MapMySection challenge task for mapping accuracy

"The winning algorithm for this category will be selected based on a quantitative assessment of accuracy of mapping to the whole mouse brain cell types from Yao et al 2023 included in the Allen Brain Cell Atlas at the "subclass" level of the hierarchy. Successful submissions will complete the mapping accuracy table that has one row for each test genetic tool and one column for each subclass found in VISp. Entries should represent the predicted fraction of cells mapping to a given subclass (e.g., rows sum to 1)."

Method for determining winner

Step 1: Organize the data into a single excel spreadsheet

The "Test Set" sheet from each of the entrants was copied to a single master sheet named by the last name of at least one of the entrants per group, with two exceptions: (1) the Razzaq/Iqbal team did not submit algorithm results and therefore were omitted and (2) Kapen submitted only the top 3 subclasses per test set, which was reformatted as frequencies and put in an appropriately formatted sheet. Finally, solutions from the "Test Set" sheet were converted into frequencies and included in a tab called "Part1_AnswerKey", with -1 values indicating a genetic tool where no SMART-seq validation was run.

Step 2: Run an R script to calculate statistics and determine winner

The script "calculate_accuracy.r" reads the solution and predictions from five entrants into R, calculates multiple (8) statistics to assess prediction accuracy, and then determines the winner by taking the average rank of the statistics. The winning entrant had an average rank of 1.5, placing first or second in 7 of 8 metrics. Importantly, for all metrics except matches_Target_top2, only the 52 out of 100 test tools with SMART-seq validation were used in determining accuracy.

The following statistics were used to assess accuracy, which were weighed equally:

1) **Cross entropy**: This is a measure of the difference between two probability distributions, quantifying the "surprise" or inefficiency of encoding one distribution using a code optimized for the other. It scales from 0 to infinity (lower is better). The average value across genetic tools was used for scoring.
2) **Mean Squares Error (MSE)** (Brier score): This is a common metric that quantifies the average of the squared differences between predicted values and actual observed values, providing a measure of the average magnitude of the errors in a set of predictions. It scales from 0 to 1 (lower is better). The average value across genetic tools was used for scoring.
3) **Pearson R squared**: This quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s) in a linear regression model, thus indicating how well the model's predictions approximate the actual data points. It scales from -1 to 1 (higher is better). The average value across genetic tools was used for scoring.

4) **Summed matched frequency** (“any_correct_hit”): This sums the total prediction frequency across all subclasses that had at least one validating cell with SMART-Seq. It scales from 0 to 1 (higher is better). The average value across genetic tools was used for scoring.
5) **Fraction of correct hits** (“fraction_correct_hits”): This is a binary call defined as TRUE (1) for genetic tools with at least 50% summed matched frequency (criteria #4) for a given method and FALSE (0) otherwise. The average value across genetic tools (corresponding to the fraction of genetic tools with correct hits) was used for scoring. A minimal offset is added to break ties.
6) **Fraction of matched top hits** (“fraction_top_match”): This is a binary call defined as TRUE (1) if predicted subclass with the highest frequency was the subclass that actually had the highest frequency and FALSE (0) otherwise. The average value across genetic tools (corresponding to the fraction of genetic tools with correctly predicted hits) was used for scoring. A minimal offset is added to break ties.
7) **Count of best subclass:** This value is the number of subclasses for which a given method is best performing (defined as having the best score in at least four of the above six criteria across all validated genetic tools for a given subclass. Thirteen of 27 subclasses had at least one validated genetic tool, and in all except “Pvalb.Gaba” there was a clear best method.
8) **Fraction of non-validation matches**: *This is the only criteria that considered the 48 genetic tools without validation data, excluding genetic tools with validation data*. This is a binary call defined as TRUE (1) if targeted subclass was one of the two predicted subclasses with highest frequency and FALSE (0) otherwise. The average value across genetic tools without validation data but with a listed cell type target was used for scoring.

These criteria were chosen to span a wide range of potential biases and definitions of “accuracy,” while still ensuring a quantitative selection of the winning entrant.