AIomics: exploring more of the proteome using mass spectral libraries extended by AI: Supporting Information

Lewis Y. Geer[1*], Joel Lapin[2,3], Douglas J. Slotta[1], Tytus D. Mak[1], Stephen E. Stein[1]

[1]Mass Spectrometry Data Center, National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Dr., Gaithersburg, Maryland 20899, United States

[2]Department of Physics, Georgetown University, Washington, DC 20057, United States.

[3]Associate, Mass Spectrometry Data Center, National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Dr., Gaithersburg, Maryland 20899, United States

Correction to the Stein-Scott Dot Product

A known issue with the Stein-Scott dot product *S* is that the significance threshold becomes higher for spectra with very few ions. The reason for this can be understood using a geometrical argument: the size of the search space for a dot product is proportional to the number of dimensions, and since the number of dimensions is low for spectra with few ions, there are more chances to make an incorrect match. While there is no closed form expression for this threshold, we can use the decoy matches to fit an exponential approximation to the empirical threshold. In the graph below, the plotted points are the 90% highest cosine score of predicted decoy spectra matched to the query spectra from our validation set, where the x axis is the number of significant ions in the query spectra.
A non-linear least squares fit of an exponential using scipy curve_fit yields the correction given in the legend of the graph. The exponential term of this correction is subtracted from the Stein-Scott dot product to give the corrected similarity score *S*. This subtraction may result in negative values to indicate lack of significance. If negative numbers are undesirable, the score can be clipped at 0 without harm.

Figure S1. Stein-Scott Correction

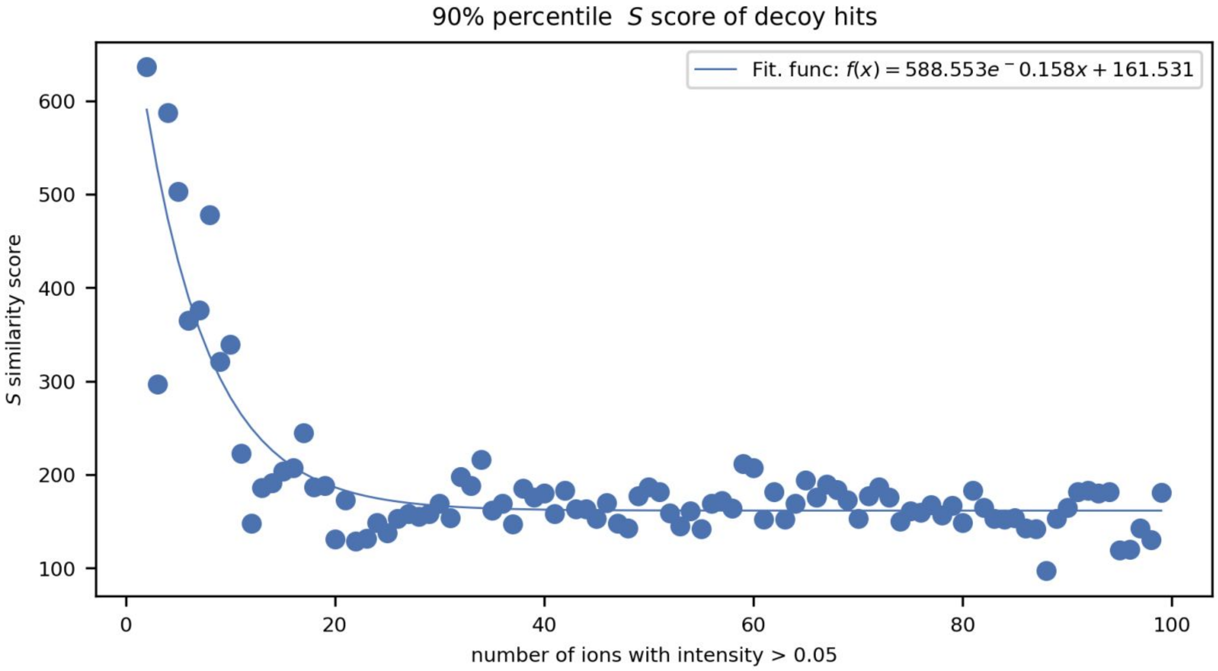

Figure S1. Plot of the 90th percentile Stein-Scott dot product of predicted decoy spectra to experimental spectra from the validation set with a given number of higher intensity ions per spectrum. An exponential function, given in the legend, is fit to the data.

The corrected Stein-Scott dot product is not normalized to the number of library spectra with matching numbers of high intensity ions. Examining how the correction is created makes this clear: for a given number of high intensity ions, we take all decoy spectra with that number of high intensity ions and take the 90% percentile of the uncorrected Stein-Scott score. Taking the 90% percentile instead of the top hit makes the correction independent of the number of peptides in the search library that have the same number of high intensity ions. This is because the underlying distribution being sampled is independent of the number of matching spectra.

Variation of the similarity score *S* over features of the test set

For the features of NCE, precursor charge, and peptide length, we computed hexagonal binning plots of the number of test spectra and their similarity scores. These plots indicate how well the network we are using predicts spectra over the entire feature range of the test set. No attempt has been made to decorrelate these features to understand their individual, uncorrelated contribution to the similarity score, but these plots indicate that the model is useful over a range of feature values typically found in proteomics experiments.

Figure S2.

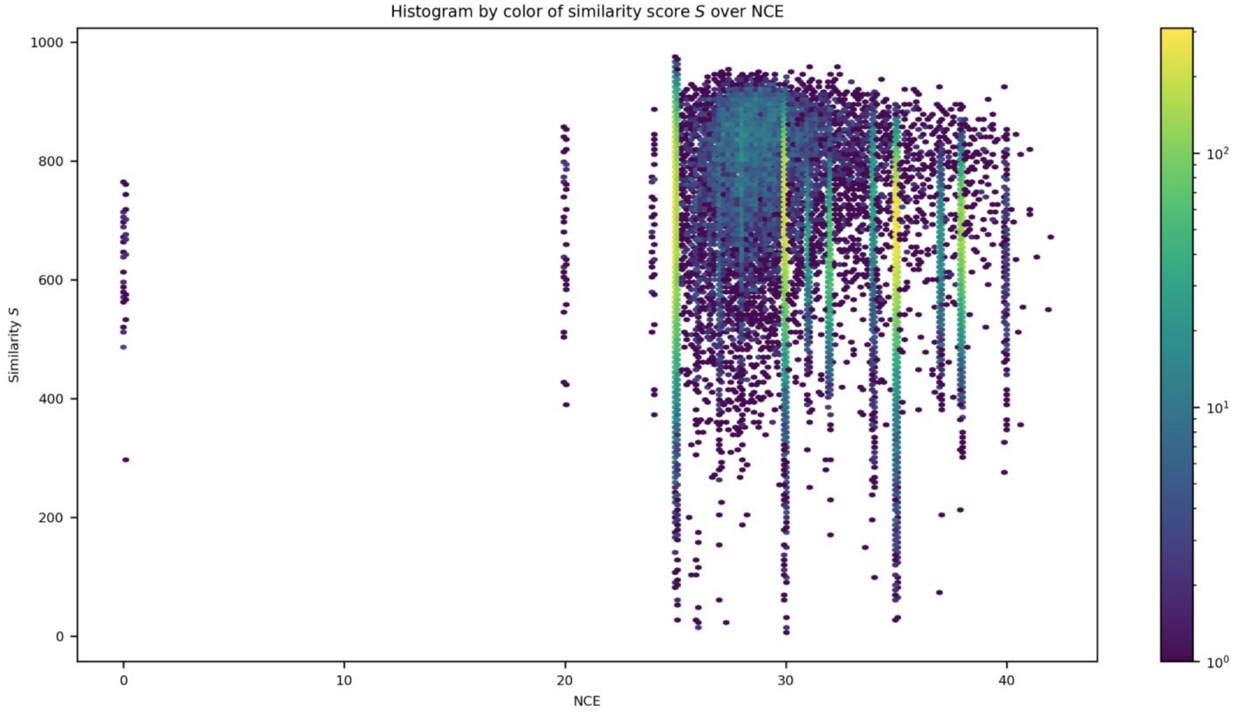

Figure S2. 2D hexagonal histogram of spectra from the test set, binned by the Stein-Scott dot product of the experimental spectra to their predicted spectra and NCE. Counts of spectra are indicated by the color legend to the right.

Figure S3.

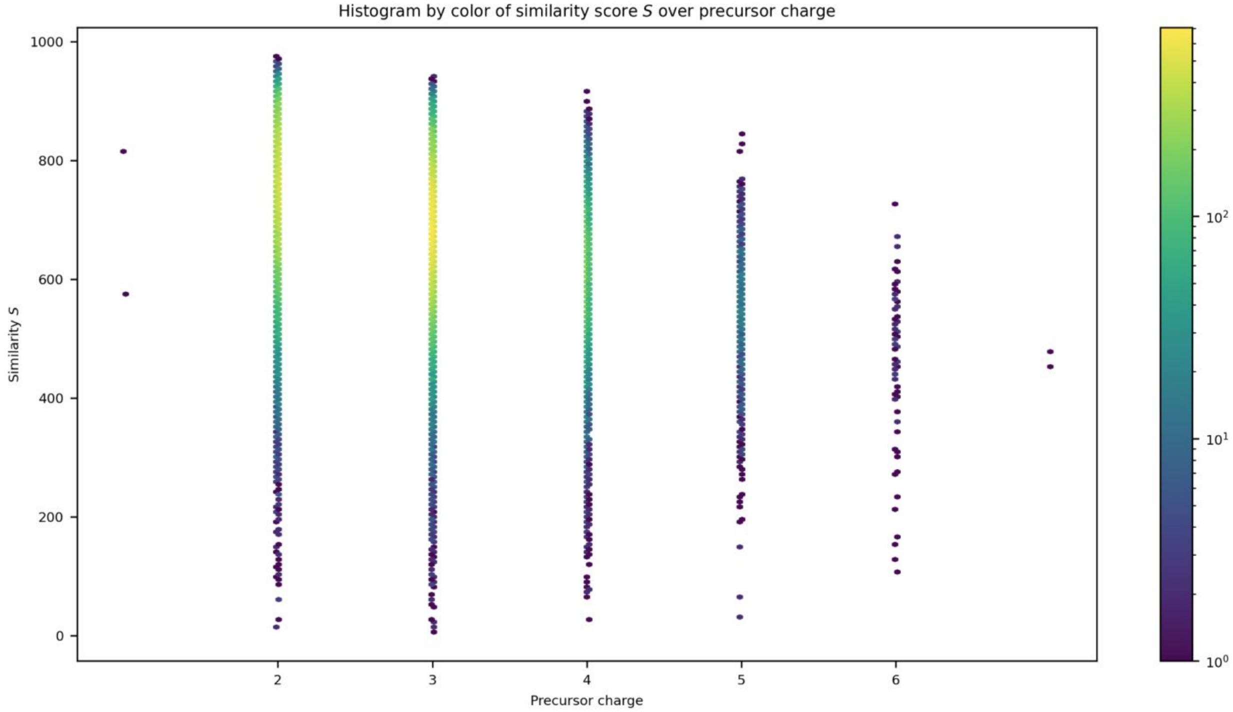

Figure S3. 2D hexagonal histogram of spectra from the test set, binned by the Stein-Scott dot product of the experimental spectra to their predicted spectra and precursor charge. Counts of spectra are indicated by the color legend to the right.

Figure S4.

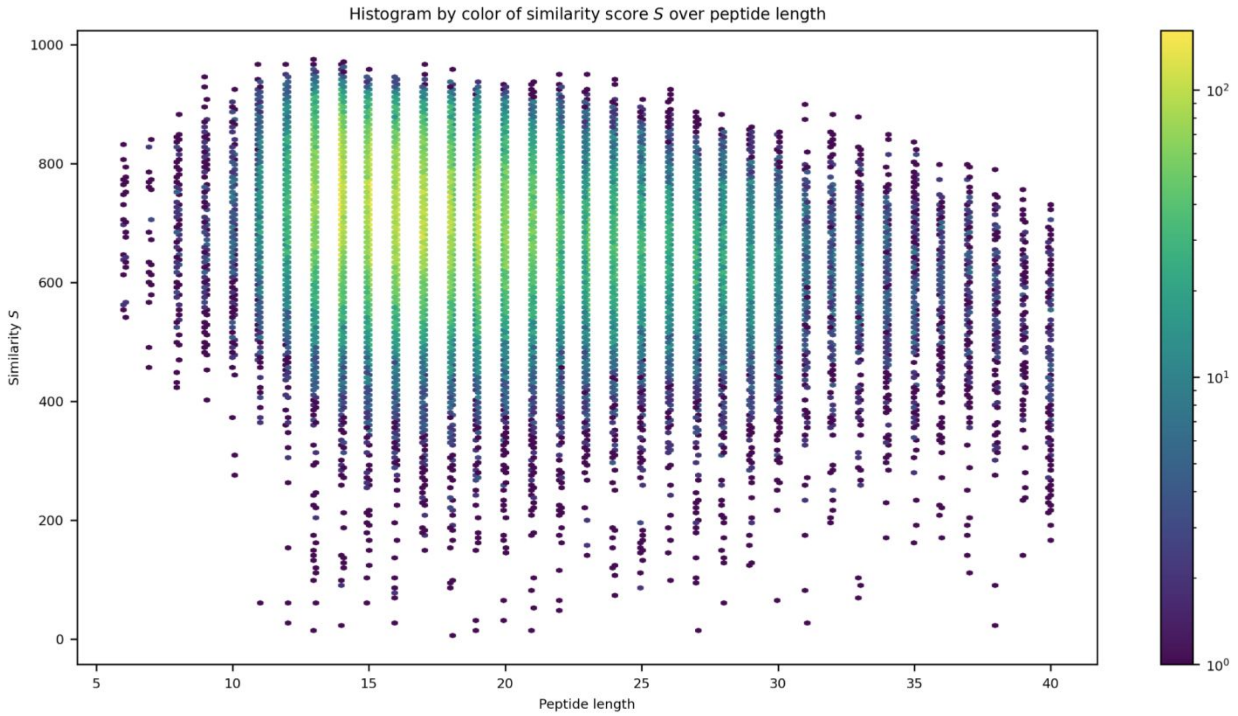

Figure S4. 2D hexagonal histogram of spectra from the test set, binned by the Stein-Scott dot product of the experimental spectra to their predicted spectra and peptide length. Counts of spectra are indicated by the color legend to the right.

Figure S5. Score histograms for predicted spectra containing only annotated ions

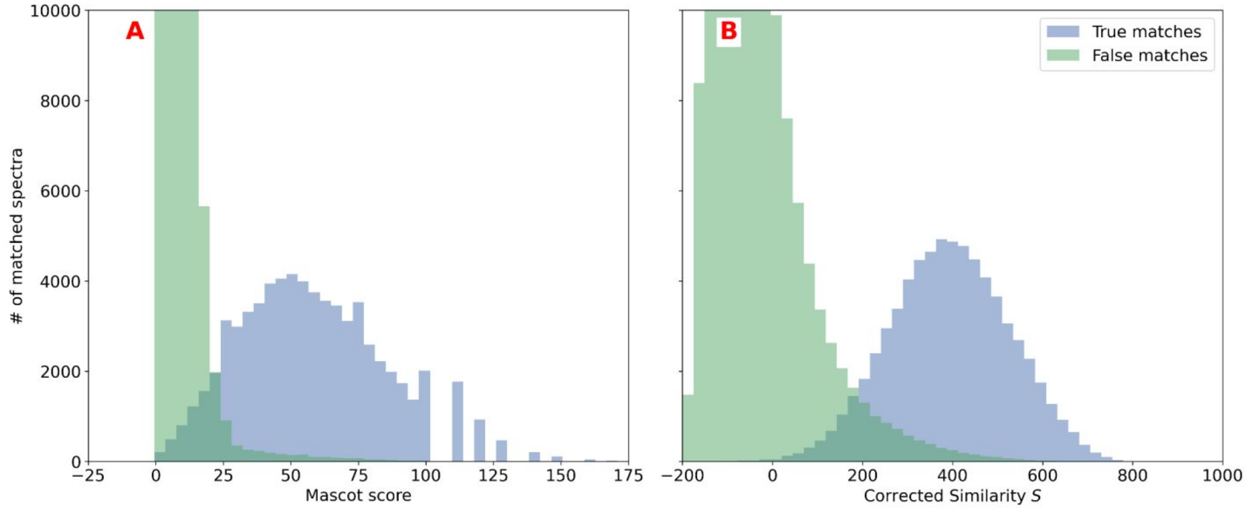

Figure S5. (A) Histogram of the Mascot ions score for both true and false matches to the test set spectra, searching against the human, mouse, and Chinese hamster proteome. (B) histogram of the corrected *S* score as applied to predicted spectra that contain only annotated ions.