Performance of different deficient variants of PepNovo.

Algorithm Variant	Average Accuracy	Predictions With Correct Subsequences of Length at Least x							
		$x=3$	$x=4$	$x=5$	$x=6$	$x=7$	$x=8$	$x=9$	$x=10$
Complete PepNovo	0.727	0.946	0.871	0.800	0.654	0.525	0.411	0.271	0.193
No double neutral losses	0.720	0.932	0.857	0.786	0.621	0.493	0.389	0.250	0.189
No flanking amino acids	0.711	0.921	0.839	0.761	0.611	0.482	0.382	0.250	0.186
No pos(m) variable	0.709	0.921	0.839	0.757	0.607	0.478	0.382	0.246	0.182
No fragment dependencies	0.693	0.904	0.818	0.736	0.593	0.471	0.378	0.243	0.182
No intensity thresholds	0.679	0.886	0.796	0.714	0.579	0.464	0.382	0.246	0.186
Simple random model	0.629[1]	0.832	0.696	0.582	0.471	0.357	0.293	0.196	0.150
Dancik scoring model	0.612	0.832	0.704	0.607	0.486	0.379	0.275	0.179	0.146

The table holds cumulative results for 280 test spectra: the average accuracy of predicted amino acids and the proportions of predictions that had a correct subsequence of length at least x, for $3 \leq x \leq 10$. The different variants of PepNovo are the following:

- **Complete PepNovo** - The full scoring method with all the features described in this article.
- **No double neutral losses** - A scoring model which does not take into account the low probability fragments corresponding to double neutral losses ($y-H_2O-H_2O, b-H_2O-NH_3$, etc.).
- **No flanking amino acids** - A scoring model that does not take into account the effect of flanking amino acids on the peak intensities (the model does not include the vertices $N-aa$ and $C-aa$).
- **No fragment dependencies** - Does not model dependencies between the different fragments (corresponds to a probabilistic network without edges connecting between the vertices).
- **No intensity thresholds** - A scoring model in which peak intensities have binary values, as opposed to PepNovo's model that has four intensity values (zero, low, medium and high).
- **Simple random model** - A scoring model in which the probability of randomly observing a peak is constant for all regions in the spectrum (as opposed to the local peak density estimation that is performed by PepNovo).
- **Dancik scoring model** - This scoring model, described in Dancik *et. al.*, includes double neutral losses, but lacks all of the other components mentioned above.

[1] There is large deterioration in performance that occurs when the simple random model is used instead of PepNovo's local peak density method (the accuracy decreases by almost 10%). This dramatic decrease can be explained by the fact that we do not filter the spectra to remove many of the low intensity peaks. In the dense regions of the spectra there are many random spurious fragment matches, which cause a high score when the simple random model is used. When the spectra were filtered, the decrease in performance that occurred because of the simple random model was not as large (a decrease of only 5.5% was obtained). Note that the same filtering caused a deterioration in the complete PepNovo's results (the accuracy decreased by 5.1%).