Mechanistic Deep Learning Framework on Cell Traits Derived from Single-Cell Mass Cytometry Data

Bo Wang[1], Sandra S. Zinkel[2,3], and Eric R. Gamazon[1, 4, 5, 6]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA

[3]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[4]Vanderbit Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[5]Data Science Institute, Vanderbilt University Medical Center, Nashville, TN

[6]Clare Hall, University of Cambridge, Cambridge, United Kingdom

Send correspondence to:

Eric R. Gamazon <eric.gamazon@vumc.org> or Sandra Zinkel <sandra.zinkel@vumc.org>

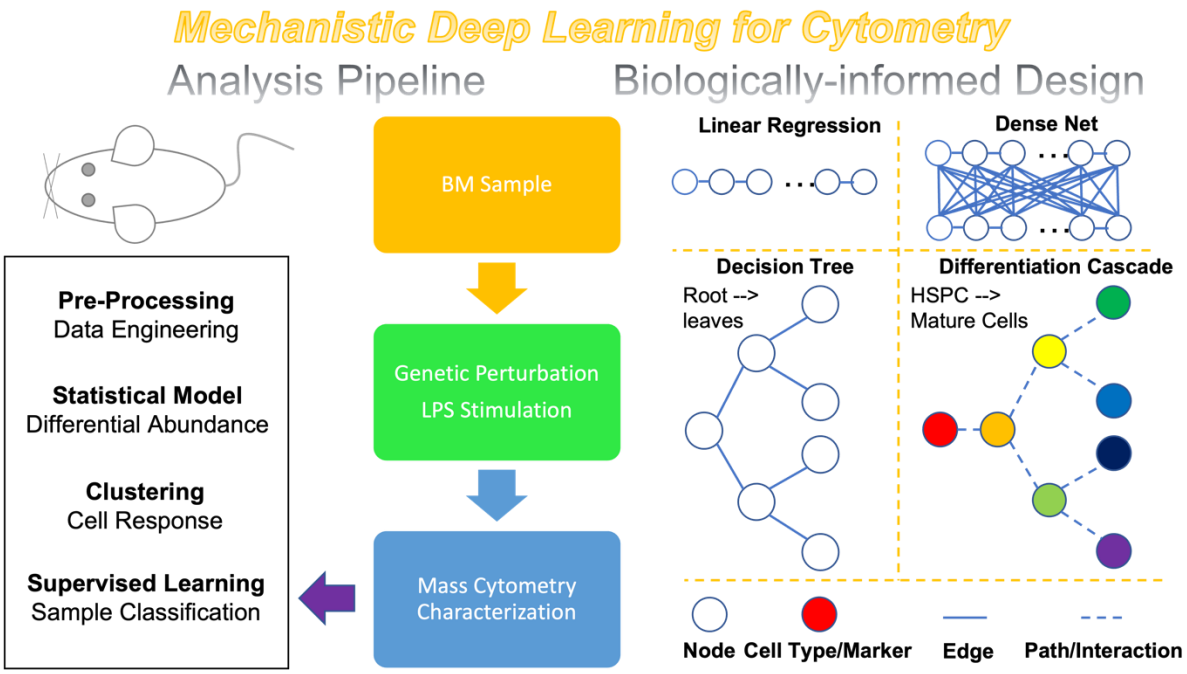
Abstract

Germline genetic variations can alter cellular differentiation, potentially impacting the response of immune cells to inflammatory challenges. Current variant- and gene- based methods in mouse and human models have established associations with disease phenotypes; however, the underlying mechanisms at the cellular level are less well-understood. Immunophenotyping by multi-parameter flow cytometry, and more recently mass cytometry, has allowed high-resolution identification and characterization of hematopoietic cells. The obtained characterization yields increased dimensionality; however, conventional analysis workflows have been inefficient, incomplete, or unreliable. In this work, we develop a comprehensive machine learning framework – MDL4Cyto – that is tailored to the analysis of mass cytometry data, incorporating statistical, unsupervised learning, and supervised learning models. The statistical modeling can be used to illuminate cell fate decision and cell-type dynamics. The unsupervised learning models along with complementary marker enrichment analyses highlight genetic perturbations that are significantly associated with alterations in cell populations in the hematopoietic system. Furthermore, our supervised learning models, including deep learning and tree-based algorithms, address the bottleneck to data pre-processing that characterizes conventional workflows and generate inferences (e.g., on marker/cell-type interactions) from raw experimental characterization. Notably, we reveal a close relationship among network design, prediction performance, and the underlying biological context. We show that the network architecture extracted from the differentiation cascade of the investigated biological system yields enhanced prediction performance. The presented methodology will enable new insights into hematopoietic differentiation at baseline and following perturbation.

Highlights

- Analysis pipeline on mass cytometry data with high-performance implementation of statistical, unsupervised learning, and supervised learning models
- Concordance of machine learning results with biological contexts
- Biologically-informed neural network designs enhance prediction performance

Graphical Abstract



Keywords

Machine Learning, Mass Cytometry, CyTOF, Hematopoiesis, Cellular Differentiation, Regulatory Cascade, Clonal Hematopoiesis of Indeterminate Potential

55 ## Introduction

56 Hematopoiesis is a hierarchical and dynamic process. Proper blood cell development requires
57 highly coordinated and tightly regulated cytokine signals, affecting cell and stage-appropriate gene
58 expression. Genetically modified mouse models are an invaluable model system for studying
59 hematopoiesis, due to their defined genetic background, wealth of reagents to precisely define
60 hematopoietic cell populations, ability to undergo precise genetic modification, and the wealth of data
61 and assays arising from decades of study. Cytometry tools have proved to be essential in simultaneously
62 analyzing labeled molecules at the single-cell level and, consequently, have been widely implemented in
63 basic and clinical research[1]. While traditional fluorescence cytometers can detect up to 35 markers, the
64 technique is limited by spectral overlap of fluorophores. Single-cell mass cytometry (Cytometry by Time-
65 Of-Flight [CyTOF]) enables multiple-parameter tracking by the use of metal-isotope-tagged antibodies,
66 overcoming the inherent limitations of spectral overlap. Therefore, a more comprehensive panel of
67 cellular functions (achieved from significantly expanded feature capacity[2]) along with intracellular
68 signaling states can be explored to identify disease-associated changes or functionally-relevant states;
69 differential expression under external perturbation or intracellular response is more easily elucidated[3].

70 Insights from cytometry characterization are traditionally extracted from manual gating.
71 Automated workflows such as FlowCore[4], PhenoGraph[5] or the recently reported Tracking Responders
72 EXpanding (T-REX)[6] have been introduced with significantly enhanced processing efficiency. However,
73 these pipelines are still limited by expert intervention into (manual) data pre-processing, where key
74 thresholds for transformation, gating or cell type clustering are tailored to the investigated systems.
75 Recently, machine learning algorithms have attracted significant attention for their extensive application
76 to the analysis of cytometry data[6–8]. Specifically, beyond functioning as embedded algorithms in
77 pipelines, machine learning tools provide “end-to-end” solutions, generating inferences or predictions

78 directly from raw experimental characterization, with potential to uncover novel associations that are
79 obscured by conventional gating strategies.

80 Hematopoiesis has served as a paradigm of stem cell biology and cellular differentiation[9].
81 Genetic perturbations within cells (in particular, Hematopoietic Stem and Progenitor Cells [HSPCs])
82 induce challenges to cellular functions, thus leading to various diseases or conditions, e.g., Clonal
83 Hematopoiesis of Indeterminate Potential (CHIP), Myelodysplastic Syndromes (MDS), and Acute Myeloid
84 Leukemia (AML). Tet2 inactivating mutations are among the early mutations that initiate clonal
85 expansion (CHIP). As the most expressed Tet gene in the hematopoietic tissue (especially in HSCs), Tet2
86 functions as a tumor suppressor whose haplo-insufficiency initiates myeloid and lymphoid
87 transformations[8]. Tet2 has been shown to regulate resolution of inflammatory responses. Tet2 KO mice
88 and cell lines display prolonged and enhanced inflammatory cytokine signaling, which impacts HSPC
89 differentiation.

90 Hematopoietic homeostasis is further controlled through tight regulation of programmed cell
91 death. The two major types of cell death, apoptosis (programmed cell death) and necrosis (uncontrolled
92 cell death), maintain homeostasis by downregulating detrimental inflammation and prevent
93 transformation by allowing efficient removal of damaged cells. It has been widely recognized that the
94 genes in the Bcl-2 family are critical in regulating apoptosis[10,11], comprising of both pro-survival
95 members and pro-apoptotic members. The functions of BID (pro-apoptotic member and multidomain
96 effector) were investigated and confirmed in our previous genomic studies[12]. Specifically, an association
97 between decreased genetically determined BID expression and myocardial infarction susceptibility was
98 identified in a discovery cohort using a gene-based association method[13], and validated by two
99 independent genome-wide association studies (GWAS). In the follow-up studies, we further tested the
100 functionalities of proteins in the family, including multidomain effectors Bax, Bak and the BH3-only
101 protein, Bid, in murine models[14]. We found that deletion of Bax/Bak inhibits both bone marrow

102 apoptosis and bone marrow necroptosis. Further deletion of BH3-only member Bid (TKO) leads to robust
103 activation of necroptosis, which results in a phenotype characterized by dysplasia and cytopenia,
104 resembling the bone marrow failure disorder MDS (myelodysplastic syndrome)[14].

105 To address the limitations of conventional cytometry data analysis pipelines[5–8,15], we developed
106 a comprehensive *Mechanistic Deep Learning for Cytometry* (MDL4Cyto) framework, implementing
107 statistical, unsupervised learning, and supervised learning models. The framework was evaluated against
108 bone marrow samples of murine models that have been genetically perturbed in the hematopoietic
109 system, including Wild Type (WT), *VavCreBaxBakBid* triple-knockout (TKO), and *VavCreTet2* knock-out
110 (Tet2 KO) mice, and validated with biological observations. Notably, building on our previous work[16,17],
111 we designed the architectures of the neural networks using the intrinsic hierarchy of the hematopoietic
112 system as a guide, establishing a *biological feature – network architecture – phenotype* relationship. We
113 show that biologically-informed network designs could function as probes on the regulatory cascade
114 with prediction performance as the evaluation metric. Our framework may inform future machine
115 learning implementations to address fundamental questions in cell biology.

116

117 Results

118 Overview of mass cytometry data and MDL4Cyto framework

119 Mice from three genetic murine models – WT, Tet2 and TKO – were treated with Phosphate-
120 buffered saline (PBS) or lipopolysaccharide (LPS). LPS is a major surface membrane polysaccharide found
121 on Gram-negative bacteria and is a typical pathogen-associated molecular pattern (PAMP).
122 Inflammation triggered by LPS stimulation activates a robust innate immune response and can induce
123 the proliferation of hematopoietic progenitor cells[18] and the production of cytokines such as TNFα,

124 interleukin and IL-2. Bone marrow from these mice was subsequently harvested and processed for
125 cytometry experiments. Besides genetic perturbations, all the samples challenged by LPS stimulation
126 were compared to the PBS vehicle control (Figure 1A). Across different types of samples (WT-PBS, WT-
127 LPS, Tet2-PBS, Tet2-LPS, TKO-PBS, and TKO-LPS), lineage marker groups (Stem Cells, Pan Leukocyte
128 Markers, T Lymphocytes, Mature Myeloid Cells, B cells, and Erythrocytes) and functional marker groups
129 (Cytokines, Cell death, and Proliferation), were tagged and monitored (Table S1). As sample data profiles
130 obtained from the mass cytometry measurements varied in length, we imputed data points at each of
131 the six sample data profiles via binning methods with 30 seconds.

132 We developed MDL4Cyto, an analysis pipeline for mass cytometry data (Figure 1B and
133 **Methods**), implementing statistical, unsupervised, and supervised learning models. Statistical tests on
134 the cell count of key lineage and functional markers were performed, aiming for a glimpse of the
135 underlying biology. As both genetic perturbations and LPS stimulation may induce inflammatory changes
136 in hematopoiesis, we applied Generalized Mixed-Effects Model (GMM) estimation to determine the
137 extent to which cell count was statistically altered by these modifications. Two established workflows,
138 namely PhenoGraph[5] and T-REX[6], were benchmarked as established pipelines capable of annotating cell
139 populations or capturing cell responses. In addition, supervised learning models were explicitly designed
140 to probe marker/cell-type interactions within the regulatory cascade, as well as to predict sample type
141 directly from the raw mass cytometry characterization. Nine machine learning models were evaluated,
142 including Linear Regression (LR), Dense Neural Network (DNN), Boosted Tree (BT), Random Forest (RF),
143 stacked Dense Neural Network - Random Forest (DNN-RF), Convolutional Neural Network in One
144 Dimension (CNN-1D), Convolutional Neural Network in Two Dimensions (CNN-2D), Recurrent Neural
145 Network with Long Short-Term Memory cell (RNN-LSTM), and Recurrent Neural Network with Gated
146 Recurrent Units cell (RNN-GRU).

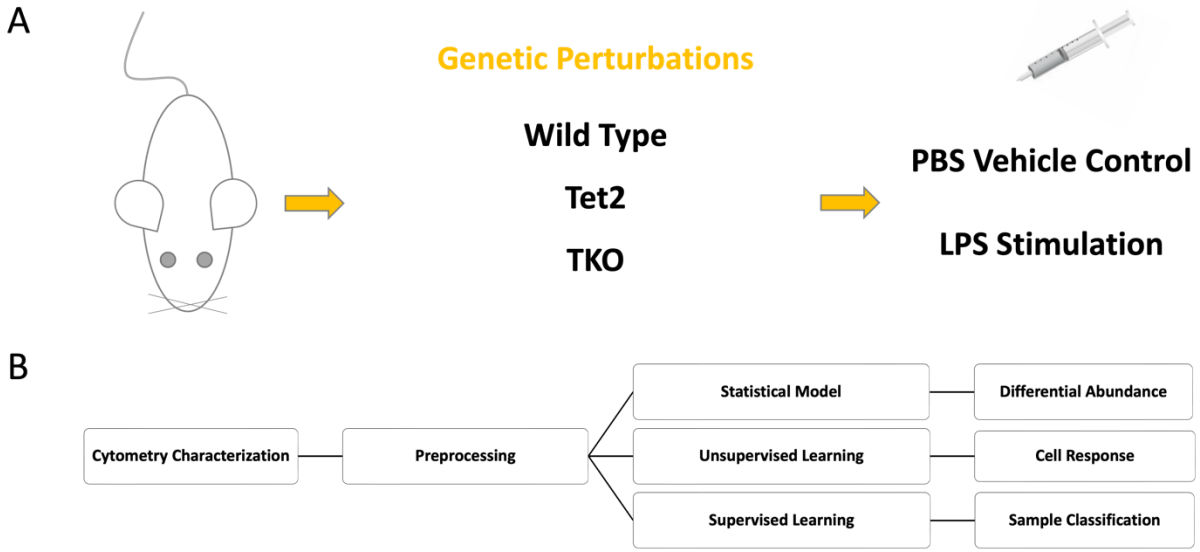


Figure 1. MDL4Cyto pipeline for analysis of single-cell mass cytometry data. Bone marrow samples (**A**) were harvested from genetically perturbed murine models (Wild Type [WT], *VavCreBaxBakBid* triple-knockout [TKO], and *VavCreTet2* knock-out [Tet2 KO]), which had been treated with LPS stimulation or PBS vehicle control. Statistical, unsupervised, and supervised learning models were implemented in our *Mechanistic Deep Learning for Cytometry* (MDL4Cyto) framework. The analysis pipeline **(B)** was benchmarked, using the reference dataset obtained from mass cytometry characterization, for the tasks of differential abundance analysis, cell response detection, and sample type classification.

Statistical modeling of genetically-determined or condition-dependent effects

The cell count summary statistic, which was calculated on the reference dataset, was found to be altered by genetic perturbations and/or LPS stimulation (Figure S1). Clear differences were observed for all markers. As reported in our previous work[14] and other studies[19], the TKO sample leads to necrosis with increasing production of cytokines. Enhanced inflammatory signaling was also observed in Tet2-deficient samples, as has also been previously described[19]. We observed a statistically significant difference in cell count for Sca1 (LSK cells), CD45 (leukocyte), TNFα (cytokines), and Caspase 3 (cell

164 death indicator) with respect to genetic perturbation (Figure 2A). Markers of cytokines were found to be

165 highly correlated with other surface markers, especially under LPS stimulation (Figure 2B).

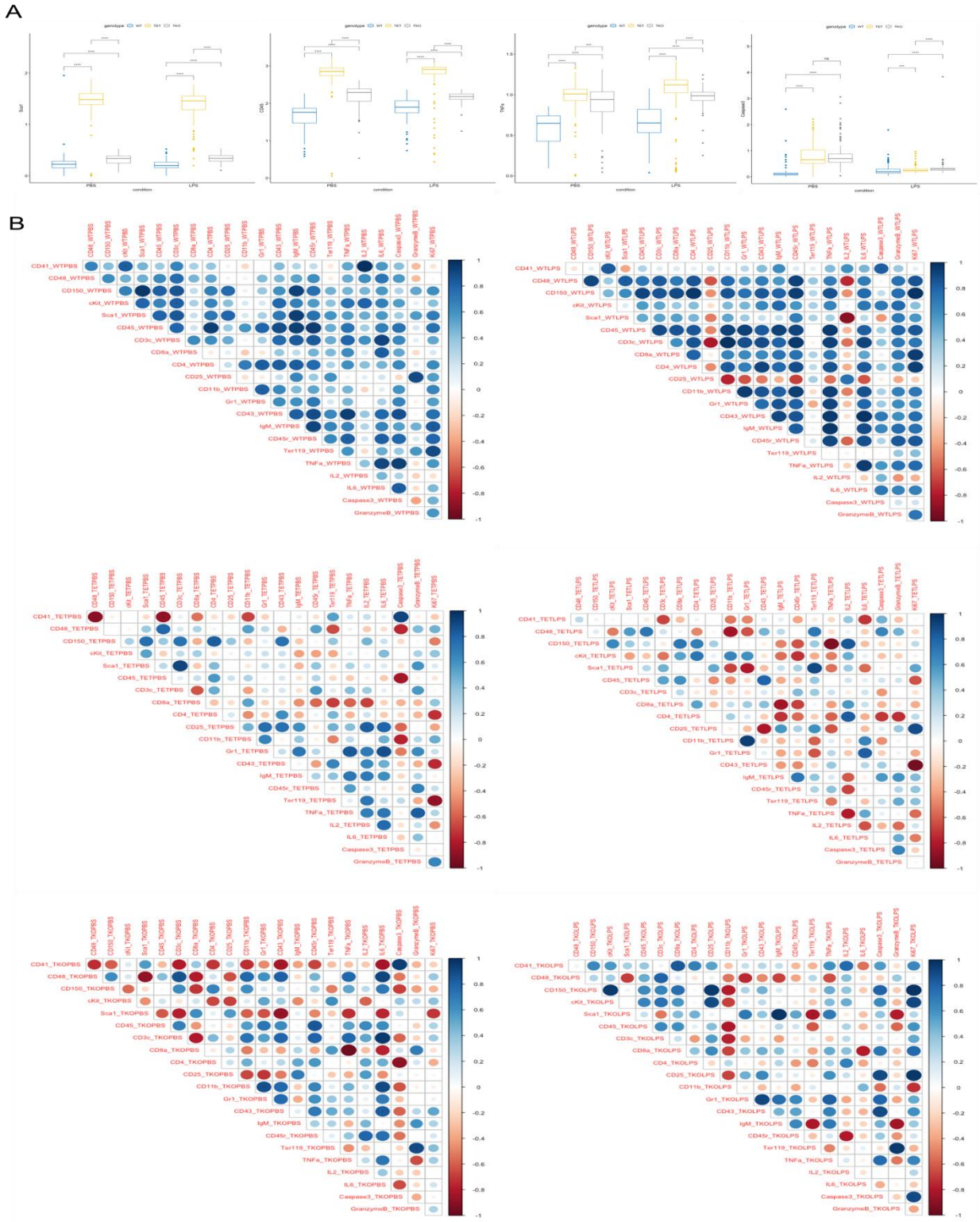
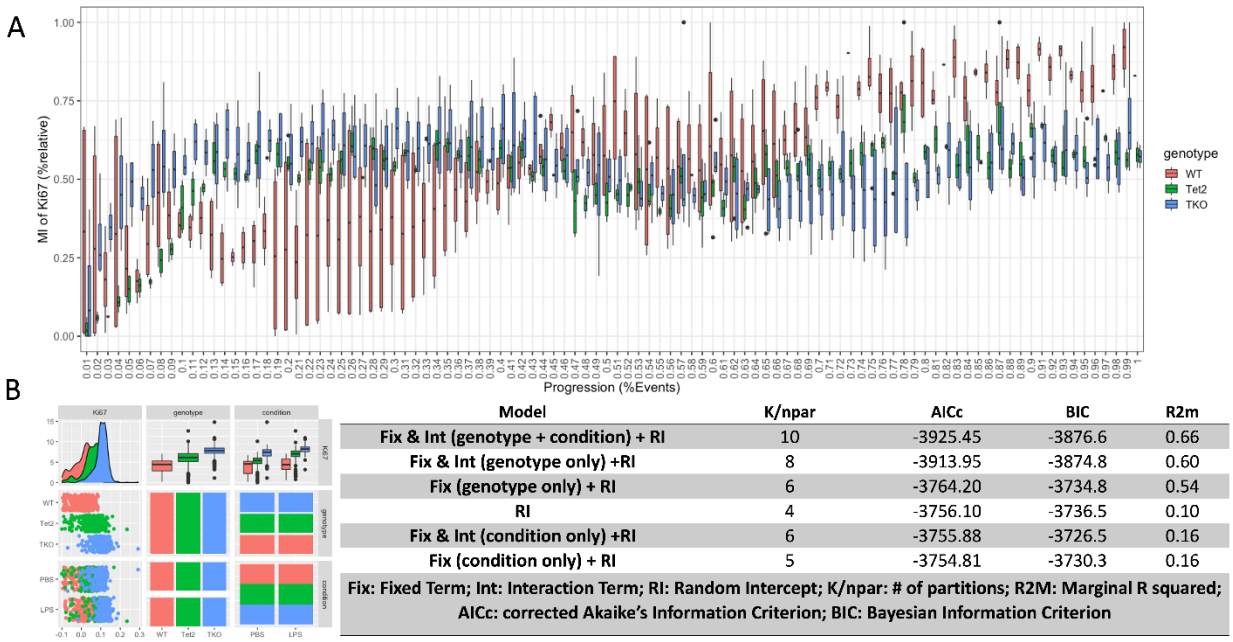


Figure 2. Impact of genetic perturbations on cell populations and correlations among the markers. A statistically significant change in cell count was observed for Sca1 (LSK cells), CD45 (leukocyte), cytokines

169 (TNFα), and Caspase 3 and (cell death indicator) in both the PBS vehicle control and LPS stimulation
170 condition, demonstrating genetic perturbations altering cell subpopulations **(A)**. Cytokines from genetic
171 knock-out samples were highly correlated with other surface markers, in particular under LPS
172 stimulation **(B)**, which is consistent with the biological context. (Wilcoxon test was performed for the
173 pairwise comparisons; Pearson correlation was calculated for the markers along with 95% confidence
174 interval. *p<.05, **p<0.01, ***p<0.001, ****p<0.0001).

175

176 The data profiles obtained from cytometry characterizations are dynamical panels, i.e., have a
177 time component. To estimate the overall effects of genotype and condition, we analyzed the
178 progression of marker Intensity for Ki67 from the aggregated data panel with 1 hour binning window
179 (Figure 3 and **Methods**). Ki67 is a proliferation protein that is increased in the S phase of the cell cycle
180 and indicates proliferating cells. We observed progression divergence according to genotype (Figure 3A)
181 and differentiated marker intensity between conditions (PBS vs LPS, Figure 3B). This is consistent with
182 our understanding that genetic perturbations and LPS stimulation impose pressures on differentiation
183 and cell cycle as dynamical processes. To achieve a better quantitative estimate of the contributions
184 arising from genotype and condition, we implemented GMM estimation. Besides the fixed and/or
185 interaction terms (**Methods**), the time term was considered in the model (Figure 3C). We applied
186 corrected Akaike's Information Criterion (AICc) and Bayesian Information Criterion (BIC) for model
187 evaluation and Marginal R squared (R2m) for variance coverage. Models built from genotype
188 demonstrated the best performance, which corroborates the fact that genetic perturbations have
189 significant impact on cellular differentiation and function. We note that the statistical modeling
190 implemented in MDL4Cyto (Figure 1B and **Methods**) is broadly applicable, which can facilitate the study
191 of cell fate decision and cell-type dynamics.



Model	K/npar	AICc	BIC	R2m
Fix & Int (genotype + condition) + RI	10	-3925.45	-3876.6	0.66
Fix & Int (genotype only) +RI	8	-3913.95	-3874.8	0.60
Fix (genotype only) + RI	6	-3764.20	-3734.8	0.54
RI	4	-3756.10	-3736.5	0.10
Fix & Int (condition only) +RI	6	-3755.88	-3726.5	0.16
Fix (condition only) + RI	5	-3754.81	-3730.3	0.16

Fix: Fixed Term; Int: Interaction Term; RI: Random Intercept; K/npar: # of partitions; R2M: Marginal R squared; AICc: corrected Akaike's Information Criterion; BIC: Bayesian Information Criterion

Figure 3. Analysis of dynamic data panel. Progression (%Events over the course of time) **(A)** and histogram **(B)** of cell count of proliferation marker Ki67 diverged with genetic background and shifted with condition. We implemented Generalized Mixed-Effects Model (GMM) estimation that considered the time term in addition to a fixed term and interaction term. While using genotype and condition as main parameters, several models were constructed, as represented by the equations in **(C)**. The performance of each model was further evaluated using the corrected Akaike's Information Criterion (AICc), Bayesian Information Criterion (BIC), and Marginal R squared (R2m). Genotype was identified as an important feature.

Quantifying cell response and population alteration via unsupervised learning

Identifying cell response from cell populations has been a key interest in mass cytometry analysis. Hematopoietic responses, induced either by genetic perturbations or by pathogen-like stimulation (LPS), were analyzed via pairwise comparisons, including WT-PBS vs TKO-PBS, WT-LPS vs TKO-LPS, WT-PBS vs Tet2-PBS, WT-LPS vs Tet2-LPS. Two established unsupervised learning algorithms, T-REX and PhenoGraph, were benchmarked in clustering our reference dataset (see Table S3 for the hyperparameters). Both algorithms apply k-nearest neighbors (kNN) for clustering and Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and visualization. As illustrated in the UMAP plot, genetic perturbations induced alterations in cell subpopulations (Figure 4A&4B). Notably,

211 differential cell count of markers under a genetic perturbation, including Sca1, CD45, TNFα, and Caspase
212 3, could be seen in the mapping of cell count to the clusters (Figure S2, T-REX algorithm). These results
213 are consistent with the comparisons from our statistical tests (Figure 2A). Compared to conventional
214 methods, unsupervised learning showed a unique advantage in efficiently capturing changes in cell
215 populations. In our framework, cell subpopulations within the clusters were further investigated in two
216 ways. While reference datasets trained here contained only a single sample for each type, we created
217 two pseudo-replicates from a single sample (**Methods**) for the PhenoGraph algorithm, thus enabling
218 differential abundance analysis for the markers within each cluster (Figure 4C). In addition, populations
219 from the T-REX algorithm were screened and enriched through Density-Based Spatial Clustering and
220 Application with Noise (DBSCAN) and Marker Enrichment Modeling (MEM)[20] (Figure 4D). Strikingly, in
221 contrast to the T-REX method, our framework uncovered a strong signal from the cytokines, i.e., IL6 and
222 TNFα, from the differential abundance analysis (Figure 4C). This observation could well be attributed to
223 the known LPS-driven inflammation, as demonstrated by the enrichment analysis from MEM (Figure
224 4D). In addition, B cells and T cells are derived from the myeloid lineage, providing a broad and general
225 defense against foreign pathogens. Collectively, the TNFα and IL6 differential abundance results are
226 consistent with myeloid-driven or innate inflammation.

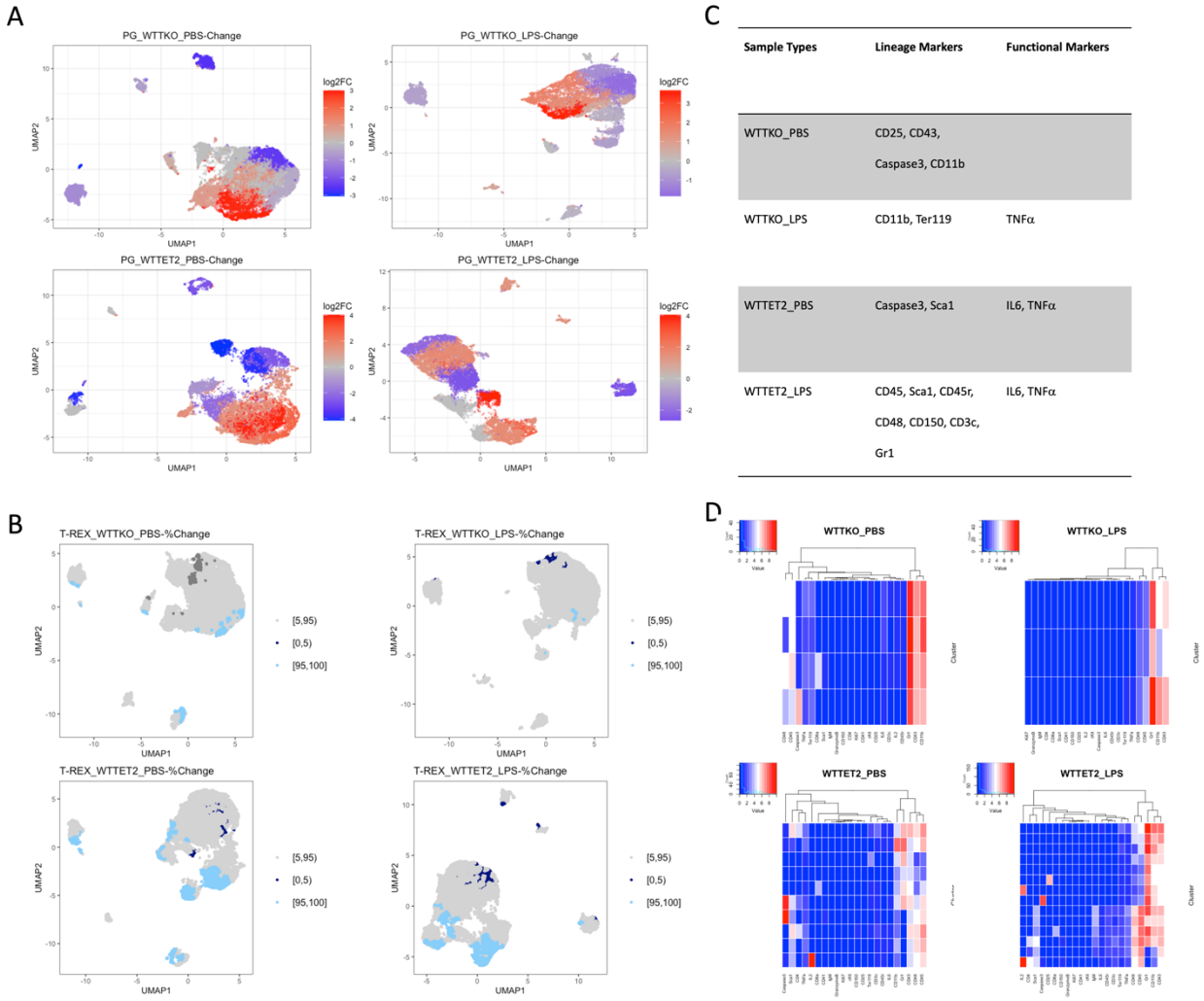


Figure 4. Cell response identified by the unsupervised learning. Clustering from Phenograph **(A)** and T-REX **(B)** algorithms were performed on the paired genetically perturbed samples. PBS vehicle control and LPS stimulation were separately explored and compared. Using both algorithms, we found that the genetic perturbations altered the cell subpopulations. Clusters of significant fold changes identified from PhenoGraph were further explored for differential abundance of markers. Markers with significant responses that were captured in more than two clusters were identified **(C).** On the other hand, cell subpopulations obtained from T-REX were screened for those with significant changes and enriched through DBSCAN and MEM **(D)**.

Probing the regulatory cascade with biologically-informed neural networks

The hematopoietic system has classically been represented as a rigid hierarchical tree, illustrating HSC progression through multipotent, oligopotent, and bipotent stages. Evaluating a

240 “dynamic” model with complex interactions, we interrogated the marker/cell-type interactions within
241 the regulatory cascade from supervised learning models, considering both the network architecture and
242 prediction performance as probes. LPS stimulation impacts all stages of HSC differentiation, thus having
243 a global impact on all surface markers[18]. Therefore, we assigned PBS vehicle control and LPS stimulation
244 as predicted labels to the samples (**Methods**). Twenty-two lineage/functional makers from nine
245 different cell types (Table S1) were encoded as essential numerical features, while genotype was loaded
246 as an external categorical/global feature.

247 Considering each node as a marker/cell type, their interactions could be represented as network
248 connections. Specifically, a linear regression model reflects the additive effect from markers/cell types; a
249 dense neural network models the complex interactions among markers/cell types; and a tree model
250 considers the binary relations between levels within the hierarchy (Figure 5A). LPS stimulation on the
251 cells (the prediction label) could be treated as a perturbation on the neural network. In this model,
252 HSPCs will be differentiated to the common lymphoid progenitors, unipotent precursor cells as well as
253 megakaryocyte-erythroid progenitor cells, eventually producing mature cells. Informed by the
254 differentiation hierarchy, we implemented Boosted Tree (BT) and Random Forest (RF) models along with
255 two reference models, Linear Regression (LR) and Deep Neural Network (DNN). The four models varied
256 in performance: Linear Regression (LR) achieved a prediction accuracy of 64.8%; the DNN model with
257 three-hidden layers enhanced the accuracy to 68.1%; the Boosted Tree (BT) model further increased the
258 prediction performance to 72.7%; and the Random Forest (RF) models topped the accuracy at 77.13%
259 (Figure 5B). This overall pattern of prediction accuracy (LR < DNN < BT < RF) suggests a non-linear,
260 hierarchical relationship among the features, consistent with the classical hematopoietic hierarchical
261 tree. Interestingly, when we stacked a Deep Neural Network on top of the Random Forest model, where
262 DNN functioned as a denoising head[21], the new stacked model showed degraded performance in
263 comparison to the RF model alone.

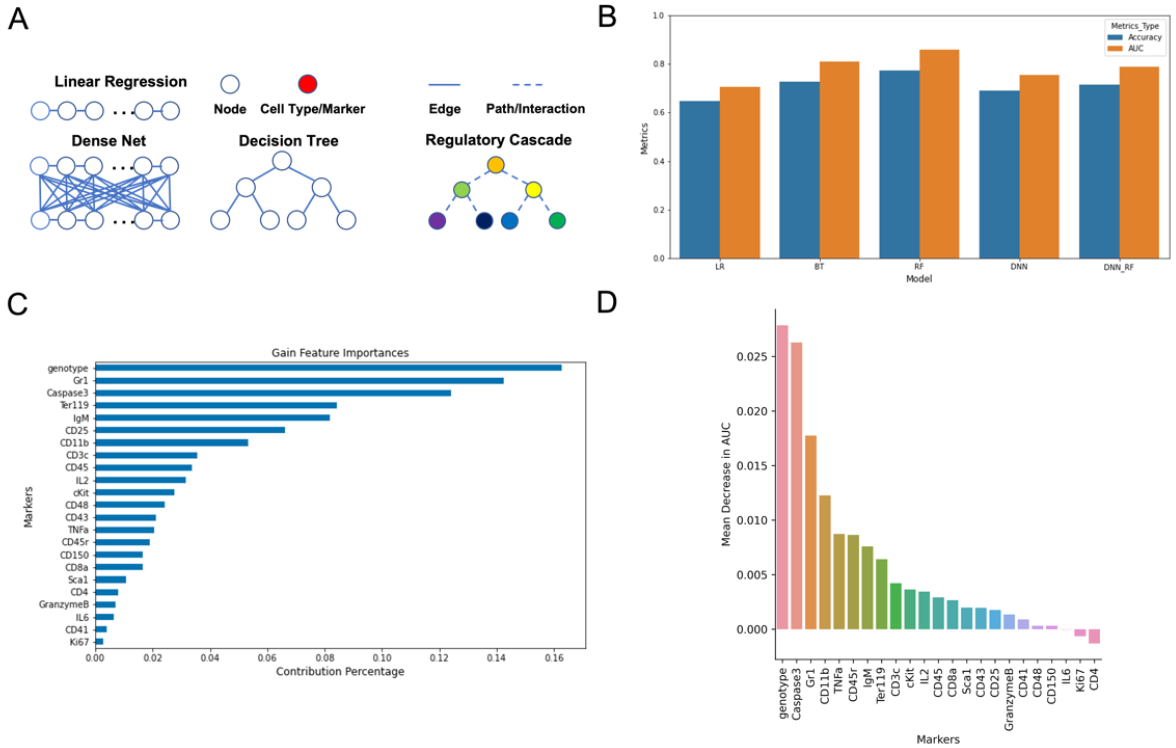


Figure 5. Mechanistic insights extracted from biologically-informed neural networks. Our neural network designs were guided by the biological context, in which we considered each node as a marker. The marker/cell-type interactions could therefore be represented as a network **(A)**, e.g., a decision tree or neural network could be linked to the regulatory cascade. With PBS vehicle control and LPS stimulation as labels, tree-based models (Boosted Tree [BT] and Random Forest [RF]) showed improved performance compared to Linear Regression (LR) and Deep Neural Network (DNN) models as well as DNN-RF stacked models **(B),** consistent with the hematopoietic tree hierarchy (i.e., representing a sequential binary branching decision process) among the markers. Feature importance of individual markers in the BT model, which implements gain-based feature importance **(C),** and RF model, which implements mean decrease in AUC **(D),** were examined. The group of markers with significant contributions would imply an overall cell response under inflammatory (LPS), e.g., pathogen-induced, challenges.

We further considered the feature contributions of individual markers using gain-based feature importance (Figure 5C) and mean decrease in AUC (Figure 5D). Gain-based feature importance reflects the relative contribution of a given feature to the model. Mean decrease in AUC indicates the decline in performance once the corresponding feature has been removed from the training dataset. We hypothesized that markers/features with the top feature contributions are biologically meaningful.

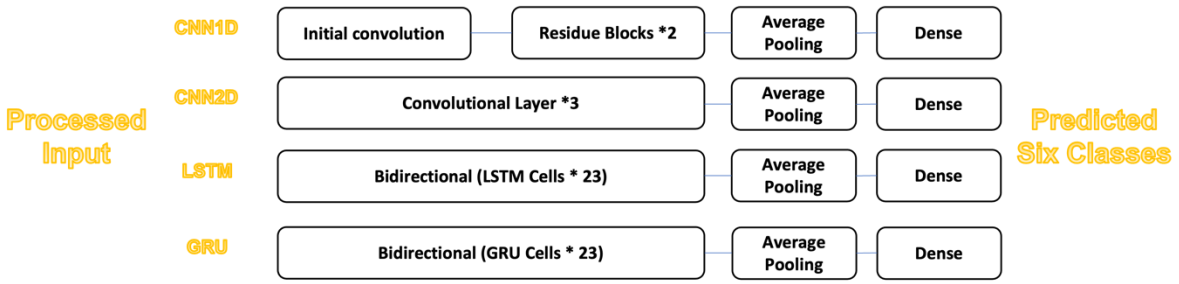
283 Indeed, the single most important contribution came from the strongest perturbation – genotype.
284 Individual markers, including Caspase3, Gr1, IgM, CD11b, and CD25, that contribute substantially to the
285 prediction of LPS stimulation, are highly biologically relevant in the full differentiation cycle of
286 hematopoiesis. Caspase 3 is an important player in the regulated cell death. Gr1 and CD11b reflect a
287 response from myeloid cells. And Ter119 indicates a strong immune response from the erythrocytes
288 under inflammatory challenges.

289

290 **Applying machine learning in sample classification**

291 While conventional methods/pipelines are limited by domain knowledge or manual
292 intervention, a predictive model that bypasses cell clustering or manual gating may provide an end-to-
293 end approach to the analysis of mass cytometry data. Here, we benchmarked a group of supervised
294 learning models, including CNN-1D, CNN-2D, RNN-LSTM, and RNN-GRU, that directly handle the raw
295 cytometry profiles for sample classification, i.e., prediction of clinical or biological outcome (Figure 6A).
296 We established multi-class output labels corresponding to six different sample types, including WT-PBS,
297 WT-LPS, Tet2-PBS, Tet2-LPS, TKO-PBS, and TKO-LPS. We used the curated ensemble of (1 hour)
298 "segments" from the initial bulk samples in evaluating the model performance. We found that CNN and
299 RNN-derived models demonstrated good prediction performance with at least 95% accuracy (Figure 6B).
300 We argue that convolutional layers and recurrent cells are efficient in learning the relevant features.
301 Algorithms such as CNN and LSTM as applied here, with cell count as the trait, captured with greater
302 fidelity the landscape of marker/cell-type interactions than the tree-based discrete models evaluated
303 above.

A



B

Model	Train Acc	Val Acc	Test Acc
CNN-1D	0.99783	0.97400	0.95887
CNN-2D	0.97580	0.95233	0.96537
RNN-LSTM	0.99819	0.96425	0.97186
RNN-GRU	1.000	0.98375	0.98052

Figure 6. Supervised learning identified sample type with high prediction performance. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) derived models (CNN-1D, CNN-2D, RNN-LSTM, RNN-GRU) were benchmarked for the task of sample type classification. The specifics of the architectures are shown in **(A)**. All tested models achieved high performance with over 95% in accuracy **(B),** demonstrating the importance of learning the complex interactions and high-order features.

Discussion

The current single-cell genomic revolution has enabled entire systems, such as the hematopoietic system, to be interrogated comprehensively at an unprecedented resolution, revealing a previously unappreciated degree of functional and phenotypic variation within cell compartments. In this work, bone marrow samples were harvested from mice and further characterized in mass cytometry – the baseline murine model (WT) was compared to genetic perturbations, including *Tet2*-knockout (*VavCreTet2*) and *VavCreBaxBakBid* triple-knockout (TKO). LPS served as an external stimulus or an innate immune response, simulating pathogen infection. Statistical modeling and unsupervised learning revealed the direct effect of genetic perturbations. Markers of interest that are typically processed individually could be combined in unsupervised learning models. Importantly, we revealed a close connection among the intrinsic hierarchy of HSPC differentiation, neural network architecture, and prediction performance. Specifically, Boosted Tree (BT) and Random Forest (RF), which closely resemble

323 the stepwise binary branching decisions of the classical hierarchical model of hematopoiesis, showed
324 superior performance in comparison with Linear Regression (LR), Deep Neural Network (DNN) or even
325 the DNN-RF stacked model. The prediction performance was substantially enhanced relative to neural
326 network models, as the model reflected the cell differentiation cascade. However, as recent studies with
327 functional assays, non-invasive lineage tracing, and omics studies at single-cell resolution have shown[22],
328 hematopoiesis is a continuous process in which obvious boundaries are lacking between cell populations
329 located at different levels of the hierarchy. Consequently, given their architecture, tree-based models
330 may have an inherent limitation in capturing transition states, even with extensive optimization,
331 constraining their ability to further improve prediction performance. From this perspective, the complex
332 relationships among the cell populations (including feed-back loops in pathways) may be better
333 elucidated by neural networks.

334 Nevertheless, results of the analyses from the machine learning models were consistent with
335 biological insights such as myeloid-driven or innate inflammation (responses from TNFα and IL6). In
336 summary, the presented framework provides a unifying methodological framework for the analysis of
337 cytometry data, combining data-driven and model-driven approaches. We constructed interpretable
338 models that are anchored in biological context, with implications for ongoing and future studies in
339 single-cell biology.

340

341 Limitations of the study

342 Machine learning models, including graphical representation learning and tree-based
343 algorithms, were trained in datasets obtained from mass cytometry (CyTOF) characterization. While we
344 designed the neural network architectures with biological information as guide, assumptions of cell
345 types as nodes and edges as their interactions within a graph-based framework are based on the classic

346 discrete cell differentiation model. Cell differentiation is a continuous process, with intermediate states
347 among various levels leading to much more complex networks.

348

349 # Methods

350 **Mass cytometry and murine samples**

351 Three generic murine models were included in this study, namely, wild type (WT), Tet2 knockout
352 (Tet2)[23], and *VavCreBaxBakBid* triple-knockout (TKO)[14]. Prior to bone marrow isolation, the mice had
353 been treated intraperitoneally with 1.5 mg/kg LPS (L4391; Millipore-Sigma) or PBS for 8 hours. The
354 obtained samples for the six genotype-condition types were subsequently characterized using mass
355 cytometry, according to previously established protocols[24]. Mass cytometry combines inductively
356 coupled plasma time-of-flight mass spectrometry and cytometry. Specifically, mass cytometry staining
357 was performed on whole bone marrow with antibodies conjugated to heavy metals. Cell responses from
358 9 different cell type and functional categories (stem cells, pan leukocyte, T lymphocytes cells, mature
359 myeloid cells, B cells, erythrocytes, cytokines, cell death, and proliferation) were recapitulated by 16
360 surface/lineage markers and 6 function/intracellular markers. Markers from pre-gating channels (DNA-1
361 and DNA-2) and instrument channels (Time/Event) were also considered (Table S1). The CyTOF data
362 generated from the six samples consisted of a maximum of 293,778 cells and a minimum of 121,124
363 cells.

364 **Data pre-processing**

365 Initial cell count data profiles obtained from the mass cytometry characterization were filtered
366 out for debris, doublets, and dead cells, thus reducing noise. They are represented as dynamical panels
367 on which the *arcsinh* transformation with co-factor = 5 and a binning window of 30 seconds was applied.
368 We sought to analyze the data in a way that maximizes robustness. The processed profiles were sliced

369 into ensembles of "segments" to achieve an optimized tradeoff between computational efficiency and
370 prediction accuracy. We evaluated an array of sizes ([2, 10, 30, 60, 120, 240, 480]), corresponding to
371 2*30S=1M, 10 * 30S = 5M, 30 * 30S = 15M, 120 * 30S = 1H, 240 * 30S = 2H, & 480 * 30S = 4H sampling
372 intervals (M: Minute, H: Hour), respectively. The obtained pile of "segments" was completely shuffled
373 and re-assembled into new ensembles. We implemented an RNN-LSTM model for this special
374 "hyperparameter search" (see prediction accuracy in Table S2) to determine the optimal segment size,
375 which was 1H. Signals of select markers, including Sca1, CD45, TNFα, Caspase 3, and Ki67, were
376 aggregated and averaged into 1 hour binning window for the statistical test and GMM estimation.

377 **Mechanistic Deep Learning Framework**

378 *Statistical test and model*

379 Wilcoxon test was performed on pairwise comparisons of marker intensity. Pearson correlation
380 coefficient was calculated for inter-marker association. We implemented a GMM on the marker Ki67.
381 We considered fixed effects and interactions among time, genotype, and condition (PBS control or LPS
382 stimulation). After testing the random slope of a sample, $(1 + \text{Time} \mid \text{Sample})$, we considered only the
383 random intercept from a sample, i.e., $(1 \mid \text{Sample})$. The models were evaluated using corrected Akaike's
384 Information Criterion (AICc), Bayesian Information Criterion (BIC), and Marginal R squared (R2M). The
385 optimal model was found to be given by the following expression:

386 $$\text{Marker Intensity} = \text{Time} + \text{Genotype} + \text{Condition} + \text{Genotype:Time} + \text{Condition:Time} + (1 \mid \text{Sample}) + \varepsilon$$

387 *Unsupervised Learning models*

388 We analyzed two established unsupervised learning pipelines, namely Phenograph[5] and T-REX[6],
389 which were applied to the preprocessed CyTOF dataset with binning window 30S. For both pipelines,
390 Uniform Manifold Approximation and Projection (UMAP)[25] was used for dimensionality reduction; k-
391 nearest neighbors (kNN)[26] for clustering; Marker Enrichment Modeling (MEM) for generating the

392 enriched heatmap. The applied hyper-parameters are presented in Table S3. For comparison, we
393 screened and enriched the cell populations derived from the T-REX algorithm through Density-Based
394 Spatial Clustering and Application with Noise (DBSCAN) and MEM.

395 *Supervised Learning models*

396 All the machine learning models were trained using the *TensorFlow (TF)* ecosystem[27], which we
397 applied to the preprocessed CyTOF dataset (with binning window size of 30S) to perform model training
398 and evaluation. Linear Regression (LR), Deep Neural Network (DNN), Boosted Tree (BT), and Random
399 Forest (RF) models were implemented on the dataset with train and test split of 8:2. The DNN was
400 implemented as a multilayer perceptron. A stacked model of Dense Neural Network and Random Forest
401 (DNN-RF) was constructed following the same splitting. Models implementing a Convolutional Neural
402 Network (CNN) or a Recurrent Neural Network (RNN) in the framework include CNN-1D, CNN-2D, RNN-
403 LSTM, and RNN-GRU. For these four architectures, pre-processed datasets were divided into train, valid,
404 and test data according to 6:2:2 ratio.

405 We trained the neural networks using mean squared error (as the loss function), *Adam* (an
406 efficient stochastic optimization method), and Leaky ReLU (as the activation function), building on our
407 previous work[16,17]. To reduce the possibility of overfitting, we implemented *dropout* assuming a dropout
408 rate of 20%.

409

410 **Acknowledgments**

411 This research is supported by National Institutes of Health (NIH) grants NHGRI R35HG010718,
412 NHGRI R01HG011138, NIA AG068026, and NHLBI R01HL133559.

413

414 **Author contributions**

415 B.W., S.S.Z., and E.R.G. conceived the project and designed the study. B.W. performed the
416 analyses and drafted the article. B.W., S.S.Z., and E.R.G. revised the article. S.S.Z. and E.R.G. supervised
417 the project and acquired funding for the study.

418

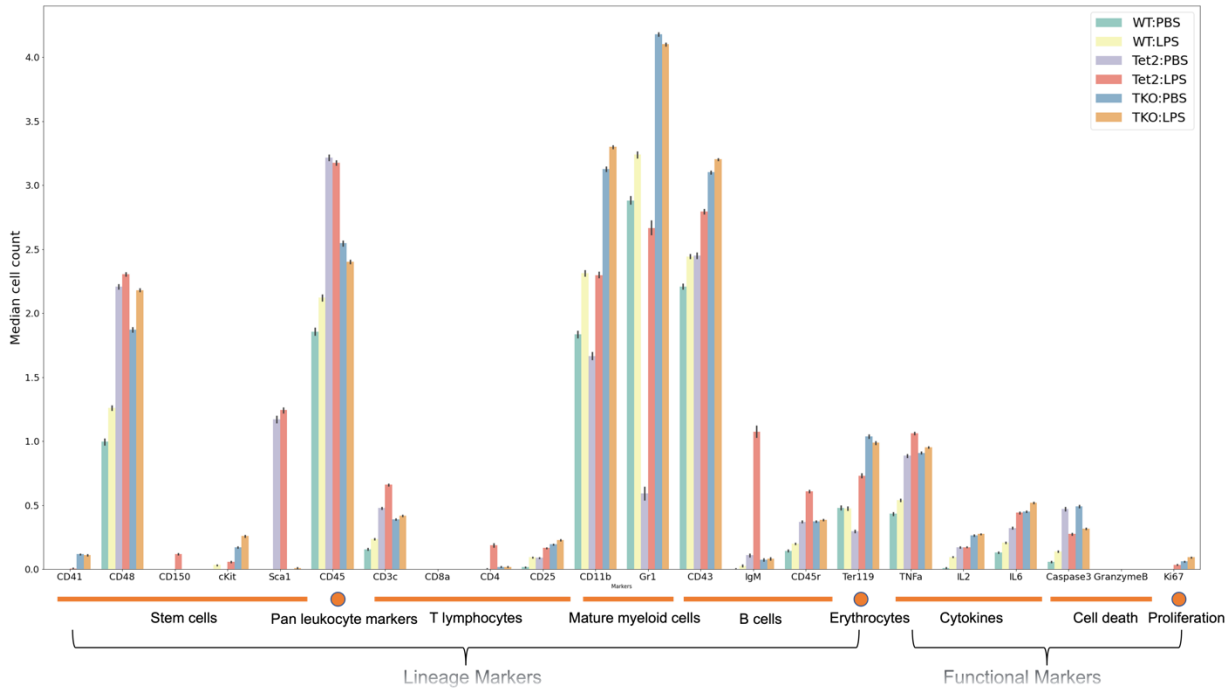
419 **References**

420 1. Davis, M. M., Tato, C. M. & Furman, D. Systems immunology: just getting started. *Nat. Immunol.* **18**,
421 725–732 (2017).
422 2. Gadalla, R. *et al.* Validation of CyTOF Against Flow Cytometry for Immunological Studies and
423 Monitoring of Human Cancer Clinical Trials. *Front. Oncol.* **9**, (2019).
424 3. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a
425 Human Hematopoietic Continuum. *Science* **332**, 687–696 (2011).
426 4. Hahne, F. *et al.* flowCore: a Bioconductor package for high throughput flow cytometry. *BMC*
427 *Bioinformatics* **10**, 106 (2009).
428 5. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that
429 Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
430 6. Barone, S. M. *et al.* Unsupervised machine learning reveals key immune cell subsets in COVID-19,
431 rhinovirus infection, and cancer therapy. *eLife* **10**, e64653 (2021).
432 7. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep
433 learning model for cytometry data. *Proc. Natl. Acad. Sci.* **117**, 21373–21380 (2020).
434 8. Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via
435 representation learning. *Nat. Commun.* **8**, 14825 (2017).

436 9. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* **182**,
437 1384–1400 (2020).
438 10. Domen, J., Cheshier, S. H. & Weissman, I. L. The role of apoptosis in the regulation of
439 hematopoietic stem cells: Overexpression of Bcl-2 increases both their number and repopulation
440 potential. *J. Exp. Med.* **191**, 253–264 (2000).
441 11. Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins: changing partners in the dance
442 towards death. *Cell Death Differ.* **25**, 65–80 (2018).
443 12. Salisbury-Ruf, C. T. *et al.* Bid maintains mitochondrial cristae structure and function and protects
444 against cardiac disease in an integrative genomics study. *eLife* **7**, e40907 (2018).
445 13. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference
446 transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
447 14. Wagner, P. N. *et al.* Increased Ripk1-mediated bone marrow necroptosis leads to myelodysplasia
448 and bone marrow failure in mice. *Blood* **133**, 107–120 (2019).
449 15. Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry
450 data. *Nat. Methods* **14**, 707–709 (2017).
451 16. Wang, B. & Gamazon, E. R. *Towards mechanistic models of mutational effects: Deep Learning on*
452 *Alzheimer's Aβ peptide*. 2021.12.19.473403
453 https://www.biorxiv.org/content/10.1101/2021.12.19.473403v2 (2021)
454 doi:10.1101/2021.12.19.473403.
455 17. Wang, B. & Gamazon, E. R. Modeling mutational effects on biochemical phenotypes using
456 convolutional neural networks: application to SARS-CoV-2. *iScience* **7**, 104500 (2022).
457 18. King, K. Y. & Goodell, M. A. Inflammatory modulation of HSCs: viewing the HSC as a foundation
458 for the immune response. *Nat. Rev. Immunol.* **11**, 685–692 (2011).

459 19. Cai, Z. *et al.* Inhibition of Inflammatory Signaling in Tet2 Mutant Preleukemic Cells Mitigates
460 Stress-Induced Abnormalities and Clonal Hematopoiesis. *Cell Stem Cell* **23**, 833-849.e5 (2018).
461 20. Diggins, K. E., Greenplate, A. R., Leelatian, N., Wogsland, C. E. & Irish, J. M. Characterizing cell
462 subsets using marker enrichment modeling. *Nat. Methods* **14**, 275–278 (2017).
463 21. Xie, J., Xu, L. & Chen, E. Image Denoising and Inpainting with Deep Neural Networks. in *Advances*
464 *in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
465 22. Zhang, Y., Gao, S., Xia, J. & Liu, F. Hematopoietic Hierarchy – An Updated Roadmap. *Trends Cell*
466 *Biol.* **28**, 976–986 (2018).
467 23. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and
468 myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
469 24. Zinkel, S. S. *et al.* Proapoptotic BID is required for myeloid homeostasis and tumor suppression.
470 *Genes Dev.* **17**, 229–239 (2003).
471 25. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for
472 Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).
473 26. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M. k-Nearest Neighbor Classification. in *Data*
474 *Mining in Agriculture* (eds. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M.) 83–106 (Springer, 2009).
475 doi:10.1007/978-0-387-88615-2_4.
476 27. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed
477 Systems. 19.
478
479
480
481

482 **Supplementary Information**



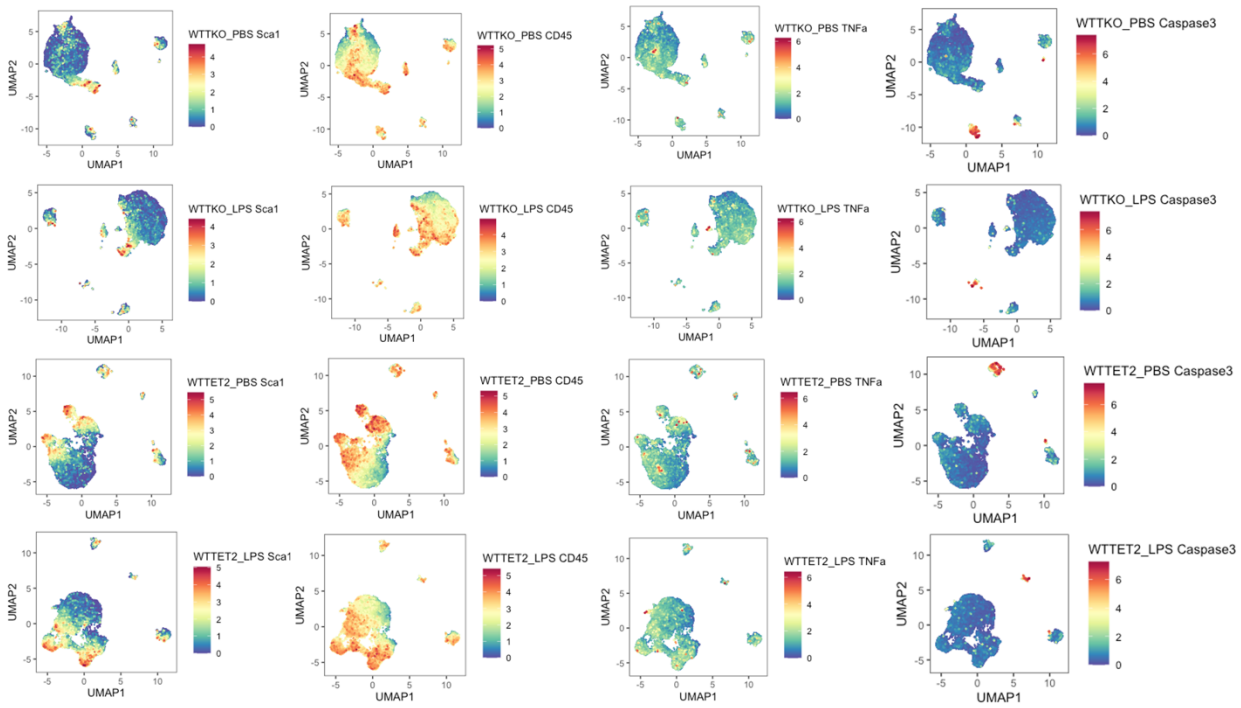
483

484 **Figure S1. Summary statistics on the cell count.** Initial sample data profiles (WT-PBS, WT-LPS, Tet2-PBS,

485 Tet2-LPS, TKO-PBS, and TKO-LPS) were obtained from mass cytometry characterization, where summary

486 statistics on the transferred cell count were computed. Lineage and functional markers are grouped

487 based on cell type.



488

489 **Figure S2. Differential cell count illustrated from the unsupervised learning.** Markers of interest under

490 genetic perturbation, including Sca1 (LSK cells), CD45 (Leukocyte), TNFα (Cytokines) and Caspase 3 (Cell

491 Death Indicator), were mapped to the clustering results from the T-REX algorithm.

492

493 **Table S1.** Cell markers investigated in this study

Marker Type	Cell Type	Marker Names
Lineage	Stem Cells	CD41, CD48, CD150, cKit, Sca1
Lineage	Pan Leukocyte	CD45
Lineage	T Lymphocytes	CD3c, CD8a, CD4, CD25
Lineage	Mature Myeloid	CD11b, Gr1
Lineage	B Cells	CD43, IgM, CD45R
Lineage	Erythrocytes	Ter119

Functional	Cytokines	TNF alpha, IL2, IL6
Functional	Cell Death	Caspase3, Granzyme B
Functional	Proliferation	Ki67
Pre-gating		DNA1, DNA2
Instrument		Cell Length, Time

494

495 **Table S2**. Searching for the optimal size "hyperparameter" via RNN-LSTM model architecture

Micro Sample Size	Train Acc	Val Acc	Test Acc
2	0.73653	0.71342	0.71033
10	0.93403	0.88667	0.87375
30	0.99791	0.95369	0.95125
60	**0.99499**	**0.9325**	**0.9575**
120	1.0	0.9598	0.975
240	1.0	0.93	0.93
480	1.0	0.77551	0.88

496

497 **Table S3.** Hyperparameters applied in PhenoGraph and T-REX

Key components in pipeline	Phenograph	T-REX
UMAP	ret_nn = TRUE	
kNN	k = 30 for single sample, k = 60 for paired sample	

498