

It might be tempting to simply count the number of peaks in $\mathcal{R}$ that are within a cutoff distance of a peak in $\mathcal{M}$, or count how many peaks in $\mathcal{M}$ have a peak in $\mathcal{R}$ within a cutoff distance. However, these approaches would allow a peak from one set to be considered a neighbor to multiple peaks in the other, thus artificially raising the count. While we do not have to solve the correspondence problem here, we have to ensure that peaks are counted at most once. We achieve this result by framing the metric as a bipartite graph, where the two peak lists form the two sets of vertices and edges connect peaks from $\mathcal{M}$ to peaks from $\mathcal{R}$ that are within the cutoff distance. The metric is then equivalent to finding the cardinality of a maximum matching on the bipartite graph (i.e. the most number of pairings between $\mathcal{M}$ and $\mathcal{R}$ where peaks are in at most one pairing). We denote the maximum matching as

$$X(\mathcal{M}, \mathcal{R}) = \{(p_m, p_r) \text{ such that } p_m \in \mathcal{M}, p_r \in \mathcal{R}, d(p_m, p_r) \leq r, p_m \text{ and } p_r \text{ not in any other pairing}\} \tag{13}$$

We denote the true positive rate as the percentage of master peaks that appear in the maximum matching

$$T(\mathcal{M}, \mathcal{R}) = \frac{|X(\mathcal{M}, \mathcal{R})|}{|\mathcal{M}|} \tag{14}$$

M4: False positive rate

This metric reports on the rate of false positives in the recovered peak list. The false positive rate, which is defined as the percentage of peaks from $\mathcal{R}$ that do not correspond to a master peak, is obtained as

$$FP_{\text{rate}}(\mathcal{M}, \mathcal{R}) = \frac{|\mathcal{R}| - |X(\mathcal{M}, \mathcal{R})|}{|\mathcal{R}|} \tag{15}$$

This rate is on the interval [0,1], but the best performance (i.e. no false positive peaks in $\mathcal{R}$) corresponds to a value of 0. The output of the metric can be simply reversed, so that the best performance is at a value of 1 and the worst is at 0. This is done by subtracting the false positive rate from 1, producing

$$\begin{aligned} F(\mathcal{M}, \mathcal{R}) &= 1 - \frac{|\mathcal{R}| - |X(\mathcal{M}, \mathcal{R})|}{|\mathcal{R}|} && (16) \\ &= \frac{|X(\mathcal{M}, \mathcal{R})|}{|\mathcal{R}|} && (17) \end{aligned}$$

M5: Valley-to-peak separation

The valley-to-peak ratio (VPR) is a measure of how well resolved a pair of peaks are, based on how close to the baseline the valley between them descends. The master peaks ($p_1^{\mathcal{M}}$ and $p_2^{\mathcal{M}}$) are injected along one of the dimensions of the spectrum so that the line connecting their peaks contains on-grid locations. The injected peaks are recovered ($p_1^{\mathcal{R}}$ and $p_2^{\mathcal{R}}$) and their (potentially off-grid) intensities are captured by the peak picker. The on-grid locations of the master peaks are used to define a connecting line in the reconstructed spectrum, which contains on-grid locations; the lowest intensity value on the line is considered the valley, and is denoted I_{valley}. The quality of the separation is defined as

$$V(p_1^{\mathcal{R}}, p_2^{\mathcal{R}}, I_{\text{valley}}) = 1 - \frac{2 \cdot I_{\text{valley}}}{|p_1^{\mathcal{R}}| + |p_2^{\mathcal{R}}|} \tag{18}$$

A value of 1 indicates separation down to the baseline and a value of 0 indicates the intensity of the valley is the same as that of the peaks. It is possible to achieve values outside of the interval [0,1], but these are extreme situations where the "valley" has a larger intensity than the peaks or if the valley has a negative intensity.

3 Scoring Functions

The metrics defined in the previous section are combined to provide single scores for fidelity and detection. Since all metrics are normalized onto the interval [0,1], it may be tempting to combine multiple metrics through averaging. However, the dynamic range of each metric is hard to pre-determine, thus a mean score could be dominated by a single term. Instead we choose to rank order all submissions for each metric and then combine metrics by averaging of the ranks. The rank of a metric's results, as opposed to its raw value, is denoted by a trailing "#" subscript. The fidelity and detection scores are defined as

$$S_{\text{fidelity}}(\mathcal{M}, \mathcal{R}) = \frac{1}{2} \left(\Delta(\mathcal{M}, \mathcal{R})_{\#} + L(\mathcal{M}, \mathcal{R})_{\#} \right) \tag{19}$$

$$S_{\text{detection}}(\mathcal{M}, \mathcal{R}) = \frac{1}{3}\left(T(\mathcal{M}, \mathcal{R})_{\#} + F(\mathcal{M}, \mathcal{R})_{\#} + V(p_1^{\mathcal{R}}, p_2^{\mathcal{R}}, I_{\text{valley}})_{\#}\right) \tag{20}$$