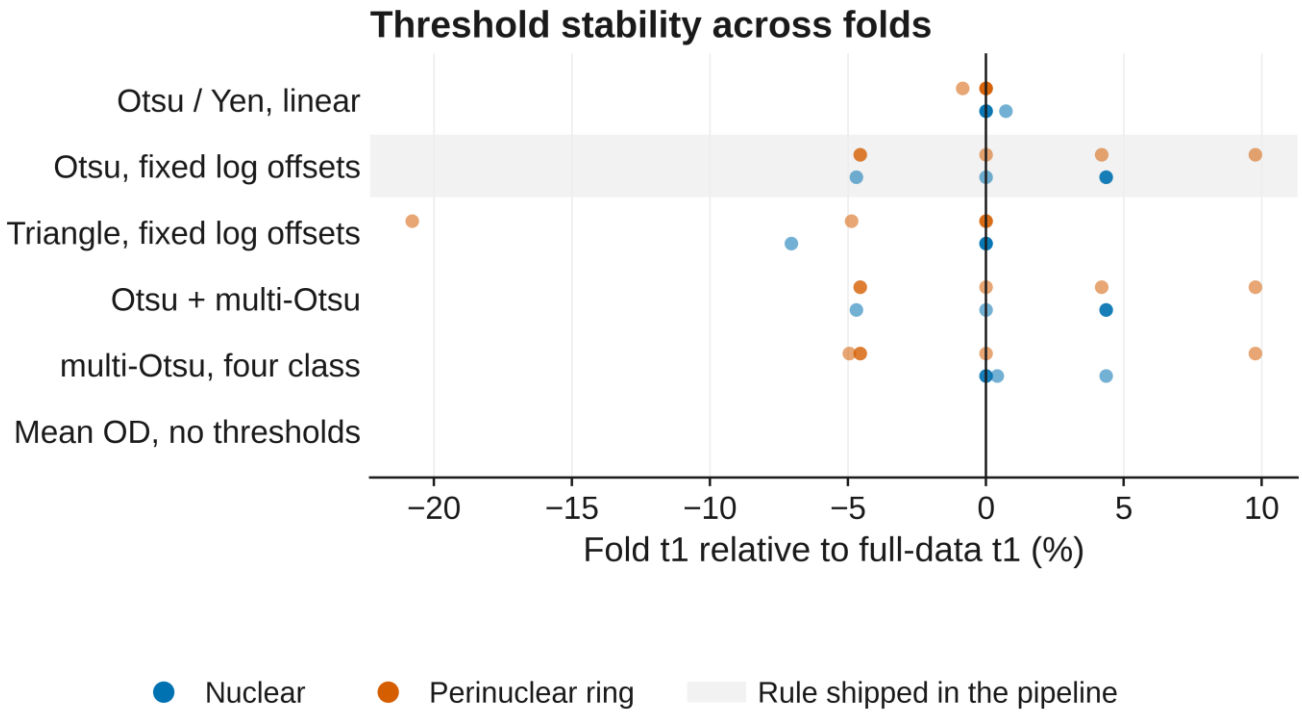


Figure S1. Stability of the first cellular-intensity threshold across cross-validation folds in the HER2-IHC-40x dataset. For each thresholding rule, the first threshold ($t1$) was fitted independently within each training fold and expressed as the percentage difference from the corresponding threshold fitted using the complete dataset. Each point represents one cross-validation fold; blue and orange points denote the nuclear and perinuclear-ring compartments, respectively. The vertical line at 0% indicates agreement with the full-data threshold, negative values indicate a lower fold-specific threshold, and positive values indicate a higher threshold. The shaded row identifies the Otsu-with-fixed-log-offsets rule implemented in the pipeline. Mean DAB optical density is shown without points because it does not require intensity thresholds. The limited fold-to-fold variation of the pipeline rule indicates threshold stability within this dataset but does not establish transferability to other staining batches, scanners, or laboratories.

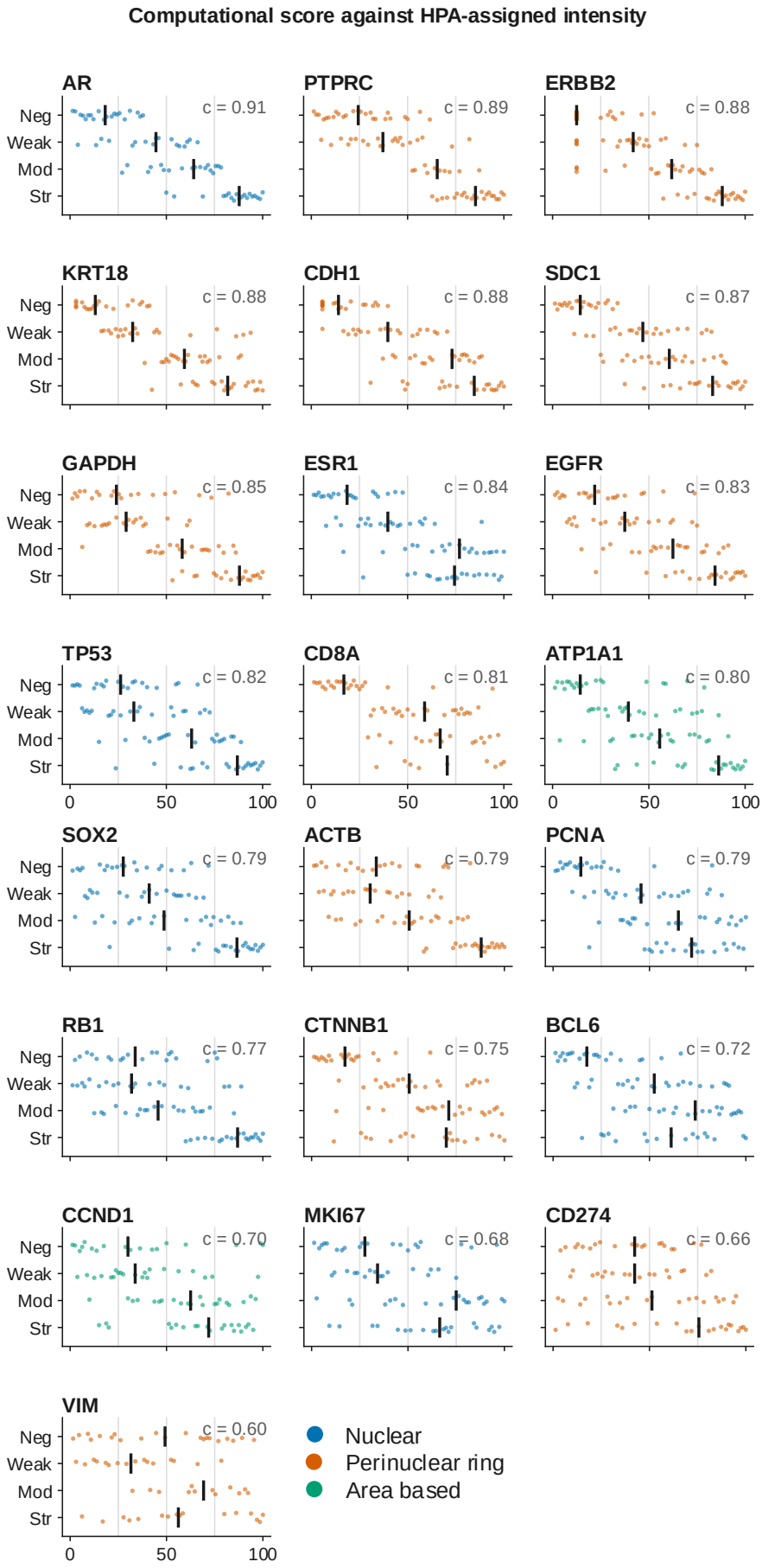


Figure S2. Computational scores by Human Protein Atlas–assigned staining intensity. Each point represents one tissue core, and black bars show class medians. Antibodies are ordered by ordinal c-index (*c*). Colors indicate nuclear, perinuclear-ring, or area-based measurement. *n* = 1,709 cores across 22 antibodies and 20 cancer types.

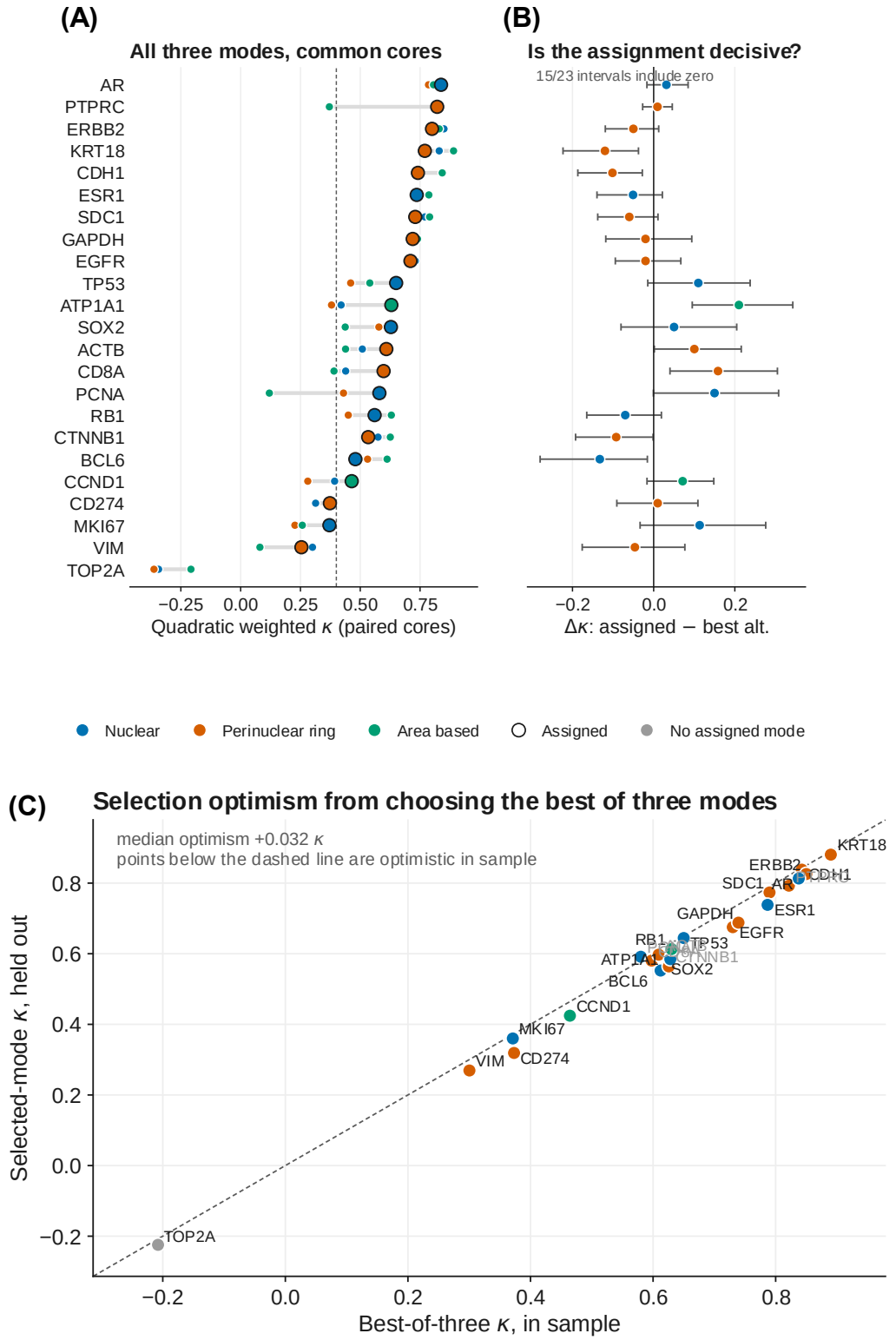


Figure S3. Effect of measurement-mode selection on agreement with Human Protein Atlas staining-intensity annotations. (A) Quadratic-weighted kappa for nuclear, perinuclear-ring, and area-based measurements, calculated using the cores available in all three modes for each antibody. Larger outlined circles indicate the prespecified measurement mode; grey symbols identify antibodies for which no mode was assigned. Horizontal lines span the results across the three modes, and the vertical dotted line marks κ = 0.40. (B) Difference in kappa between the assigned mode and the best alternative mode (Δκ = assigned − best alternative). Positive values favour the assigned mode, negative values favour an alternative mode, and horizontal error bars show 95% confidence intervals; 15 of 23 intervals included zero. Colours identify the assigned measurement mode. (C) Optimism introduced by selecting the best-performing mode using the evaluation data. In-sample best-of-three kappa is plotted against the corresponding held-out kappa after mode selection. The dashed diagonal represents equal in-sample and held-out performance; points below it indicate optimistic in-sample estimates. Median optimism was 0.032 κ. The analysis includes the 22 antibodies in the primary evaluation and TOP2A as a separately reported unresolved case.

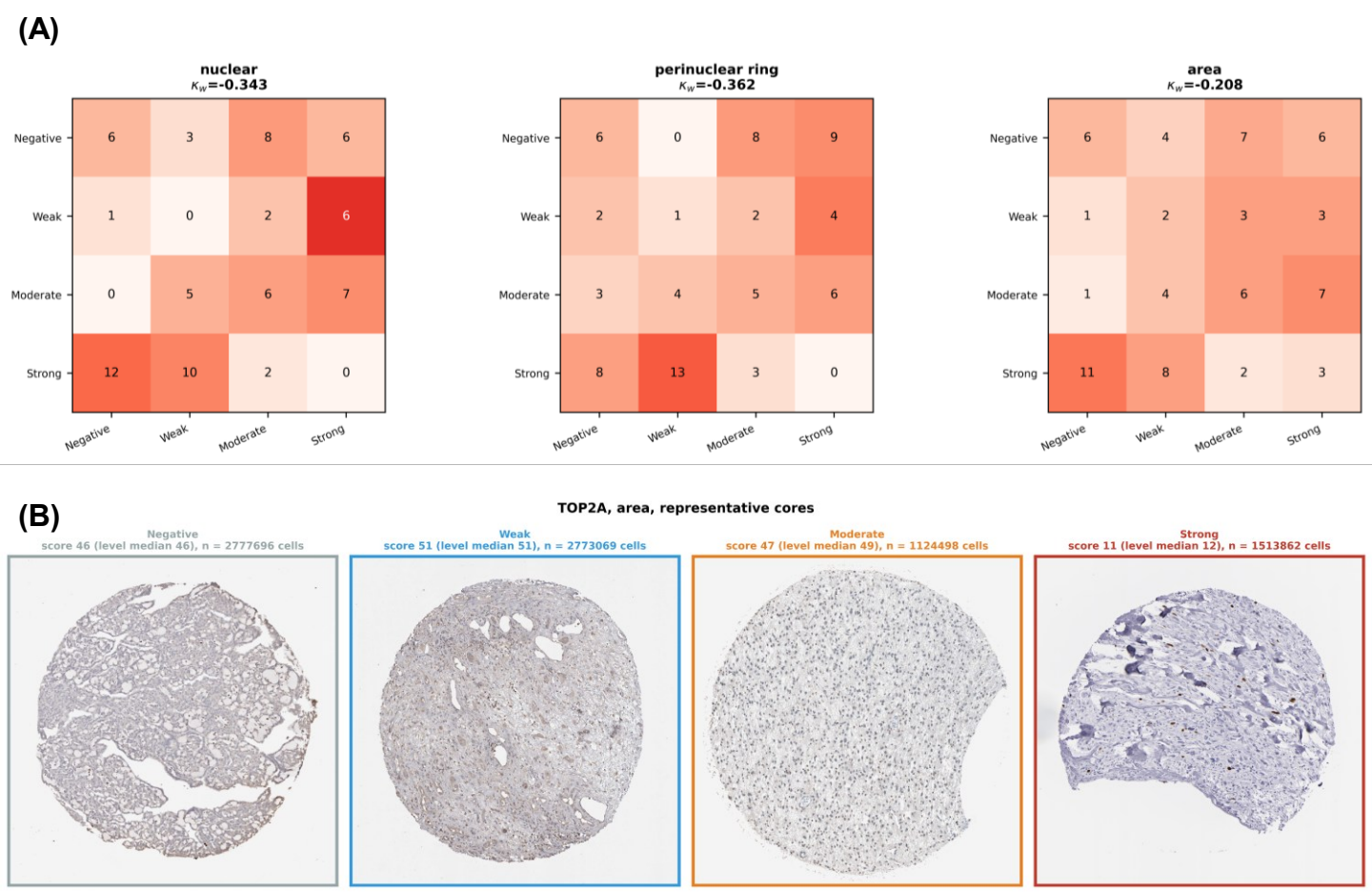


Figure S4. Unresolved TOP2A validation across measurement modes. (A) Confusion matrices comparing computationally predicted intensity classes with Human Protein Atlas (HPA)–assigned classes for nuclear, perinuclear-ring, and area-based measurements. Numbers indicate core counts. Quadratic-weighted kappa was negative for all three modes (nuclear, κw = −0.343; perinuclear ring, κw = −0.362; area based, κw = −0.208), indicating disagreement in the direction of the ordinal classes rather than successful grade separation. Rows represent HPA-assigned classes and columns represent computationally predicted classes. (B) Representative TOP2A cores from the four HPA-assigned intensity classes, selected near the median area-based computational score within each class. The computational scores did not increase with the HPA-assigned intensity, consistent with the inverted agreement observed in panel A. TOP2A was therefore not assigned a measurement mode and was reported as an unresolved limitation of chromogen optical-density scoring.

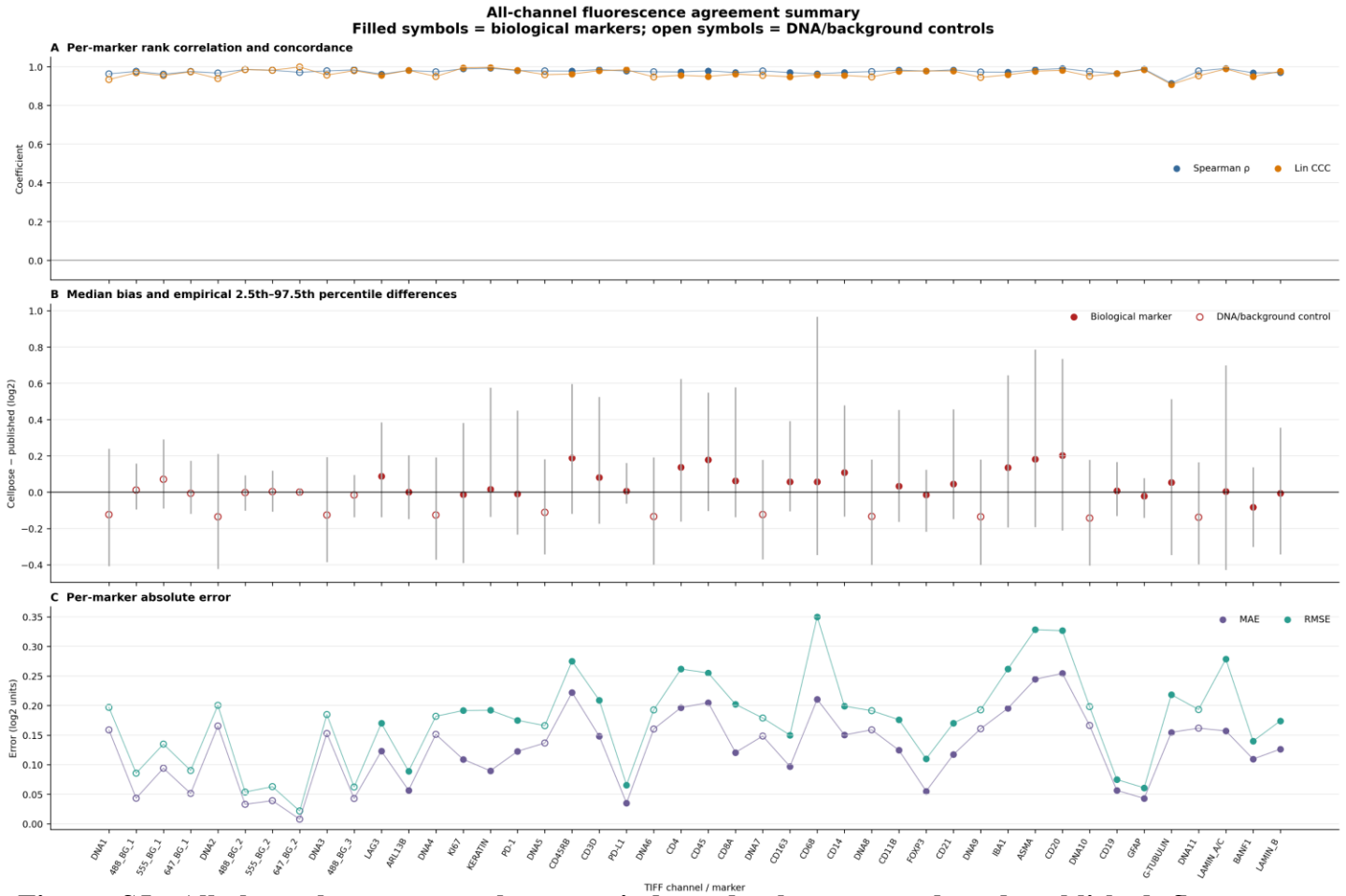


Figure S5. All-channel agreement between independently measured and published fluorescence intensities in the TONSIL-1 external validation. (A) Spearman's ρ and Lin's concordance correlation coefficient for all 44 channels. (B) Median bias (Cellpose − published, log2) with empirical 2.5th–97.5th percentile differences; filled symbols are biological markers, open symbols are DNA/background controls. Membrane and cytoplasmic markers (e.g., CD45RB, CD45, CD4, ASMA, IBA1, CD20) show consistently positive bias, while nuclear markers (e.g., BANF1, KI67, FOXP3, LAMIN_B) show near-zero or negative bias, reflecting the single nuclear mask geometry used for all channels. (C) Mean absolute error (MAE) and root-mean-square error per channel (RMSE).