# Consensus Machine Learning for Gene Target Selection in Pediatric AML Risk

Jenny Smith[1,8], Sean K. Maden[2,3,8], David Lee[4,8], Ronald Buie[5], Vikas Peddu[6], Ryan Shean[6], Ben Busby[7,9]

**1 Clinical Research Division, Fred Hutch, Seattle, WA**
**2 Computational Biology Program, Oregon Health & Science University, Portland, OR, 97201, USA**
**3 Dept. of Biomedical Engineering, Oregon Health & Science University, Portland, OR, 97239, USA**
**4 California Lutheran University, Thousand Oaks, CA, 91360, USA**
**5 Dept. of Biomedical Informatics and and Health Education, University of Washington, Seattle, WA, 98195, USA**
**6 Department of Virology, University of Washington Department of Laboratory Medicine, Seattle, WA, 98195, USA**
**7 National Library of Medicine, National Center for Biotechnology Information, Bethesda, MD, 20892, USA**

## Summary

Acute myeloid leukemia (AML) is a cancer of hematopoietic systems that poses high population burden, especially among pediatric populations. AML presents with high molecular heterogeneity, complicating patient risk stratification and treatment planning. While molecular and cytogenetic subtypes of AML are well described, significance of subtype-specific gene expression patterns is poorly understood and effective modeling of these patterns with individual algorithms is challenging. Using a novel consensus machine learning approach, we analyzed public RNA-seq and clinical data from pediatric AML patients (N = 137 patients) enrolled in the TARGET project.

We used a binary risk classifier (Low vs. Not-Low Risk) to study risk-specific expression patterns in pediatric AML. We applied the following workflow to identify important gene targets from RNA-seq data: (1) Reduce data dimensionality by identification of differentially expressed genes for AML risk (N = 1984 loci); (2) Optimize algorithm hyperparameters for each of 4 algorithm types (lasso, XGBoost, random forest, and SVM); (3) Study ablation test results for penalized methods (lasso and XGBoost); (4) Bootstrap Boruta permutations with a novel consensus importance metric.

We observed recurrently selected features across hyperparameter optimizations, ablation tests, and Boruta permutation bootstrap iterations, including *HOXA9* and putative cofactors including *MEIS1*. Consensus feature selection from Boruta bootstraps identified a larger gene set than single penalized algorithm runs (lasso or XGBoost), while also including correlated and predictive genes from ablation tests.

We present a consensus machine learning approach to identify gene targets of likely importance for pediatric AML risk. The approach identified a moderately sized set of recurrent important genes from across 4 algorithm types, including genes identified

[8]These authors contributed equally
[9]Lead Contact, Correspondence: ben.busby@gmail.com

across ablation tests with penalized algorithms (*HOXA9* and *MEIS1*). Our approach mitigates exclusion biases of penalized algorithms (lasso and XGBoost) while obviating arbitrary importance cutoffs for other types (SVM and random forest). This approach is readily generalizable for research of other heterogeneous diseases, single-assay experiments, and high-dimensional data. Resources and code to recreate our findings are available online.

# Introduction

Acute leukemia is the most prevalent childhood cancer, accounting for 30% of childhood cancers overall [Arber et al., 2016, Tarlock and Cooper, 2019]. Major subtypes of pediatric acute leukemia include acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), accounting for 15% and 85% of these leukemia cases, respectively [Tarlock and Cooper, 2019]. Despite improving survival rates, pediatric AML remains deadlier than ALL [Tarlock and Cooper, 2019]. AML is a heterogeneous cancer of the blood and bone marrow myeloid stem cells that presents with numerous molecular subtypes actionable for stratification and treatment. These subtypes are often based on cytogenetics, molecular data, and other characteristics [Ley et al., 2013, Bolouri et al., 2018]. By contrast to adult AML, pediatric AML is characterized by rare somatic mutations, absence of common adult AML mutations, and relatively frequent structural variants [Bolouri et al., 2018]. These findings indicate the importance of age-based targeted therapies for AML treatment, and the potential for molecular assays to further our understanding of how gene expression relates to pediatric AML risk, prognosis, and treatment. We utilized RNA-seq expression data to better understand its relation to pediatric AML risk, which remains poorly understood.

Interest in identification of biomarker and gene target sets of cancer risk using RNA-seq data has endured for over a decade [Saeys et al., 2007]. For statistical rigor and clinical utility, reduction of high-dimensional, whole-genome expression sets of tens of thousands of genes is vital. Differential gene expression (DEG) analysis is typically used to achieve dimensionality reduction by selecting loci with maximal expression contrast between sample groups. This is typically followed by fitting and optimization of models to these reduced sets of DEGs, further narrowing focus to loci showing the greatest contrast and most predictive qualities between sample sets. For the present work, we consider this cumulative process of dimensionality reduction, model fitting, and optimization as a problem of gene feature selection.

Selection of important genes from expression data remains challenging for biomedical research, partly because the commonly applied cross-sectional case/control study design confounds results interpretability. Further, underlying biological dynamics can be nuanced and complex in disease processes, especially for molecularly heterogeneous cancers like AML. These problems can be tractable with modern machine learning approaches, which include the recently developed eXtreme Gradient Boosting (XGBoost) algorithm and Boruta permutation method [Kursa and Rudnicki, 2010, Chen et al., 2019]. With computational advances, these and other methods are more robust, efficient, and accessible to quantitative researchers than ever before. However, these improvements don't address the need to reconcile disparate findings from applying multiple distinct algorithm types to biomedical data. For this task, it is useful to devise a formalized consensus approach that leverages feature importance metrics across algorithms to arrive at a consensus important feature set. Far from straightforward, development and formalization of consensus feature selection methods with machine learning presents its own challenges. Researchers must reconcile results interpretability, model performance variations, and selection of important features across disparate algorithms and their respective assumptions, strengths, and weaknesses. Further, vital

properties of consensus feature selection methods, especially best practices for their use, have yet to be established for biomedical research. Nevertheless, development of such methods is warranted and could become a boon for biomedical research.

The present work is a starting point for addressing aforementioned obstacles for identifying consensus important gene features that help elucidate how gene expression differences relate to pediatric AML risk. We used clinical and RNA-seq data from pediatric AML samples (N = 137 patients) provided by the TARGET consortium. We focused on achieving consensus from 4 distinct algorithms, including lasso, random forest, support vector machines (SVM), and XGBoost [Friedman et al., 2010, Liaw and Wiener, 2002, Meyer et al., 2018]. These represent a variety of algorithm types, each with distinct assumptions, strengths, and weaknesses. Random forest and SVM do not natively differentiate important from non-important features, necessitating an importance or weight cutoff be set to identify the most important features. By contrast, lasso and XGBoost perform penalized regression and ensemble learning, respectively, which returns greatly restricted feature subsets, though at the cost of feature exclusion bias (see Results). We addressed these issues by bootstrapping Boruta permutations with a novel consensus importance metric based on relative feature importance rank across these 4 algorithms.

# Results

## Pediatric AML Risk Group Demographics

This study focused on whether gene expression could be used to predict pediatric AML risk, as defined using the classical cytogenetic and molecular classification scheme [Arber et al., 2016]. We initially identified TARGET pediatric AML patients with primary blood or bone cancer samples (N = 137 patients) and defined a binarized risk group (BRG) classifier as either low risk or not-low risk, where the latter category combines "standard" and "high" risk patients (Figure 1). Summary statistics indicated binarized risk was approximately balanced for important demographic characteristics, including age at first diagnosis, gender, and bone marrow leukemic blast percentage and peripheral blasts (see also Table S1). However, the "not-low" risk group had a significantly lower median for white blood cell count at diagnosis (29.3 [range: 1.30-519] in not-low versus 53.5 [range:1.60-253] in low-risk, p = 0.032). We randomly divided samples into training (N = 96 samples) and test (N = 49 samples) subsets, at a ratio of 2:1, preserving BRG sample group frequencies in each subset. We used training data to calculate differentially expressed genes (DEGs), and the train and test set classifications to fit and assess fitted models below.

## Dimensionality Reduction with Differentially Expressed Genes (DEGs)

We pre-filtered the RNA-seq gene expression dataset to limit the number of features included in the initial model training. Using the training dataset, gene expression for standard or high risk patient (not-low risk, BRG = 1, N = 55 samples) were contrasted to patients at low risk (low risk group, BRG = 0, N = 38 samples) using differential expression analysis. From approximately 60,000 genes assayed, we identified 1,984 differentially expressed between risk groups (—log2FC— <1, p-adj. >0.05, Figure 1B and 1C, see also Table S2, Methods). This increased the mean of normalized expression differences from 0.50 to 1.71 (median increase from 0.32 to 1.51). Mean of variance differences also increased from 0.76 to 2.19 (median increase from 0.31 to 2.19).

## Algorithm Hyperparameter Optimization

We performed hyperparameter optimization with four distinct algorithm types (lasso, random forest, SVM, and XGBoost) to determine optimal values to use in following ablation and consensus tests (Figure 2, Table 1). These algorithms include two ensemble methods (random forest and XGBoost) two penalized methods (XGBoost and lasso) and two unpenalized methods (SVM and random forest). These algorithms quantify feature importance in the following ways: 1. Lasso assigns beta-value coefficient (positive, negative, or null/0) for use in penalized regression; 2. SVM assigns a feature weight (positive or negative) for inclusion in kernel-based estimator; 3. XGBoost assigns importance (positive or null/0) from gain across splits; 4. Random forest assigns importance using mean decrease in Gini index (positive value). For each algorithm, we tested at least 3 distinct hyperparameter value sets (Table 1 column 3), and compared model performances. We observed a variety of model performance fluctuations across optimizations for each algorithm (see also Supplemental Information). All fitted XGBoost and lasso models showed uniformly high performance. For SVM, radial kernel tests showed worse performance than linear kernel tests. Where there were clear performance benefits, we selected the optimal hyperparameter sets for inclusion in our consensus importance metric (Figure 4, see also Supplemental Information).

## Ablation Tests and Exclusion Bias with Lasso and XGBoost

Unlike Random Forest and SVM, lasso and XGBoost penalize uninformative and/or correlated features, resulting in 0 or null importance assignment for most genes. Lasso assigns a beta-value coefficient for regression, and XGBoost estimates gain from fractional contributions to splits. Feature omission can reduce data dimensionality and overfitting risk, though this is likely not optimal in biomedical research settings where the objective is to identify a set of gene targets. Exclusion of correlated features can obscure gene sets or pathways of importance, constituting an omission bias. We performed ablation studies using lasso and XGBoost. For each iteration of ablation, we excluded all features selected from prior iterations before refitting lasso or XGboost models, respectively (Figure 3, Figures S1 and S2, Methods).

In the absence of an omission bias, we expected consistent decline in fitted model performance with successive ablation iterations. Instead, we observed oscillation between performance recovery and decline across successive ablation iterations, with gradual performance decline across 15 and 70 ablation iterations of lasso and XGBoost, respectively (Figure 3C and Figure S1). Interestingly, models fitted in later iterations could recover substantial performance, and this trend was even more exaggerated for XGBoost than lasso ablation iterations. This trend likley reflects signal gain and loss of alternative predictive and related or correlated gene sets and pathways, which are unrepresented in sets from earlier ablation iterations. As iteration increases, gene members of alternate functional sets may be successively selected then exhausted, resulting in initial performance recovery followed by successive performance loss. These findings highlight the importance of carefully evaluating iterations of penalized methods in biomedical research, and the utility of ablation tests.

We observed substantial correlated expression, both positive and negative, across genes selected in the first 3 and 4 ablation iterations for lasso and XGboost, respectively (Figure 3C and Figure S2). Correlated expression could result from direct or indirect functional interactions or relatedness. We observed evidence for functional similarity across these selected gene sets, especially shared HOX pathway membership. Surprisingly, *HOXA9* was selected in the first iteration of lasso ablation, but not until the fourth iteration of XGBoost ablation (Figure 3A, Figure S2). *HOXA9* is known to be co-expressed in multiple pediatric AML subtypes, and its activity can be used to

predict patient risk [Brumatti et al., 2013, Collins and Hess, 2016b] (Figure 3A). We further note substantial positive correlation between *HOXA9* and the HOX family gene *MEIS1*, which was selected in iteration 3 of lasso ablations. *MEIS1* expression is linked to hematopoietic stem cell development, and *HOXA9-MEIS1* complexes were found to correlate with AML subtype and outcome [Rozovskaia et al., 2001, Wang and Kamps, 2007, Collins and Hess, 2016a, Mohr et al., 2017].

## Consensus Important Gene Features from Boruta Permutation Bootstraps

We designed and applied a consensus machine learning algorithm to identify recurrent important gene features. We used a consensus importance metric ("nrank", Figure 4, Figure S3-S10), which returns a normalized median absolute rank after calculating the algorithm-specific importance metrics from lasso, random forest, XGBoost, and SVM. We then permuted this calculation in the Boruta method for 1,000 bootstraps, with redraw of 2/3rds of pediatric AML samples in each bootstrap, to simulate redraw of the training sample subset ( [Kursa and Rudnicki, 2010], Supplemental Information). For comparison, we also used single-algorithm importance for 4 algorithms in Boruta permutations across 1,000 bootstraps apiece (Supplemental Information). Across these tests, we evaluated importance calculation histories (Figure S4), gene-wise summaries of label assignments (either "rejected","tentative", or "confirmed" in each Boruta bootstrap iteration, Figures S5-S9), and finally extent of consensus across Boruta bootstrap runs using each respective importance metric (Figure 4C, Figure S10).

To understand these results, it is necessary to summarize the Boruta method. In each permutation, this method calculates observed importance for "real" features and importance distribution of "shadow" features. Shadow features are obtained by random reassignment of expression values to samples, which breaks correlation of expression with class (AML risk). Real features are rejected if their importance is sufficiently similar to the shadow feature importance distribution. Remaining features are then retained in following permutations. Ultimately a label of "rejected" (non-important), "tentative" (marginal features), or "confirmed" (high-confidence important features) to each gene feature (Boruta citation). Evaluating real and shadow feature importance across Boruta permutations for a sampling of bootstraps, we observed a range of behavior across the various importance metrics used (Figure S4). XGBoost showed no exclusion of rejected features across permutations. By contrast, the remaining methods, including nrank consensus, showed progressive retention of confirmed or tentative features and omission of rejected ones.

We studied recurrent selected genes in each test by setting progressively more stringent cutoffs (e.g. gene was labeled confirmed in >1, >20% or 200/1,000 bootstraps, or >50% or 500/1,000 bootstraps). We observed a range in the total sizes of confirmed feature sets across runs, and total recurrent confirmed feature sets from the consensus nrank run fell in the middle of this range. Interestingly, about 50% of consensus nrank confirmed features overlapped with recurrent confirmed features from random forest, SVM, and lasso runs, though not XGBoost (Figure 4C, Figure S10). Certain confirmed genes, including *HOXA9* and *MEIS1*, were present in the final recurrent confirmed gene set (Table 2). These genes were identified in the first 4 ablation iterations with lasso and XGBoost, and their inclusion in the consensus gene set indicates our approach mitigates exclusion bias of independent penalized algorithms.

## Discussion

We present a consensus feature selection strategy, including a novel consensus rank importance metric and implementation with bootstraps of Boruta permutations. This consensus approach can mitigate possible algorithm feature exclusion biases of penalized algorithms (lasso and XGBoost) while obviating the need to set arbitrary importance cutoffs with algorithms not natively performing feature selection (random forest and SVM). Prior studies characterized numerous molecular subtypes in pediatric AML, reflecting heavy utilization of whole genome sequencing, methylation, and other assay types, with less utilization of RNA-seq gene expression data ( [Bolouri et al., 2018, Ley et al., 2013]). We focused primarily on data from RNA-seq, which may be underutilized for characterization of pediatric AML and clinical risk. Our consensus gene feature set validates prior literature and demonstrates how a single-assay approach can be used to characterize clinical risk in a molecularly heterogeneous cancer.

Among consensus important genes for pediatric AML risk, we identified numerous potential therapeutic targets. *HOXA9* has been implicated in MLL (KMT2A)-rearranged AMLs and *MLL-HOXA9* fusion has been shown to induce leukemogenesis in xenograph and mouse models [Milne, 2017]. Interestingly, *SLTRK5* and *ITGB3* are both highly and aberrantly expressed on the cell-surface. This suggests these genes may be good potential targets for antibody and CAR-T cell therapies. *SLTRK5* has been shown to be aberrantly expressed in nearly 80% of AML and coincides with high/standard risk clinical features, allowing one the potential to improve outcomes for AMLs with poor prognosis [Miller et al., 2013].

Our consensus machine learning approach can and should be formalized and fine-tuned for better performance, efficacy, and generalizability. This can be achieved in several possible ways, including iterative recalculation of DEGs, bootstrapping in hyperparameter optimization, and inclusion of alternative consensus importance metrics besides the "nrank" method used here (Results, Supplemental Information). Finally, best practices for consensus feature selection, such as minimal data size or optimal test/train split for a given effect magnitude, have yet to be established for biomedical data. To this end, our results here are promising, and we have provided sufficient notebooks and scripts such that our consensus method can be generalized to research other diseases or datasets.

## Limitations of the Study

While the present work describes a thorough application of algorithm and machine learning to elucidate expression-based gene targets in AML risk, we must note certain limitations. The TARGET pediatric AML dataset is moderate in size, and further limited when training and test subsets are used, increasing risk of model overfitting. We note that our consensus approach could mitigate the effects of individual model overfitting because its output is a gene list rather than a fitted model.

## Author Contributions

Conceptualization, Funding Acquisition, and Supervision, B.B.; Methodology, Investigation, and Writing - Original Draft, J.S., S.K.M., and D.L.; Writing - Review & Editing, J.S., S.K.M., R.B., V.P., and R.S.

## Acknowledgments

We thank organizers and participants of NCBI Hackathons for supporting this project with feedback and input from its inception.

## Declaration of Interests

The authors declare no competing interests.

## References

## References

Arber et al., 2016. Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., Bloomfield, C. D., Cazzola, M., and Vardiman, J. W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405.

Bolouri et al., 2018. Bolouri, H., Farrar, J. E., Triche, T., Ries, R. E., Lim, E. L., Alonzo, T. A., Ma, Y., Moore, R., Mungall, A. J., Marra, M. A., Zhang, J., Ma, X., Liu, Y., Liu, Y., Auvil, J. M. G., Davidsen, T. M., Gesuwan, P., Hermida, L. C., Salhia, B., Capone, S., Ramsingh, G., Zwaan, C. M., Noort, S., Piccolo, S. R., Kolb, E. A., Gamis, A. S., Smith, M. A., Gerhard, D. S., and Meshinchi, S. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nature Medicine*, 24(1):103–112.

Brumatti et al., 2013. Brumatti, G., Salmanidis, M., Kok, C. H., Bilardi, R. A., Sandow, J. J., Silke, N., Mason, K., Visser, J., Jabbour, A. M., Glaser, S. P., Okamoto, T., Bouillet, P., D'Andrea, R. J., and Ekert, P. G. (2013). HoxA9 regulated Bcl-2 expression mediates survival of myeloid progenitors and the severity of HoxA9-dependent leukemia. *Oncotarget*, 4(11):1933–1947.

Chen et al., 2019. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and implementation), X. c. b. X. (2019). xgboost: Extreme Gradient Boosting.

Collins and Hess, 2016a. Collins, C. T. and Hess, J. L. (2016a). Deregulation of the HOXA9/MEIS1 Axis in Acute Leukemia. *Current opinion in hematology*, 23(4):354–361.

Collins and Hess, 2016b. Collins, C. T. and Hess, J. L. (2016b). Role of HOXA9 in leukemia: dysregulation, cofactors and essential targets. *Oncogene*, 35(9):1090–1098.

Friedman et al., 2010. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Kursa and Rudnicki, 2010. Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(1):1–13.

Ley et al., 2013. Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., Robertson, A. G., Hoadley, K., Triche, T. J., Laird, P. W., Baty, J. D., Fulton, L. L., Fulton, R., Heath, S. E., Kalicki-Veizer, J., Kandoth, C., Klco, J. M.,

Koboldt, D. C., Kanchi, K.-L., Kulkarni, S., Lamprecht, T. L., Larson, D. E., Lin, L., Lu, C., McLellan, M. D., McMichael, J. F., Payton, J., Schmidt, H., Spencer, D. H., Tomasson, M. H., Wallis, J. W., Wartman, L. D., Watson, M. A., Welch, J., Wendl, M. C., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y., Chiu, R., Chu, A., Chuah, E., Chun, H.-J., Corbett, R., Dhalla, N., Guin, R., He, A., Hirst, C., Hirst, M., Holt, R. A., Jones, S., Karsan, A., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, K., Parker, J., Pleasance, E., Plettner, P., Schein, J., Stoll, D., Swanson, L., Tam, A., Thiessen, N., Varhol, R., Wye, N., Zhao, Y., Gabriel, S., Getz, G., Sougnez, C., Zou, L., Leiserson, M. D. M., Vandin, F., Wu, H.-T., Applebaum, F., Baylin, S. B., Akbani, R., Broom, B. M., Chen, K., Motter, T. C., Nguyen, K., Weinstein, J. N., Zhang, N., Ferguson, M. L., Adams, C., Black, A., Bowen, J., Gastier-Foster, J., Grossman, T., Lichtenberg, T., Wise, L., Davidsen, T., Demchok, J. A., Shaw, K. R. M., Sheth, M., Sofia, H. J., Yang, L., Downing, J. R., and Eley, G. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine*, 368(22):2059–2074.

Liaw and Wiener, 2002. Liaw, A. and Wiener, M. (2002). Classication and Regression by randomForest. 2:6.

Meyer et al., 2018. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., C++-code), C.-C. C. l., and C++-code), C.-C. L. l. (2018). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Miller et al., 2013. Miller, P. G., Al-Shahrour, F., Hartwell, K. A., Chu, L. P., Järås, M., Puram, R. V., Puissant, A., Callahan, K. P., Ashton, J., McConkey, M. E., Poveromo, L. P., Cowley, G. S., Kharas, M. G., Labelle, M., Shterental, S., Fujisaki, J., Silberstein, L., Alexe, G., Al-Hajj, M. A., Shelton, C. A., Armstrong, S. A., Root, D. E., Scadden, D. T., Hynes, R. O., Mukherjee, S., Stegmaier, K., Jordan, C. T., and Ebert, B. L. (2013). In Vivo RNA Interference Screening Identifies a Leukemia-Specific Dependence on Integrin Beta 3 Signaling. *Cancer cell*, 24(1):45–58.

Milne, 2017. Milne, T. A. (2017). Mouse models of MLL leukemia: recapitulating the human disease. *Blood*, 129(16):2217–2223.

Mohr et al., 2017. Mohr, S., Doebele, C., Comoglio, F., Berg, T., Beck, J., Bohnenberger, H., Alexe, G., Corso, J., Ströbel, P., Wachter, A., Beissbarth, T., Schnütgen, F., Cremer, A., Haetscher, N., Göllner, S., Rouhi, A., Palmqvist, L., Rieger, M. A., Schroeder, T., Bönig, H., Müller-Tidow, C., Kuchenbauer, F., Schütz, E., Green, A. R., Urlaub, H., Stegmaier, K., Humphries, R. K., Serve, H., and Oellerich, T. (2017). Hoxa9 and Meis1 Cooperatively Induce Addiction to Syk Signaling by Suppressing miR-146a in Acute Myeloid Leukemia. *Cancer Cell*, 31(4):549–562.e11.

Rozovskaia et al., 2001. Rozovskaia, T., Feinstein, E., Mor, O., Foa, R., Blechman, J., Nakamura, T., Croce, C. M., Cimino, G., and Canaani, E. (2001). Upregulation of ***meis1*** and ***hoxa9*** in acute lymphocytic leukemias with the t(4 : 11) abnormality. *Oncogene*, 20(7):874–878.

Saeys et al., 2007. Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–2517.

Tarlock and Cooper, 2019. Tarlock, K. and Cooper, T. M. (2019). Acute myeloid leukemia in children and adolescents - UpToDate.

Wang and Kamps, 2007. Wang, G. G. and Kamps, M. P. (2007). Survival signaling in HoxA9/Meis1 AML. *Blood*, 109(9):3619–3620.

# Table and Figure Legends

## Table 1

**Results of hyperparameter optimization tests for 4 algorithms (SVM, lasso, Random Forest, XGBoost). Performance metrics of each fitted model are shown for true negatives (TN), false negatives (FN), true positives (TP), false positives (TP), true positive rate (TPR), true negative rate (TNR), false discovery rate (FDR), false omission rate (FOR), log loss (LL), and error (E). Model evaluations were performed on test data subset after models were fitted to training data.**

## Table 2

**Recurrent consensus important gene features from label assignments of bootstrapping (N = 1,000 bootstraps) Boruta permutations with consensus importance (nrank). Columns 1-3 show gene info, and 4-6 show extent of recurrence where "nrank.conf.n1" is confirmed in ¿= 1 bootstrap, "nrank.conf.5perc" is confirmed in ¿= 5% (or N = 50) bootstraps, and "nrank.conf.20perc" is confirmed in ¿20% (or 200) bootstraps.**

## Figure 1

**Clinical and demographic information summary for TARGET pediatric AML dataset. A. Survival in Low (binary risk group, BRG = 0), compared to Not-Low (BRG=1) clinical risk group. B Volcano plot of differentially expressed genes (DEGs), x-axis is log2 fold-change, y-axis is -1 times log10 of unadjusted p-value from t-tests, significance threshold (horizontal line) set at ¡0.01 p-adjusted and (vertical lines) —log2FC—¿1. C Heatmap of DEG expression (Z-score of normalized expression) with sample-wise clinical annotations ("cto" is primary ctogenetic subtype).**

## Figure 2

**Results of hyperparameter optimizations. A Results of 3 lasso iterations varying alpha from 0.5-1.2 (shows all genes with not-null coefficients). B SVM 4 iterations, varying linear and radial kernel, and no weight filter versus top 50% weight filter (shows features with top 99th quantile absolute weight). C XGBoost ("XGB") 3 iterations, varying steps (shows all with not-null importance). D Random forest ("RF") 3 iterations varying ntrees from 5-15k (shows top 99th quantile importance).**

## Figure 3

**Lasso ablation test results. A. Beta-value coefficients (non-zero) from first lasso iteration. B. Correlation of selected gene feature expression (Spearman Rho, whole dataset) from iterations 2 and 3 with iteration 1 features expression. C. Fitted model performance across 15 ablation iterations, showing (top) true positive (TPR, red) and true negative (TNR,**

**blue) rates, and (bottom) false discovery (FDR, green) and false omission (FOR, purple).**

### Figure 4

**Methods to determine consensus important gene features from Boruta bootstraps (N = 1,000). A Workflow calculating "nrank" consensus importance, or normalized median absolute importance rank, across 4 algorithms (lasso, SVM, random forest, and XGBoost). B Feature (green is confirmed, red is rejected, yellow is tentative) and shadow feature-wise (blue lines) importance (rank, y-axis) across Boruta permutations (x-axis, max = 100). C Upset plot of recurrent confirmed features (present in at least 20% or 200/1,000 bootstraps) across Boruta bootstrap analyses with 5 distinct importance metrics (XGB = XGBoost importance, SVM = SVM importance, Nrank = consensus importance nrank, Lasso = lasso importance, RF = random forest importance). Red genes and data are shared across consensus, lasso, and random forest runs, purple is confirmed genes unique to consensus run, blue is confirmed genes shared only by consensus and lasso runs.**

## Supplemental Information

Please see Supplemental Information document for methods details and supporting figures and tables.

### Figure S1

**Correlation of gene expression (entire dataset) across gene features selected in iteration 1 vs. iterations 2-4 of XGBoost ablation tests (see Results, Supplemental Information).**

### Figure S2

**XGBoost fitted model performances across 70 iterations of ablation. (Top) True positive (TPR) and true negative rate (TNR). (Bottom) False discovery (FDR) and false omission rate (FOR).**

### Figure S3

**Recurrent important genes from Boruta bootstraps with 5 importance metrics, showing genes confirmed in at least 1 (A) or 50 (B) out of 1,000 total bootstraps.**

### Table S1

**Summary descriptive statistics table by groups of binary risk group (BRG, Low = 0, Not-low = 1).**

### Table S2

**Differentially expressed genes (DEGs) from training set comparison (binary risk group, sBRG 0 vs 1).**

## Table S3

**Standardized output table showing gene-wise importance across 4 algorithms. Ensembl gene ID (column 1), gene symbol (2), lasso beta coefficients (3:5), random forest importance (6:8), XGBoost importance (9:11), and SVM weights (12:15).**

## Table S4

**Manifest for data download.**