

clinTALL: machine learning-driven multimodal subtype classification and treatment outcome prediction in pediatric T-ALL

Lukas Stoiber¹, Željko Antić¹, Stefano Rebellato², Grazia Fazio^{2,3},
Annika Rademacher⁴, Lennart Lenk⁴, Franco Locatelli^{5,6}, Adriana Balduzzi^{3,7},
Gunnar Cario⁴, Carmelo Rizzari⁷, Giovanni Cazzaniga^{2,3}, Jiangyan Yu^{1*},
Anke Katharina Bergmann^{1,8,9,10*}

1 Institute of Clinical Genetics and Genomic Medicine, University Hospital Würzburg, Würzburg, Germany.

2 Tettamanti Center, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy.

3 School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy.

4 Department of Pediatrics I, Pediatric Hematology and Oncology, ALL-BFM Study Group, University Medical Center Schleswig-Holstein, Campus Kiel Kiel, Germany.

5 IRCCS Bambino Gesù Children's Hospital, Rome, Italy.

6 Catholic University of the Sacred Heart, Rome, Italy.

7 Pediatrics, Fondazione IRCCS San Gerardo dei Tintori, Monza, Germany.

8 National Center for Tumor Diseases WERA (NCT WERA), Germany.

9 Comprehensive Cancer Center Mainfranken (CCC MF), Germany.

10 Bavarian Cancer Research Center (BZKF), Germany.

*Corresponding author(s). E-mail(s): Yu_J@ukw.de; Bergmann_A@ukw.de

Supplementary Methods

SHAP Values

SHAP values [1] were computed to quantify feature-level contributions to the trained TabM models (variants, clinical and gene expression modalities). For gene expression features, we used SHAP's GradientExplainer with 50 integration steps, whereas clinical and variant features were analyzed using SHAP's PermutationExplainer with max_evals = 1000. SHAP values were computed separately for the survival and classification heads of the model, and feature importance was summarized as the mean absolute SHAP value across explained samples. To improve numerical stability and reproducibility, attribution analyses were repeated across 5 folds. For each explained sample, the background set consisted of all remaining samples in a leave-one-out manner.

Transcriptional Group Analysis

Pairwise transcriptional similarity between samples was quantified using Pearson correlation across genome-wide expression profiles. For each driver class and subtype, an aggregate similarity score was computed as the mean pairwise similarity between all samples belonging to the corresponding groups. For each driver, the average similarity to its expected subtype(s) (analog to 'assigning proxy labels' in methods) was compared with the average similarity to all other subtype classes. The difference between these values defined a separation margin, reflecting how specifically a driver-associated expression profile

resembled its expected subtype context relative to unrelated subtypes. This margin was further normalized relative to the magnitude of the expected similarity to account for scale differences across drivers. An overall predictability score was then calculated by averaging subtype specificity (mean_expected) and relative margin, such that higher scores corresponded to drivers showing strong and selective association with their expected transcriptional subtype programs.

References

[1] Scott M Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems 30 (NIPS 2017)