RESEARCH BRIEF

Deep Morphology Learning Enhances *Ex Vivo* Drug Profiling-Based Precision Medicine

Tim Heinemann[1], Christoph Kornauth[2], Yannik Severin[1], Gregory I. Vladimer[3], Tea Pemovska[3,4], Emir Hadzijusufovic[4], Hermine Agis[4], Maria-Theresa Krauth[4], Wolfgang R. Sperr[4,5], Peter Valent[4,5], Ulrich Jäger[4], Ingrid Simonitsch-Klupp[2], Giulio Superti-Furga[3,6], Philipp B. Staber[4], and Berend Snijder[1]

ABSTRACT

Drug testing in patient biopsy-derived cells can identify potent treatments for patients suffering from relapsed or refractory hematologic cancers. Here we investigate the use of weakly supervised deep learning on cell morphologies (DML) to complement diagnostic marker-based identification of malignant and nonmalignant cells in drug testing. Across 390 biopsies from 289 patients with diverse blood cancers, DML-based drug responses show improved reproducibility and clustering of drugs with the same mode of action. DML does so by adapting to batch effects and by autonomously recognizing disease-associated cell morphologies. In a *post hoc* analysis of 66 patients, DML-recommended treatments led to improved progression-free survival compared with marker-based recommendations and physician's choice–based treatments. Treatments recommended by both immunofluorescence and DML doubled the fraction of patients achieving exceptional clinical responses. Thus, DML-enhanced *ex vivo* drug screening is a promising tool in the identification of effective personalized treatments.

SIGNIFICANCE: We have recently demonstrated that image-based drug screening in patient samples identifies effective treatment options for patients with advanced blood cancers. Here we show that using deep learning to identify malignant and nonmalignant cells by morphology improves such screens. The presented workflow is robust, automatable, and compatible with clinical routine.

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [2]Department of Pathology, Medical University of Vienna, Vienna, Austria. [3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. [4]Department of Medicine I, Division of Hematology and Hemostaseology, Medical University of Vienna, Vienna, Austria. [5]Ludwig Boltzmann Institute for Hematology and Oncology, Medial University of Vienna, Austria. [6]Center for Physiology and Pharmacology, Medical University of Vienna, Vienna, Austria.

T. Heinemann and C. Kornauth contributed equally to the article.

P.B. Staber and B. Snijder contributed equally as co-senior authors of this article.

Current address for G.I. Vladimer: Exscientia GmbH, Vienna, Austria.

Corresponding Author: Berend Snijder, Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Otto-Stern-Weg 3, 8093 Zurich, Switzerland. Phone: 41-44-633-71-49; E-mail: snijder@imsb.biol.ethz.ch

Blood Cancer Discov 2022;3:502–15

doi: 10.1158/2643-3230.BCD-21-0219

INTRODUCTION

Precision medicine aims to identify the best evidence-based treatment for each individual patient (1), historically focusing on molecular markers to guide patient treatment. Examples of successful genetically guided precision medicine for blood cancers include the BCR-ABL1 inhibitor imatinib for patients with chronic myeloid leukemias (CML; ref. 2), and FLT3 inhibition for patients suffering from *FLT3*-mutated acute myeloid leukemia (AML; ref. 3). However, genetically stratified precision medicine is currently estimated to benefit around 10% of all patients with cancer (4), indicating a clear need for additional methods to identify effective treatments.

A complementary route to identifying effective treatments is to examine how cells from the patient respond to drugs in a lab test (4–12). Although conceptually simple, there are many ways to perform such 'functional precision medicine' (FPM) tests, and prospective clinical evidence is increasingly showing that they can bring patient benefit (4, 6, 10, 11). One such FPM approach relies on directly exposing cells isolated from patient biopsies to a panel of drugs, and, after drug incubation, immunofluorescent staining of cells for diagnostic markers (immunofluorescence; IF) and automated microscopy to determine cell type and viability of each cell in the complex biopsies. This approach, which we call pharmacoscopy, is fast, automatable, high throughput, and can test hundreds of treatments from small peripheral blood or bone marrow biopsies. The single-cell resolution and marker-based cell-type classification allows comparison of the drug response of cancer cells with that of healthy cells, providing a patient-internal toxicity control. Furthermore, the spatial and morphologic resolution provided by microscopy captures more complex drug responses, such as immune cell activation and engagement with target cells in response to immunomodulatory drugs (13–15).

In our recently reported clinical study, 56 patients with advanced aggressive hematologic malignancies received treatments guided by pharmacoscopy (4, 6, 10, 16). Of these, 30 patients (54%) achieved a more than 1.3-fold improved progression-free survival (PFS) compared with their previous therapy (4, 6). Furthermore, 12 patients (21%) achieved PFS that lasted three times longer than expected for their respective disease, referred to as "exceptional responses" (4, 17). Cells were identified as malignant ("cancer") or nonmalignant ("healthy") by diagnostic marker IF. *Post hoc* analyses confirmed that the clinical predictions became more accurate when also considering the drug toxicity on the healthy cells within the tested patient sample (4, 6). Given that IF marker–based cell-type identification is sensitive to cancer heterogeneity, antigen loss, and limited antibody specificity, improving cell-type identification in such image-based data could thus further increase the clinical predictive power of pharmacoscopy.

Assessment of cell morphology is fundamental for the diagnosis of cancer (18, 19) and provides rich information for deep learning–based diagnostic tools (20, 21). Recent studies have shown successes using deep learning, particularly convolutional neural networks (CNN), for cell class and state identification in the context of the hematopoietic system in health and disease (14, 22, 23): For example, CNN-based deep learning on live-cell imaging of human hematopoietic progenitor cells predicted lineage commitment outcome several cell divisions prior to the commitment event being evident at the molecular level (22). CNN-derived features predicted genetic aberrations from bone marrow histopathology imaging for patients with myelodysplastic syndrome (23). And, in the context of automated microscopy of peripheral blood mononuclear cells (PBMC), CNNs enabled deconvolution of multiplexed immunofluorescence staining in a high-throughput setting, based in part on morphologic differences between cell types (14, 15). CNNs consist of different "layers" representing different mathematical transformations, the combination of which makes up the "architecture" of the CNN. Given the large number of possible and published CNN architectures, an open question in CNN research is how to identify the most suitable architecture for a given problem.

In this study, we set out to investigate the use of deep learning using CNNs to morphologically classify nonmalignant and malignant cells (termed "deep morphology learning"; DML) in the context of *ex vivo* drug screening. We evaluate different CNN architectures and identify a relatively small CNN (mCNN) with good DML performance. Throughout the study, we contrast DML by mCNN to DML by a larger and state-of-the-art CNN (ResNet; ref. 24). Analyzing drug-response data across 1.3 billion single cells from 390 individual drug screens, we find improved performance for DML compared with marker-based analysis by IF and show that this is associated with robustness to batch effects and autonomous identification of disease-typical cell morphologies. The impact of DML on the clinical predictive power of *ex vivo* drug screening is assessed in a *post hoc* analysis of 66 patients with multiyear clinical follow-up data. This reveals improved PFS for treatments recommended by DML-based drug screening compared with the IF-based recommended treatments. Significant enrichment of exceptional responders is further observed for patients receiving treatments recommended simultaneously by IF and DML-based analyses, indicating diagnostic markers and cell morphology are complementary in cancer cell identification. Thus, DML deepens the insights derived from image-based *ex vivo* drug screening and enhances its clinical predictive power for the personalized treatment of relapsed and refractory hematologic malignancies.

RESULTS

We performed image-based drug screening in 390 real-time biopsies from 289 individual patients, collecting multiyear clinical follow-up for a subcohort of 66 prospectively treated patients reported previously (4, 6, 13, 16, 25). The combined data set comprises a total of 1.3 billion patient cells across 136 *ex vivo* tested drugs (Supplementary Tables S1–S4). The screens were performed over a period of 3 years and across diverse and heterogeneous hematologic diagnoses, including AMLs, T-cell lymphomas, diffuse large B-cell lymphomas (DLBCL), chronic lymphocytic leukemias (CLL), and multiple myeloma (MM). Diagnosis and tumor cell content were confirmed by routine clinical pathology. Real-time patient samples were incubated for 24 hours in 384-well format microtiter plates containing the drug library, and afterward fixed and stained. The drug screens were imaged by automated

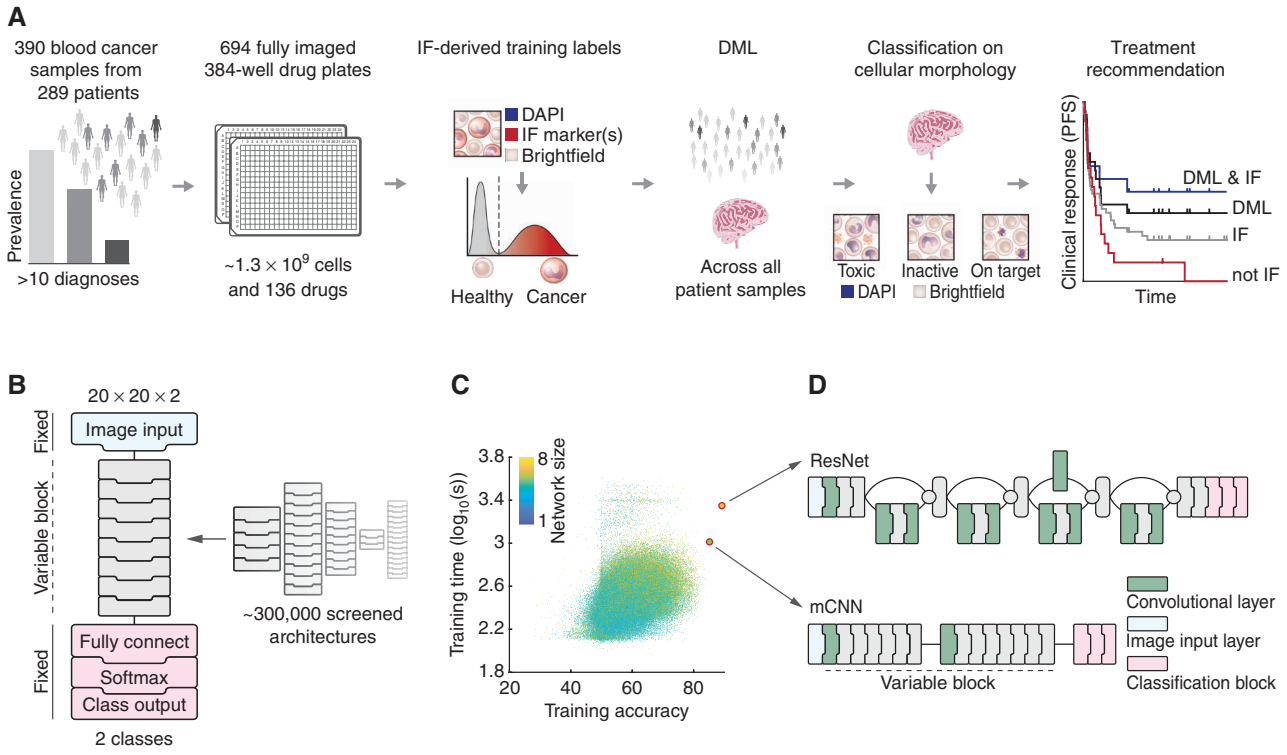


Figure 1. Screening neural network architectures for DML. **A,** Workflow and data set used for DML. 390 blood cancer samples were screened *ex vivo* on a library of 136 drugs, followed by automated confocal microscopy and single-cell image analysis. The final data set encompasses 696 fully imaged 384-well plates, imaging over 1.3 billion single cells in 5 channels: brightfield, DAPI, and three channels used for IF staining of up to three markers identified by clinical diagnostics for each sample. DML uses CNNs trained to recognize malignant and healthy cells from just the DAPI and brightfield channels. CNN training is weakly supervised by marker immunofluorescence. DML-based drug scores were used to stratify patient PFS in a *post hoc* analysis. **B,** Graphical presentation of the randomized architecture screen. mCNN was selected based on training time, network size, and test accuracy from over 290,000 screened architectures. **C,** Scatter plot of training accuracy and corresponding training time of screened architectures. Storage size of the network data structures are color-coded. **D,** Graphical comparison of the mCNN and ResNet architectures.

confocal microscopy, recording a DNA dye (DAPI) capturing the nuclear morphology, brightfield (BF) imaging capturing cytoplasmic cell morphologies, as well as IF against up to three disease-matched diagnostic markers (see Supplementary Table S4 for the markers used per sample).

Weakly Supervised CNN Training for Label-Free Cancer and Healthy Cell Classification

We set out to develop and use deep CNNs to classify cancer and healthy cells strictly by their nuclear and cytoplasmic morphologies, termed DML (Fig. 1A). In the absence of a single-cell level ground truth, we use the diagnostic marker expression measured by IF to automatically label single-cell centric image crops (20 × 20 pixels per channel) as either "cancer" or "healthy," and train the CNNs to reproduce these labels from just the DAPI and brightfield image crops (i.e., 20 × 20 × 2 input images). Such so-called weak supervision allows training at a scale not easily attainable with manual curation, which can help CNNs learn robust DAPI- and brightfield-derived features that differentiate cancer from healthy cells. The increased training data set size can further protect against CNN overfitting on experimental, technical, or lineage-specific features and avoid annotator bias.

Many CNN architectures with excellent performance in image classification tasks have been previously reported (26); however, none have been optimized for this weakly supervised classification task. We systematically explored different possible CNN models by training 291,800 different neural network architectures on 5,000 cells labeled as cancer cells and 5,000 cells labeled as healthy. For each CNN, we evaluated their training time, model size, and training accuracy, i.e., how well the CNN-predicted labels match the IF-derived labels on the training data (Fig. 1B). As a control, we trained ResNet, a previously published and considerably more complex CNN architecture with state-of-the-art performance for image classification purposes (ref. 24; Fig. 1C and D; Supplementary Fig. S1). None of the randomized architectures outperformed the training accuracy of ResNet (Fig. 1C). However, the analysis identified a CNN architecture (referred to as mCNN) that was considerably smaller than ResNet, while reaching a similar test accuracy on 1,000 previously unseen cells per sample and class (85% for mCNN vs. 89% for ResNet; Fig. 1C and D). Smaller models with fewer parameters are generally less prone to overfitting, i.e., performing well on training data but poorly on data on which it was not trained. We, therefore, continued our analyses with both mCNN and ResNet.

Given that the used diagnostic markers were predominantly cell lineage markers (Supplementary Table S4), training a CNN on the data of a single sample stained with a single marker (marker-sample pair) might lead to the identification

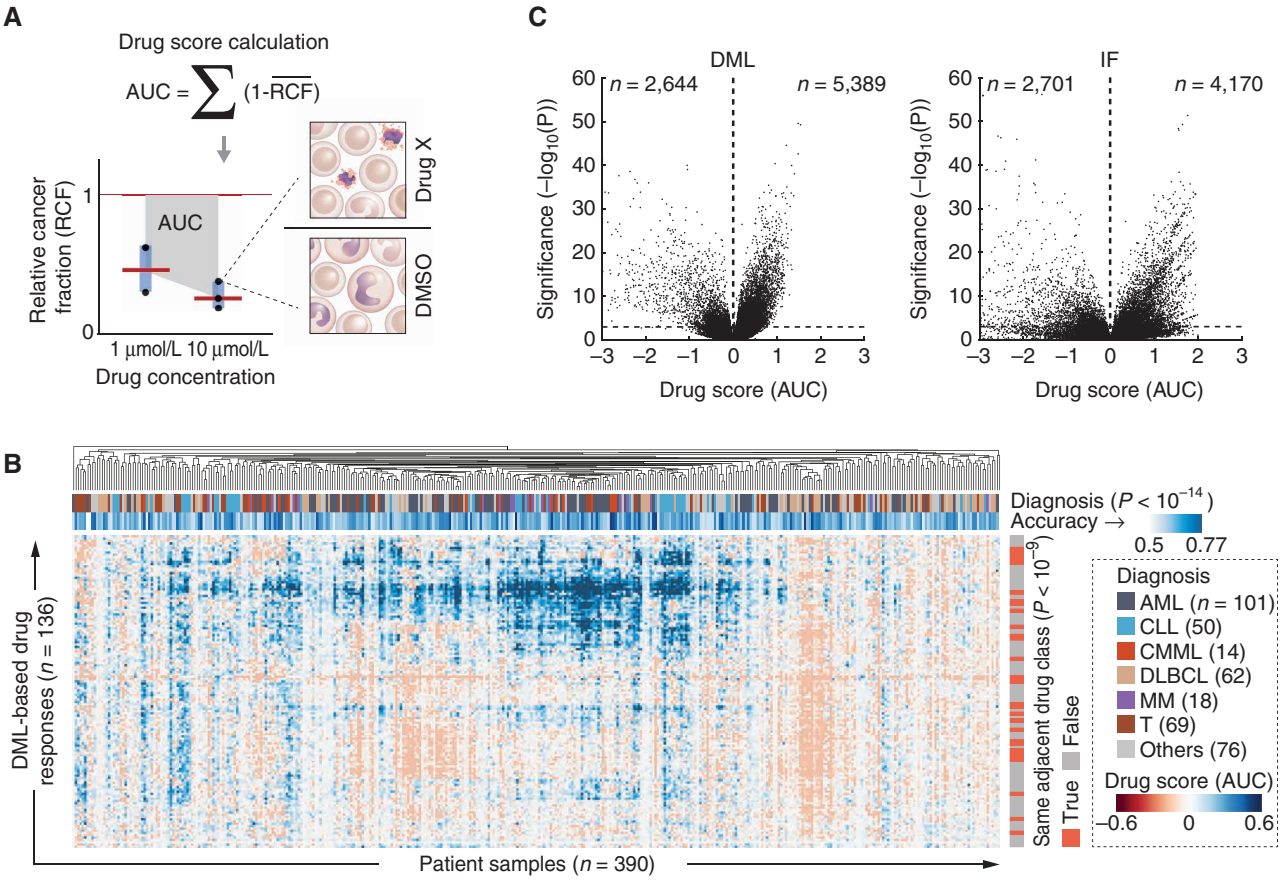


Figure 2. DML-based cancer cell identification improves results from massively parallel *ex vivo* drug screening. **A,** Schematic example of the drug score calculation (AUC = area under the curve). **B,** Heat map of clustered DML-based drug-response scores (AUCs; see color bar) for drugs (rows) and patient samples (columns). Drug signature similarity is calculated with 1 – the Pearson correlation and graphically represented as node distance in a hierarchical binary tree above the heat map. DML was performed using the mCNN architecture. Adjacent signatures associated with the same drug class are indicated on the right side of the heat map in red. Each sample-associated test accuracy and corresponding diagnosis are color-coded on top of the heat map. The accuracy is calculated as the percentage of matching DML and IF single-cell labels assigned to a test set (unseen during training) of 2,000 randomly selected cells per sample, equally split across both classes. *n* = number of patient samples/drugs. The *P* values are derived from hypergeometric testing. **C,** Volcano plots showing the significance [$-\log_{10}$(*P* value); *y*-axis] and effect size (AUC; *x*-axis] per drug for all drugs and samples, analyzed either by DML (left) or IF (right). Significance was calculated by a two-sided Student *t* test comparing the drug score replicate wells with the negative control DMSO wells per sample. *n* = number of drug scores that are located in the corresponding quadrant as confined by the black dashed outlines. (*continued on next page*)

of the cell lineage as opposed to the identification of malignant cells. Training a CNN across heterogeneous samples—although more challenging—is less likely to lead to the identification of just lineage features, as it presents malignant and nonmalignant cells from across diverse lineages. To evaluate this, we compare the accuracy of mCNN trained per marker-sample pair (defined as drug screen of one patient sample analyzed by a single marker), with that of an mCNN trained across cells from all samples (Supplementary Fig. S2A). Individual mCNNs were trained per marker-sample pair on 5,000 cancer and 5,000 healthy labeled cells. The pan-sample mCNN was trained on 586,500 cells (0.05% of all imaged cells) evenly subsampled from all 390 samples. This training data set was equally split across malignant and nonmalignant cell labels, and labels were based for each sample on the marker with the best performance in the per marker-sample pair mCNN training. Strikingly, training per marker-sample pair only achieved modestly higher test accuracies compared with mCNN trained across all samples, and accuracies per sample were correlated (r_{SP} = 0.63; Supplementary Fig. S2A and S2B). This shows that mCNN classification accuracy would not be improved much by limiting the training data to cells from a single sample.

DML Improves Drug-Response Characteristics

The labels on which we trained mCNN are, however, themselves imperfect in distinguishing malignant cells from nonmalignant cells. As a result, the interpretation of classification accuracies in this so-called weakly supervised setting has limited value. Therefore, we next asked, notwithstanding these uncertainties, whether our DML-based cancer cell classification could improve our ability to identify cancer cell response to drugs (Fig. 2A) across all 390 samples and 136 drugs (Fig. 2B; Supplementary Table S5). As a drug-response metric, we calculated the area under the curve (AUC) across concentrations, previously reported to be a robust drug effect estimator (ref. 27; Fig. 2A). Positive AUCs denote an on-target drug-induced reduction in the fraction of cancer cells

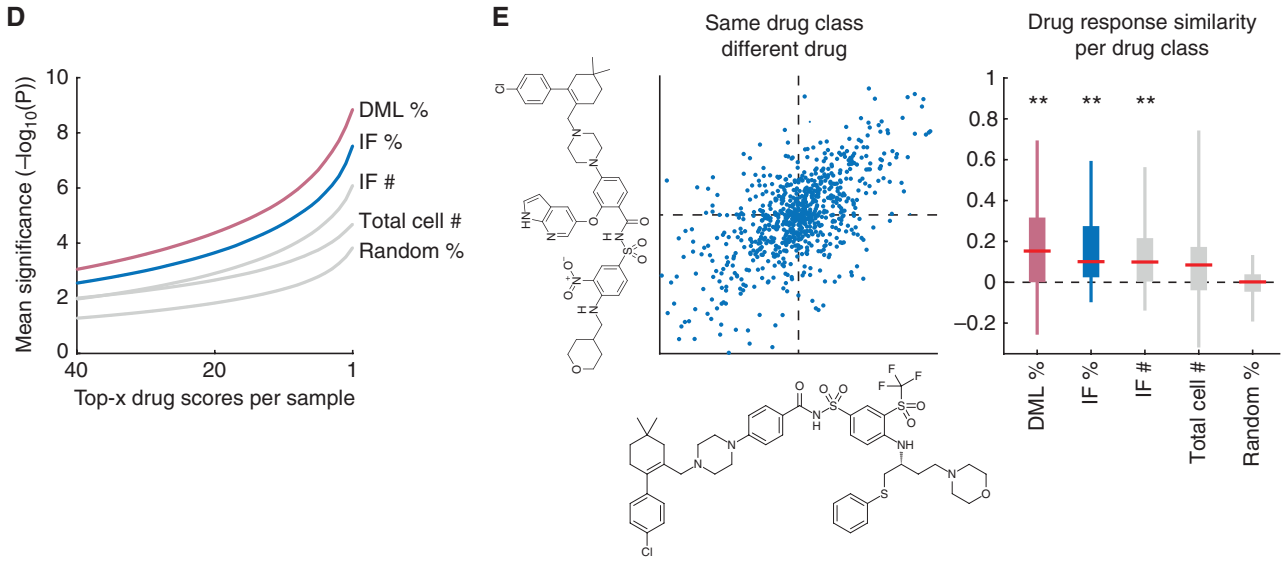


Figure 2. ***(Continued)*** **D,** Mean significance [–$\log_{10}$(*P* value); *P* values are derived as in **C**; *y*-axis] of the top-x strongest drug responses per sample (*x*-axis) compared for different drug-screen readouts. IF indicates marker-based cancer cell identification. % indicates change in the target cell fraction relative to control. # indicates change in the number of target cells relative to control. "Total cell #" indicates change in total cell number relative to control. Random indicates randomized cancer cell classification. **E,** Box and whisker plots comparing the similarity in drug responses (pairwise Pearson correlation) across samples for drugs sharing the same class. Only drug classes with at least two associated drugs were considered. Red bar depicts the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. **, $P \leq 0.01$. The *P* values indicate the significance of drug-response similarity and were derived from two-tailed Student *t* tests.

compared with DMSO control, whereas AUCs around 0 indicate no change, and negative AUCs indicate a relative increase in the cancer cell fraction compared with DMSO. Hierarchical clustering across this DML-based drug-response landscape revealed several striking observations. We noted a significant grouping of samples by their diagnosis ($P < 10^{-14}$; Fig. 2B), which was not observed for repeated testing on randomly permuted data (Supplementary Fig. S2C). Furthermore, drugs with the same mechanism of action (annotated in Supplementary Table S1) significantly clustered across the landscape ($P < 10^{-9}$; Fig. 2B). And the significance of the strongest on-target drug responses (or "hits"; measured by Student *t* test on the biological repeat measurements per condition against the DMSO-based negative controls of each screen) was better for DML than for IF (Fig. 2C). Given fixed effect size and significance thresholds, more such on-target hits were obtained for DML ($n = 5{,}389$) compared with IF ($n = 4{,}170$) across the full set of image-based drug screens (Fig. 2C). Improved reproducibility for DML was also observed when comparing the mean significance of top on-target drug responses across samples (Fig. 2D). In this analysis, we included drug responses based on the relative drug scores of either DML (DML %) or IF (IF %), as well as on the total number of marker-positive cells relative to control (IF #), the total cell number measured in each condition relative to control (total cell #), and single-cell level randomized drug responses (random %). Across all different readouts, DML-based readouts showed the most significant top hits (Fig. 2D).

For comparison, we also trained ResNet on the same pan-sample data set of 586,500 cells that mCNN was trained on. For clarity, please note that we use "DML" to denote deep morphology learning using the mCNN architecture trained on all 586,500 cells, unless indicated otherwise. In the comparison between DML by mCNN and DML by ResNet, we observed that top hits from ResNet were even more significant (Supplementary Fig. S2D). Lastly, clustering of drugs with the same mode of action was also strongest for DML-based results compared with the other screening readouts (Fig. 2E). In this test, DML by mCNN and by ResNet showed equally good performance (Supplementary Fig. S2E). Thus, cancer cell classification by DML, irrespective of the CNN architecture, resulted in drug-response profiles that outperformed marker-based cancer cell identification by IF in orthogonal technical and biological evaluation criteria.

Opening the Black Box: Analyzing Features by Which Cells Are Clustered in mCNN's Latent Space

CNNs learn a multidimensional data representation called the latent space, in which images that look similar to each other group closer together. Analysis of latent-space clustering of images is thus an effective way to get a better understanding of what a CNN has learned. We, therefore, asked whether morphologically similar samples and similar diagnoses were clustered together in the multidimensional latent space of the CNN. At the technical level, we extracted the output from mCNN's last fully connected layer per cell (Fig. 3A; Supplementary Fig. S1), and calculated the pairwise sample similarity in this latent-space representation. We defined sample similarity as statistical enrichment in the K-nearest neighbors (KNN) between the cells of two samples, which we refer to as the deep morphologic similarity (DMS) score (Fig. 3A). Hierarchical clustering of the DMS scores showed strong separation between cancer and healthy cell classes, each further divided into three subclusters containing samples from

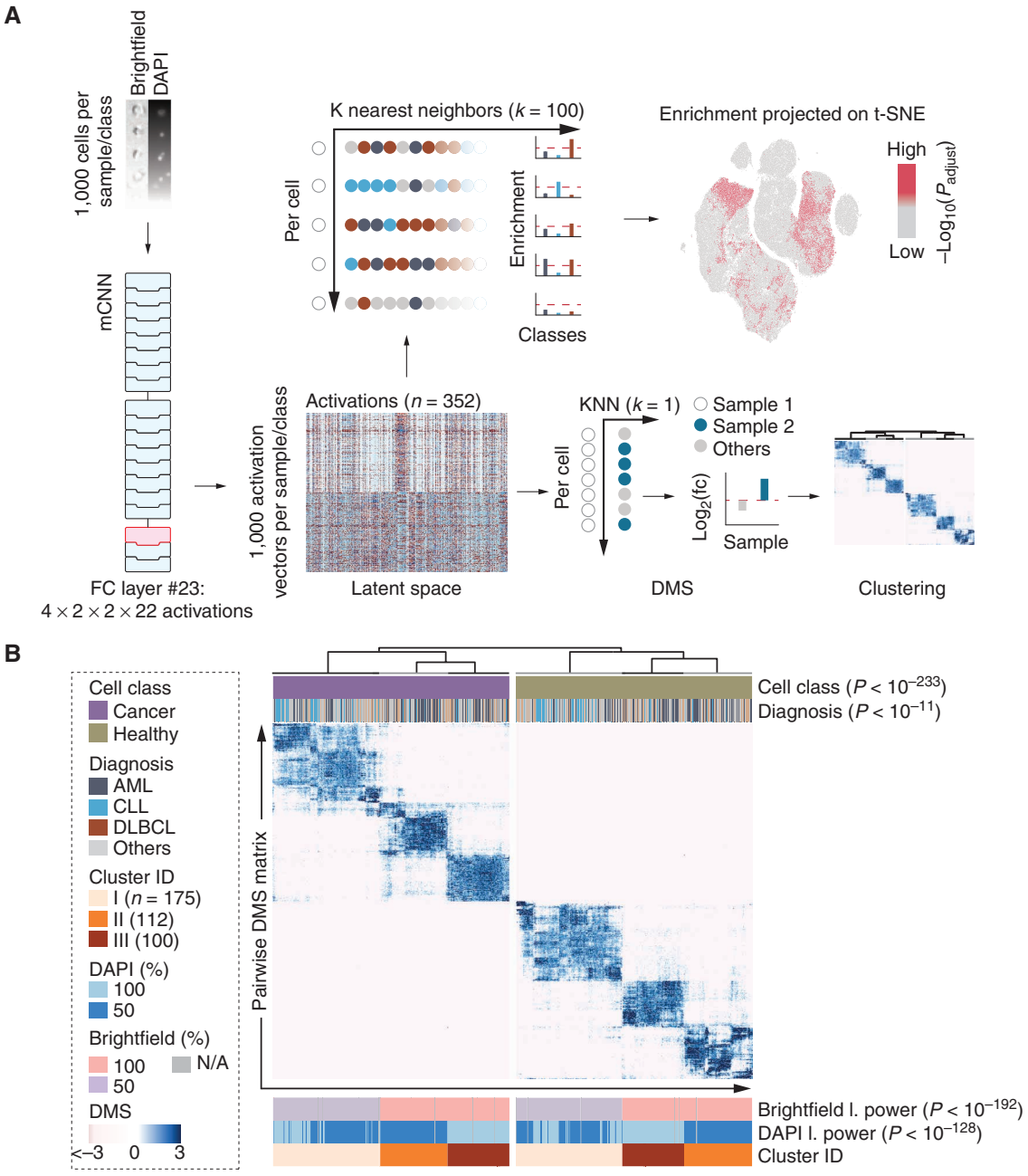


Figure 3. DML autonomously recognizes diagnosis-enriched cell morphologies. **A,** Workflow for the analysis of mCNN's latent space (left) by DMS (bottom right) and single-cell KNN enrichment (top right). Feedforward-propagation-derived activations of the last fully connected (FC) mCNN layer are analyzed. Bottom right, DMS is calculated by quantifying hypergeometric enrichment in the latent-space nearest neighbors between two samples, among the full set of samples. 1,000 cells per class and sample are used. Top right, Hypergeometric enrichment for categories (e.g., diagnosis) in the 100 nearest neighbors in latent space per cell. Enrichments are projected onto the t-SNE embedding. fc = fold change. **B,** Hierarchical clustering of the pairwise DMS matrix for 1,000 cells per class and sample. Leaf identity is color-coded according to the corresponding cell class (cancer, healthy), diagnosis, the laser power settings in the brightfield and DAPI channels, and the assigned cluster ID. The *P* values indicate cluster significance and are derived from hypergeometric testing. (*continued on next page*)

multiple diagnoses (Fig. 3B). A closer investigation indicated that these three subclusters reflected power settings of the light source used during brightfield ($P < 10^{-192}$) and DAPI ($P < 10^{-128}$) imaging, which had been stepwise altered over the course of 3 years of continuous biopsy screening (Fig. 3B; Supplementary Fig. S3A). Given that the cell class separation was stronger than the batch-associated clustering, mCNN had recognized and learned to overcome technical batch effects in the data.

DML Autonomously Learns Diagnosis-Associated Cell Morphologies

The clustering of DMS scores additionally showed significant grouping of samples with the same diagnosis ($P < 10^{-11}$;

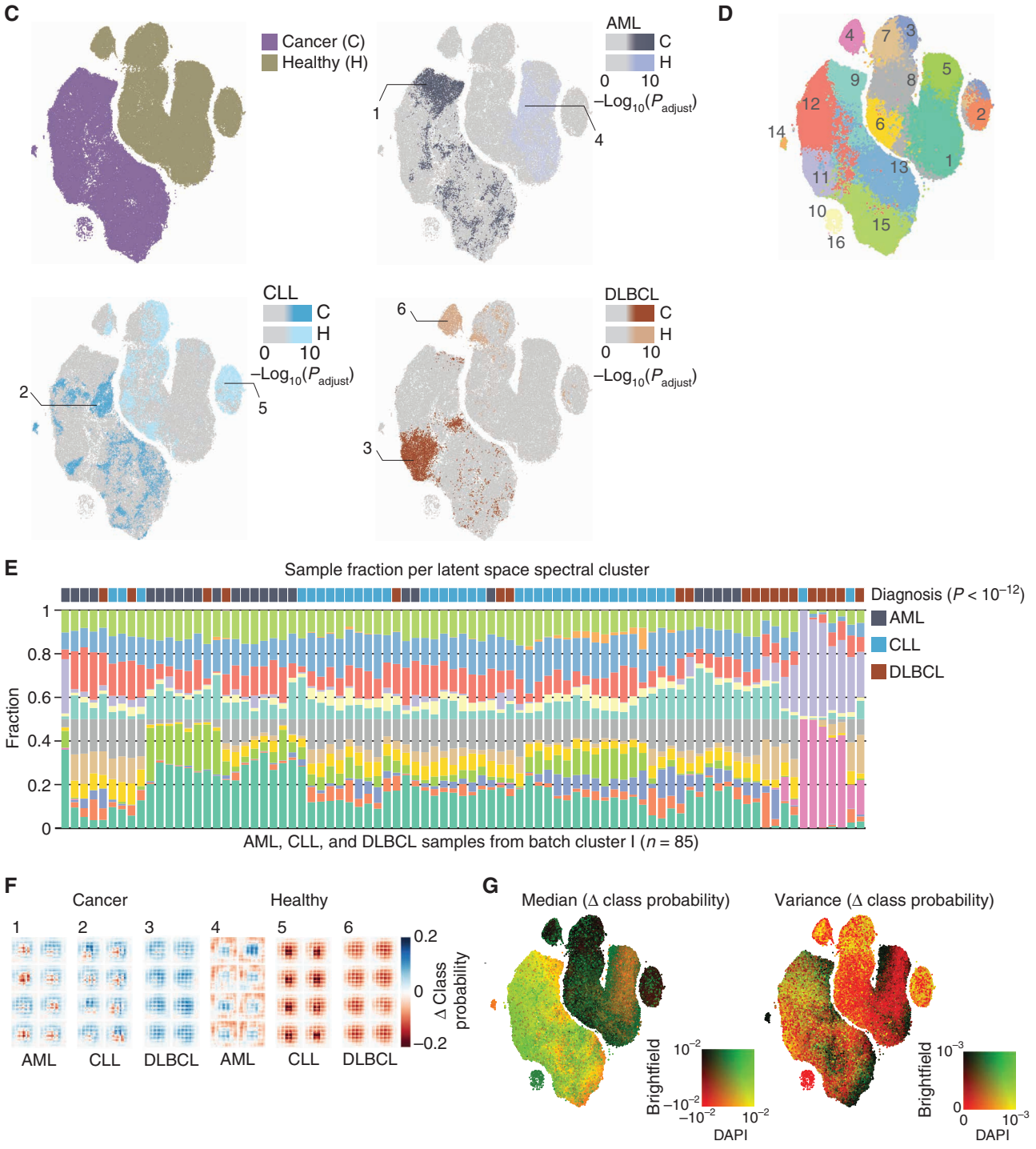


Figure 3. ***(Continued)*** **C,** t-SNE embedding of mCNN activations for 350,000 cells (1,000 cells per class from 175 patient samples from cluster I in **B** colored according to cell class (top left) or to the diagnosis/class nearest neighbor enrichment in latent space for AML (top right), CLL (bottom left), or DLBCL (bottom right). P_{adjust} values are derived from hypergeometric testing in the 100 nearest neighbors per cell. **D,** Spectral clustering of mCNN's latent-space projected onto the t-SNE embedding. **E,** Bar graph depicting fractions of cells calculated per sample and latent-space spectral cluster from **D**, for AML, CLL, and DLBCL samples from batch cluster I ($n = 85$). The corresponding clinical diagnosis for each sample is indicated above. The *P* value denotes cluster significance and is derived from hypergeometric testing. **F,** mCNN's delta class probabilities per image subregion for eight example cell images from each of the diagnosis- [AML (1, 4), CLL (2, 5), DLBCL (3, 6)] and class- [cancer (1, 2, 3), healthy (4, 5, 6)] enriched t-SNE regions in **C**. For every input image, each pixel was masked with a square (5 × 5 pixels) of equal intensity in the brightfield and DAPI channel. **G,** Median and variance of the delta class probabilities per cell image projected on the t-SNE–clustered mCNN latent space. Metrics were calculated for each image individually in the DAPI and brightfield channels, and t-SNE embedding is colored according to a 2D color map reflecting results for each channel.

Fig. 3B). As leukemias and lymphomas are morphologically distinct (28, 29), this hinted at the possibility that mCNN had learned to recognize diagnosis-associated cell morphologies. To further investigate this, we visualized the t-SNE embedding of the single-cell latent space, only considering samples from cluster ID I to reduce batch effects as possible confounding factors (Fig. 3B). The t-SNE embedding suggested considerable subclustering within both healthy and malignant classes, which we investigated in the context of our three most frequent diagnoses in our cohort: AMLs, CLLs, and DLBCLs (Fig. 3C). KNN enrichment analysis in the latent space (Fig. 3A) showed distinct regions enriched for cells from each diagnosis (visualized on the t-SNE embedding in Fig. 3C), despite AMLs and DLBCLs comprising a variety of subtypes (30, 31). Notably, this latent-space clustering of cells from the same diagnosis was significantly stronger for the mCNN architecture than for the deeper ResNet architecture ($P < 0.009$; Supplementary Fig. S3B). We further clustered the cells in latent space using a graph clustering approach (called spectral clustering) and quantified the fraction of cells per cluster and sample (Fig. 3D). This confirmed that these morphologic single-cell signatures were not an artifact of a few samples, but were observed across all samples (Fig. 3E). Furthermore, analyzing the sample similarity based on these cluster frequencies resulted in highly significant grouping of samples with the same diagnosis ($P < 10^{-12}$; Fig. 3E). Finally, we could confirm that these latent-space clusters partially captured differences in interpretable cellular features, including DAPI intensity and nucleus size. For example, we found that the AML-enriched cancer subcluster #9 was characterized by particularly large nuclei (Supplementary Fig. S3C and S3D).

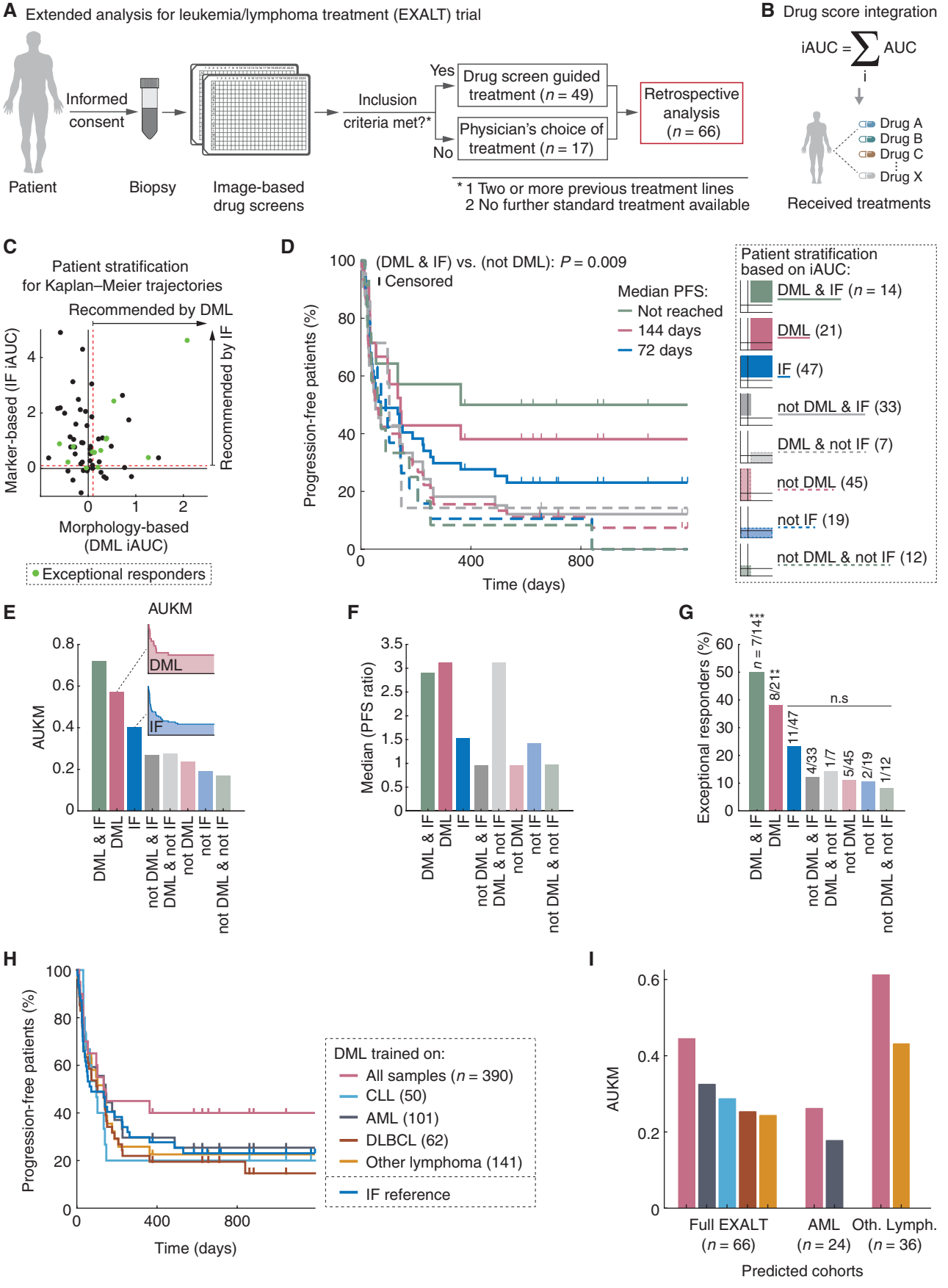
The presence of distinct diagnosis-enriched cell morphologies detected by DML was further investigated by measuring the response of mCNN to partially masking the input images. This showed that mCNN's classification confidence was sensitive to masking distinct image subregions and channels in a cell class- and diagnosis-dependent manner (Fig. 3F; Supplementary Fig. S3E). Projecting the median and variance of the change in confidence per cell crop onto the t-SNE embedding of the latent space indicated between- and within-class sensitivity differences (Fig. 3G). For example, healthy cell classification was strikingly more sensitive to masking of the brightfield channel than to masking of the DAPI channel, independent of the patient diagnosis. This indicates that the brightfield channel adds important confidence to a healthy cell classification. Thus, while learning to classify healthy and malignant cells across the entire patient cohort, the tailored mCNN architecture had recognized and adapted to batch effects and autonomously learned to recognize diagnosis-associated cell morphologies.

Treatments Recommended by Both IF and DML Double Fraction of Patients Achieving Exceptional Clinical Responses

On-target drug responses measured by pharmacoscopy, evidenced by positive AUC scores, can be taken as a treatment recommendation for patients who have exhausted standard therapies. Indeed, we have previously shown that treatments recommended by IF-based pharmacoscopy for patients enrolled in the Extended Analysis for Leukemia/Lymphoma Treatment (EXALT) study (Fig. 4A) led to improved PFS compared with the patient's own response to their prior treatment (4, 6, 10, 16). This positive association between on target *ex vivo* drug response and good clinical response was also previously confirmed by *post hoc* analysis (4). Here, we use this same *post hoc* analysis strategy of the 66 patients included in both the EXALT trial and the current study to see how well the patients responded to the DML-recommended treatments compared with IF-recommended treatments. The 66 patients included 24 patients suffering from relapsed/refractory AML and 36 patients suffering from relapsed/refractory B- or T-cell lymphomas, further characterized in Supplementary Table S2. As the treatments that the patients received following pharmacoscopy contained multiple drugs for most patients, we quantified the *ex vivo* support for such combination treatment as the integrated drug responses (integrated AUC score; iAUC; Fig. 4B). We take positive iAUCs above 0.1 to indicate *ex vivo* support for the treatment, whereas lower scores indicate the treatment is not recommended by pharmacoscopy.

Comparing the IF- and DML-based treatment recommendations for the 66 patients showed that 47 patients received treatments recommended by IF-based pharmacoscopy, whereas only 21 patients received treatments supported by DML (Fig. 4C). *Post hoc* analysis indicated that treatments recommended (iAUC >0.1) by either DML alone, or simultaneously by DML and IF, showed clinical benefit by several metrics. DML-supported treatments were associated with prolonged median PFS compared with IF-supported treatments (although significance not reached by the Wilcoxon rank test; Fig. 4D; Supplementary Fig. S4A), with a near doubling of median PFS, from 72 days for IF to 144 days for DML, and median not reached for treatments recommended by both DML and IF (Fig. 4D; Supplementary Fig. S4A). As the survival curves were characterized by heavy tails that were not captured by the median PFS, we further quantified clinical performance by three metrics: (i) the normalized "area under the Kaplan–Meier curve" (AUKM; Fig. 4E) as an ad hoc measure of integrated clinical benefit; (ii) the median PFS ratio, which compares the patient's prior PFS response to PFS achieved following testing (Fig. 4F); (iii) and enrichment for so-called exceptional responders, previously defined as patients achieving responses lasting three times longer than expected for their respective disease (ref. 4; Fig. 4G). By all three measures, treatments recommended by either DML alone, or by both DML and IF, outperformed treatments recommended by IF alone. Enrichment for exceptional responders was particularly striking: 38% (8 of 21) of patients receiving DML-recommended treatments ($P < 0.0024$) and 50% (7 of 14) of patients receiving treatments recommended by both DML and IF ($P < 0.00041$) led to exceptional clinical responses.

DML-recommended treatments further significantly improved median PFS compared with treatments not recommended by DML ($P < 0.025$; Fig. 4D and E; Supplementary Fig. S4A). And patients receiving treatments that were recommended by IF but not by DML showed poor PFS, indicating DML identified a number of false-positive treatment recommendations from IF (Fig. 4D and E; Supplementary Fig. S4A). Nonetheless, both


A Extended analysis for leukemia/lymphoma treatment (EXALT) trial
Patient
Informed consent
Biopsy
Image-based drug screens
Inclusion criteria met?*
Yes
No
Drug screen guided treatment (n = 49)
Physician's choice of treatment (n = 17)
Retrospective analysis (n = 66)
* 1 Two or more previous treatment lines
2 No further standard treatment available
B Drug score integration
iAUC = ∑ AUC
i
Drug A
Drug B
Drug C
Drug X
Received treatments
C
Patient stratification for Kaplan–Meier trajectories
Recommended by DML
Recommended by IF
Marker-based (IF iAUC)
Morphology-based (DML iAUC)
Exceptional responders
D
(DML & IF) vs. (not DML): P = 0.009
Censored
Median PFS:
Not reached
144 days
72 days
Progression-free patients (%)
Time (days)
Patient stratification based on iAUC:
DML & IF (n = 14)
DML (21)
IF (47)
not DML & IF (33)
DML & not IF (7)
not DML (45)
not IF (19)
not DML & not IF (12)
E
AUKM
DML
IF
DML & IF
DML
IF
not DML & IF
DML & not IF
not DML
not IF
not DML & not IF
F
Median (PFS ratio)
G
Exceptional responders (%)
n = 7/14***
8/21**
11/47
4/33
1/7
5/45
2/19
1/12
n.s
H
Progression-free patients (%)
Time (days)
DML trained on:
All samples (n = 390)
CLL (50)
AML (101)
DLBCL (62)
Other lymphoma (141)
IF reference
I
AUKM
Full EXALT (n = 66)
AML (n = 24)
Oth. Lymph. (n = 36)
Predicted cohorts

IF- ($P < 0.01$; consistent with the EXALT study; refs. 4, 6) and DML- ($P < 0.002$) recommended treatments showed improved clinical responses compared with prior treatment based on the physician's choice (Supplementary Fig. S4B).

The lower number of treatments recommended by DML coinciding with their improved clinical performance thus suggests that morphologic cancer cell classification is robust to IF-based false-positive treatment recommendations, possibly stemming from drug-induced loss of marker expression. Such differences in false positives between IF and DML are a likely explanation for why their intersection is associated with the best clinical responses.

Network Architecture and Pan-Diagnosis Training Are Critical for DML Performance

Given the clinical heterogeneity of the 66 patients included in the retrospective analysis, we analyzed if the advantages resulting from DML-based cancer cell classification were observed across different patient subgroups. Indeed, both DML alone and the combination of DML and IF identified the most effective personalized treatments for the 24 included patients with AML (Supplementary Fig. S4C), for the 36 included patients with lymphoma (Supplementary Fig. S4D), and for fitter patients (as identified by ECOG performance status ≤1; Supplementary Fig. S4E, F). These improvements were consistently seen using either Kaplan–Meier curves, AUKMs, or median PFS as clinical performance metrics.

Lastly, we evaluated if the CNN architecture and training regime influenced the clinical response associated with DML-recommended treatments. Despite the state-of-the-art ResNet architecture achieving a higher training accuracy than the smaller mCNN architecture (Fig. 1C), mCNN outperformed ResNet with regard to the PFS achieved in response to their recommended treatments, both as a standalone recommendation and when combined with IF-recommended treatments (Supplementary Fig. S4G and S4H). Furthermore, training DML only on the samples of an individual diagnosis (as tested for AML, CLL, DLBCL, and other lymphomas) entirely annulled the improvements in PFS and AUKM achieved by DML-based treatment recommendations compared with IF (Fig. 4H and I). Even when evaluating the PFS of patients treated for the same diagnosis as DML was trained on (for either AML or lymphomas), no improvements over IF-based treatment recommendations were observed (Fig. 4H and I; Supplementary Fig. S4I and S4J). Thus, as originally postulated, weakly supervised pan-diagnosis training with a slim neural network architecture had enabled DML to rise above the limitations inherent to marker-based cancer cell identification, thereby improving *ex vivo* image-based drug-response testing for personalized treatment identification.

DISCUSSION

There is an urgent need for methods that identify effective therapies for patients suffering from relapsed and refractory cancer. Here, we introduce an automated workflow for weakly supervised DML to classify malignant and healthy cells in patient biopsies and show that it leads to highly actionable clinical treatment recommendations from image-based *ex vivo* drug screening for hematologic cancers.

Our results are at first sight counterintuitive: DML by a small CNN architecture (mCNN) analyzing only two of our five imaging channels outperforms both patient-tailored marker expression measured by immunofluorescence as well as DML based on a deeper and more accurate ResNet classifier. mCNN-based drug-response profiles showed an increased number of significant on-target hits, superior clustering of drugs with the same mechanism-of-action, stronger identification of diagnosis-enriched cell morphologies, and its treatment recommendations were associated with better clinical outcomes.

mCNN achieved these improvements by weakly supervised training on 586,500 (out of 1.3 billion) imaged cells from 390 drug screens in real-time patient biopsies, labeled by the marker expression of each individual cell. Given that ResNet achieved higher test accuracies with a deeper residual network architecture, the smaller mCNN likely induced it to better generalize features of malignant and healthy cells across diagnoses. This interpretation is supported by (i) the fact that mCNN's latent-space representation of the data more strongly clustered cells from the same diagnosis, (ii) that mCNN *post hoc* identified more effective treatments across both leukemias and lymphomas, and (iii) that the improved performance of mCNN was lost when trained on multiple samples with the same diagnosis, rather than across multiple diagnoses. Furthermore, the custom mCNN architecture resulting from the architecture screen might have helped it adapt to technical batch effects present in the training data. Thus, we show that optimizing deep learning just on accuracy

Figure 4. DML-based pharmacoscopy identifies clinically effective personalized treatment options. **A,** Outline of the EXALT trial (4, 6). **B,** Graphic presentation of the integrated drug score calculation (iAUC) per patient. AUC scores are integrated for all tested drugs each patient subsequently received in the clinic. **C,** Scatter plot of marker- (IF; *y*-axis) and morphology-based (DML; *x*-axis) iAUCs per patient. Selected strong negative values not shown for readability. Dashed red lines indicate the iAUC thresholds of 0.1, used to determine if patient treatments were supported by either IF and/or DML-based pharmacoscopy. Exceptional responders are colored in green. **D,** Kaplan–Meier plot showing the percentages of progression-free patients with treatment regimens supported by iAUCs > 0.1 for: DML and IF (green, $n = 14$ patients); DML (pink, $n = 21$); IF (blue, $n = 47$); IF but not DML (light gray, $n = 33$); DML but not IF (dashed gray, $n = 7$); not DML (dashed pink, $n = 45$); not IF (dashed blue, 19); or not DML and not IF (dashed green, $n = 12$); see legend on the right. Tick marks indicate censored data points (i.e., ongoing responses). The *P* value indicates Wilcoxon rank test significance. **E,** Area under the Kaplan–Meier curve (AUKM; see insets on the right for DML and IF examples) corresponding to the patient subgroup PFS curves shown in **D**. **F,** Median PFS ratio comparing the patients' current treatment response (following pharmacoscopy testing) and the responses to their previous treatment, stratified as in **D**. **G,** Percentage of patients achieving exceptional clinical responses, stratified as in **D**. **H,** Kaplan–Meier curves comparing the PFS of EXALT patients ($n = 66$) receiving treatments *post hoc* recommended by either DML trained on all patient samples (pink; $n = 390$) or DML trained on all CLLs (light blue; $n = 50$), AMLs (gray; $n = 101$), DLBCLs (brown; $n = 62$), or all other lymphoma samples (dark yellow; $n = 141$). The PFS curve of patients receiving treatments recommended by IF is shown as reference (blue). **I,** AUKM values corresponding to **H** (left) or corresponding to the subset of EXALT AML patients ($n = 24$; middle) or the subset of EXALT "other lymphoma" patients ($n = 36$; right). Bar colors corresponding to DML trained on different sample subsets as shown in **H**. **A**, **D**, **G–I**, n = number of patients. **G,** **, $P \leq 0.01$; ***, $P \leq 0.001$. *P* values are derived from hypergeometric testing.

is not always ideal. Rather, screening CNN architectures to balance the trade-off between classification accuracy and network size yielded the best biological and clinical performance. This concept, and the developed computational workflow, will likely apply to other biomedical questions where training data labels are imperfect.

Functional assays that probe drug sensitivity and resistance directly in patient biopsies are increasingly used to guide patient treatment and complement molecular tumor profiling for clinical decision support (4–7, 9–12, 32–35). These functional screens stand out as they can provide evidence for treatments that lack molecular response predictors, identify drug sensitivities for cancers that do not harbor actionable molecular markers, and prioritize among therapies associated with the same biomarker, even on an $n = 1$ patient basis (8, 36). Our concomitant clinical study shows that our image-based approach is feasible in clinical routine and beneficial for the treatment of aggressive hematologic malignancies (4, 6). With the development of DML, we here demonstrate the power of combining cell morphology, a bedrock of clinical pathology, with deep learning across massively parallel image-based *ex vivo* drug screening for precision medicine. Several features of DML make it highly attractive for inclusion in future clinical studies: DML works in parallel and complementary to established experimental protocols; it can learn to be robust to technical and biological variation; it can learn from previously generated data sets; training data can be expanded by experimentation as well as manual curation; and retraining is fast. Furthermore, by training the deep learning on single-cell classification and deriving the clinical treatment recommendations from the resulting drug-screen analysis, rather than directly training the network on the identification of effective treatments, we avoid overfitting-associated limitations that have hampered the adoption of deep learning in clinical routine (37, 38). Thus, our results encourage further clinical studies testing the impact of deep morphology learning-augmented *ex vivo* drug screening for personalized treatment selection in hemato-oncology.

METHODS

Clinical Cohort and Sample Preparation

In this study, we examined 390 samples from patients suffering from various hematologic malignancies, including, among others, 101 AML, 62 DLBCL, 50 chronic lymphocytic (CLL), and various T-cell lymphoma samples ($n = 69$; Fig. 2B; Supplementary Tables S3 and S4). A subset of samples were from 66 patients enrolled in the EXALT trial (4, 6), for which clinical outcome information with a median follow-up time of 23.9 months following pharmacoscopy testing was previously documented (Supplementary Table S2). All samples were collected and processed with written informed consent from the participants. For individual late-stage patients who met the EXALT trial inclusion criteria (4, 6), pharmacoscopy-guided therapy was provided as an individual healing attempt, conducted in accordance with the Declaration of Helsinki and the International Conference on Harmonization Guidelines for Good Clinical Practice. Ethical approval was granted by the Ethics Commission of the Medical University of Vienna (Ethik Kommission 1830/2015, 2008/2015, and 1895/2015).

Cancer cell–containing samples were acquired by either biopsy, bone marrow aspirate, or peripheral blood draws and were processed fresh or frozen (Supplementary Table S4). Tumor cell content was determined in clinical routine by fluorescence-activated cell sorting and/or by differential blood counts for liquid samples (blood and bone marrow) and by microscopic, IHC evaluation of slides for the respective tumor markers by a pathologist. Procured biopsies were either purified using Ficoll density gradient (bone marrow aspirates, peripheral blood, pleural effusion, ascites) or filtered through a 70-μm mesh filter (lymph tissue) into single cells resuspended in RPMI containing 10% FBS and 1% penicillin–streptomycin. Where possible, 20,000 cells per well were seeded in 384-well imaging plates. Depending on cell numbers, most samples (304 of 390) were distributed over two plates containing small compound libraries including 136 different drugs in two concentrations (1 μmol/L and 10 μmol/L) in, respectively, 2 and 3 technical replicate wells (Supplementary Tables S1 and S4). Compound annotations were retrieved from the database of Chemical Entities of Biological Interest (CHEBI; http://www.ebi.ac.uk/chebi/), as well as from the Kyoto Encyclopedia of Genes and Genomes Compound database (http://www.genome.jp/kegg/compound/) and DrugBank (https://go.drugbank.com/). Based on these resources, a "simplified drug class" annotation was generated manually. Samples were incubated in the drug plates overnight (18–24 hours, 37°C, 5% CO_2), subsequently fixed with 0.5% formaldehyde and 1:1,000 Triton X in phosphate-buffered saline, stained with the sample-specific set of antibodies and 4′,6-diamidino-2-phenylindole dihydrochloride (DAPI, Thermo Fisher Scientific) for later nuclear identification. Depending on the patient's diagnosis, different fluorescence-labeled primary antibodies were chosen for their ability to detect the malignant target cancer cell populations in each sample and cells were subsequently stained for 1 hour. Selected antibodies are listed in Supplementary Table S4 and include CD3 (HIT3a), CD19 (HIB19), CD20 (2H7), CD79a (HM47), CD34 (4H11), CD117 (104ED2), and CD138 (DL-101; eBioscience; Thermo Fisher). Next, all samples were imaged by automated microscopy (brightfield (650–760 nm), DAPI/Nuclear signal (435–480 nm), GFP/Green signal (500–550 nm), PE/Orange signal (570–630 nm), and APC/Red signal (650–760 nm) by Opera Phenix (PerkinElmer). Cell detection, fluorescence quantification (CellProfiler v2; Broad Institute of Harvard and the Massachusetts Institute of Technology, Boston, MA; ref. 39) and data processing (Matlab; version R2020a) were conducted as described previously (6, 13). In short, cell segmentation was performed by locating cell nuclei based on their DAPI signal followed by several rounds of nuclear expansion to detect the cell outline and background regions for local background correction. Mis-segmented cells, contaminants, or image artifacts were excluded based on thresholding on the DAPI intensity and the segmented area.

CNN Architecture Screen and CNN Training

CNNs have repeatedly demonstrated leading-edge performance in learning meaningful biological information for the automated analysis of microscopy images (40–42). With an improving computational infrastructure and newly emerging network building blocks, CNN architectures can be more easily adapted to specific classification problems in terms of network complexity, layer type, and parameter count. To determine the optimal architecture for the classification of our small single-cell image crops, we screened 291,800 randomly generated CNN architectures. For each sampled network, a fixed scaffold of image input and a three-layered classification block (fully connected, softmax, and class output; Fig. 1B) was complemented with a random collection of up to 25 layers drawn with equal probability from a set of nine layer types (including the sampling of layer-specific parameters): (i) convolutional layer (filter size: 3–10, stride: 1–3, padding: 0–2, number of filters: 5–100); (ii) fully connected layer (number of nodes: 4–100); (iii) rectified linear unit layer; (iv) leaky rectified unit layer; (v) batch normalization; (vi) cross-channel normalization (window size: 1–5); (vii) dropout layer; (viii) average pooling layer (window size: 1–5, stride: 1–3); (ix) max pooling layer (window size: 1–5, stride: 1–3). Before training, all networks were tested for their

basic functionality to discard networks with invalid layer sequences or incorrect parameterizations (like mismatching input/output sizes).

Functional networks were trained on 20 × 20 × 2 pixels (52 μm × 52 μm × 2 channels; DAPI and brightfield) subimage crops extracted from the DAPI and brightfield images, where each image contains a cell of interest at the image center. Cells overlapping with the original image border were excluded from training and classification. The cells' class labels (cancer or healthy) were automatically assigned by thresholding the fluorescence intensity of the staining antibody marking the malignant target population. The threshold optimally separating the marker-positive cancer cell population from the marker-negative healthy cells was determined by fitting a Gaussian mixture model with two components to the marker intensity distribution and identifying the intersection of both Gaussians (see Matlab's Statistics and Machine Learning Toolbox). The image resolution was constrained by a 10× magnification and subsequent binning of the pixel information to allow for faster imaging. Apart from image normalization with the in-built Matlab zero-center functionality, additional image preprocessing steps were conducted as described previously (4, 6). The training data ratio was set to 0.7/0.2/0.1 (training/validation/test cells). Network-layer weights and biases were initialized randomly before training. To avoid overfitting, L2 regularization with 0.005 was applied. Furthermore, images were randomly rotated in 45-degree steps and flipped vertically or horizontally in each iteration. CNNs were trained up to 30 epochs with a fixed learning rate of 0.0001. As a parameter optimization algorithm, the stochastic gradient descent implementation from Matlab was used with a momentum of 0.9.

Screened network candidates were ranked according to the highest training accuracy achieved on an image set from one of the patient samples (5,000 randomly sampled marker-positive and negative cell images per class; mini batch size: 250). Architecture, network size, training time, and training accuracy of the top-scoring CNN (mCNN) and an adapted state-of-the-art ResNet architecture are displayed in Fig. 1D. For the pan-sample versus per marker-sample pair training strategy comparison, we first trained up to three mCNNs per sample, one for each sample-specific target population (5,000 cells per sample and class; mini batch size: 250). Subsequently, mCNN and ResNet were trained on 586,500 cell images (750 randomly sampled images per sample and class; mini batch size: 2,000) pooled from all patient samples and across all drug conditions. For the pan-sample training image labels, we selected the marker signal that led to the highest per marker-sample pair training accuracy. To benchmark how the training strategy affects network performance, we collected a reference test set of 1,000 cell images per sample and class and compared the test accuracies achieved by the pan-sample and per marker-sample pair-trained mCNNs on this reference set (Supplementary Fig. S2A and S2B). CNN training and classification were done with Matlab (version 2020a).

Drug Score Calculation

For quantification of the *ex vivo* drug responses per sample and condition, we compared several drug score readouts that each rely on the classification of cancer and healthy cells. In total, we considered six measures: two CNN-based drug scores (DML by mCNN% and DML by ResNet%), where neural networks predicted the cell class labels, two immunofluorescence-based readouts (IF% and IF#), where a fluorescence threshold separated marker positive from negative cells, and two control measures based on the total cell number (total cell#), and a class vector, where the cell labels were assigned randomly (random%). The drug scores were reported either as absolute cell counts normalized to control (labeled with #), or by taking the relative cancer cell fraction (RCF) into account (labeled with %), i.e., the fraction of cells labeled as cancer in drug-treated wells divided by the average fraction of cancer-labeled cells measured in dimethyl sulfoxide (DMSO)-containing control wells (Fig. 2A). For the latter, we determined the AUC by calculating 1−the mean RCF value averaged over technical replicates and summed over both drug concentrations as described previously (4, 6, 13). Consequently, AUC drug scores above zero (corresponding with small RCF values) indicate on-target responses, values around zero refer to no significant drug effects and negative scores can be interpreted as chemoresistance. Statistical analyses were performed with Matlab (version R2020a).

Morphologic Profiling in CNN Latent Space

During training, CNNs learn an abstract interpretation of the input data, which is accompanied by an automated selection of those image features that desirably lead to an optimal classification of the image content. At the core, this learning process is an optimization problem, where the intrinsic parameters of each artificial neuron are systematically altered to improve class prediction. As CNN architectures are often hierarchically organized, i.e., neurons are arranged in layers, which have connections only to the preceding and the following layers, each layer constitutes a separate intermediate representation of the inputs and thus forms a so-called latent space. Consequently, latent spaces encode cellular morphology and vice versa; morphologic similarities appear closer also in latent space. Assuming that layers located deeper in the architecture contain a more concise representation of the learned features, we extracted the latent space either from the 23rd mCNN layer or from the 36th ResNet layer. Specifically, we computed the outputs of all neurons (so-called activations) that span the mCNN and ResNet latent space, respectively, via feed-forward propagation for 1,000 single-cell images per patient sample and class. All cell images were randomly selected from DMSO-control wells. For a better comparison of the mCNN and ResNet performance, we aligned the latent space dimensionality by conducting a principal component analysis and selected the first 20 components, respectively.

To explore the morphologic similarity between samples, we calculated the KNN of each cell in the extracted latent space. Each queried cell was assigned with the sample identity of the most similar cell from another sample (k = 1, Euclidean distance). Subsequently, we calculated the DMS score between two samples by computing the log ratio ($\log_2$ fold change) between the actual and the expected number of neighboring cells from those two samples.

Local enrichment of diagnosis-specific morphologies was computed by hypergeometric testing, i.e., by calculating the probability to randomly find at least n cells of diagnosis X in the defined neighborhood using a hypergeometric cumulative distribution function. This takes into account the total number of cells in the KNN-neighborhood (k = 100), the total number of cells sampled, and the total number of cells of diagnosis X. The enrichment probability is assigned to the corresponding cell in a t-distributed stochastic neighbor embedding (t-SNE) of the CNN latent space. In the t-SNE calculation (Barnes-Hut algorithm, Matlab 2020a), a standardized Euclidean distance metric, a perplexity of 1,000, and an exaggeration parameter of 50 were applied. All P values were corrected for multiple testing (Bonferroni correction), i.e., by the number of total cells (i.e., tests) in the analysis.

Possible Routes for Implementing DML-Based Pharmacoscopy into Clinical Routine

There are several different routes by which DML-based pharmacoscopy could be implemented in clinical routine. Sample testing by pharmacoscopy could be either done in a centralized lab or decentralized. Subsequently, computational analysis could be done either on local infrastructure or on decentralized "cloud" infrastructure. Analysis by DML operates out-of-the-box as described in the downloadable example code (https://www.snijderlab.org/deep_morphology_learning/). If samples were processed according to the provided pharmacoscopy protocol with similar imaging settings and modalities and patient diagnoses as the data mCNN was trained

on, the provided pretrained mCNN network can classify single-cell images from new samples into cancer or healthy without further network modifications. Those class labels, in turn, allow for drug-response quantification per condition, drug ranking, and subsequent treatment recommendation. However, the pretrained mCNN will not work out-of-the-box for stainings, imaging settings or modalities and diagnoses that were not included in its training data. In this case, it would be required to retrain mCNN via transfer learning on additional in-house samples. Possible control samples would be those for which IF identifies cancer from healthy cells with a 100% accuracy.

Data Availability

Due to privacy regulations, data storage was conducted on secured servers at the Medical University of Vienna, the Research Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), and the ETH Zurich. A pretrained instance of mCNN and example implementation is available at https://www.snijderlab.org/deep_morphology_learning/. All drug-response data and anonymized clinical annotations required to reproduce presented results are provided in the Supplementary tables accompanying this article.

Authors' Disclosures

T. Heinemann reports grants from Deutsche Forschungsgemeinschaft (DFG; German Research Foundation project number 389640585) and grants from the Swiss National Science Foundation (PP00P3_163961) during the conduct of the study. G.I. Vladimer reports other support from Exscientia GmbH and Allcyte GmbH during the conduct of the study; other support from Exscientia GmbH outside the submitted work; in addition, G.I. Vladimer has a patent for US2017356911A1 issued and licensed to Exscientia GmbH and a patent for US2021181183A1 pending and licensed to Exscientia GmbH. T. Pemovska reports grants from the European Molecular Biology Organization during the conduct of the study. W.R. Sperr reports grants and personal fees from Pfizer, personal fees from AbbVie, Jazz, BMS, and Stemline and personal fees and nonfinancial support from Novartis outside the submitted work. P. Valent reports grants from Allcyte during the conduct of the study; grants and personal fees from Celgene-BMS, AOP Orphan, personal fees from Novartis, Blueprint, Pfizer, Incyte, and Stemline outside the submitted work. Dr Jäger reports grants and personal fees from Novartis, personal fees from AbbVie, Roche, MSD, Gilead, Janssen, and BMS/Celgene during the conduct of the study; personal fees from Miltenyi, Amgen, Sanofi, and Incyte outside the submitted work. G. Superti-Furga reports nonfinancial support from Exscientia during the conduct of the study; nonfinancial support from Proxygen and Solgate outside the submitted work; in addition, G. Superti-Furga has a patent for pharmacoscopy issued to Exscientia. P.B. Staber reports personal fees from Amgen, grants and personal fees from Roche, Janssen, Gilead, Incyte, Morphosys, CTI, BMS, AbbVie, Takeda, and Beigene outside the submitted work. B. Snijder reports grants from the Swiss National Science Foundation and the European Research Council during the conduct of the study; grants from Roche, personal fees from Novartis, AbbVie, GSK, and other support from Exscientia outside the submitted work; in addition, B. Snijder has a patent for US Patent App. 15/514,045 issued. No disclosures were reported by the other authors.

Authors' Contributions

T. Heinemann: Data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **C. Kornauth:** Data curation, formal analysis, investigation. **Y. Severin:** Software, investigation. **G.I. Vladimer:** Data curation, formal analysis, investigation. **T. Pemovska:** Investigation. **E. Hadzijusufovic:** Investigation. **H. Agis:** Investigation. **M.-T. Krauth:** Investigation. **W.R. Sperr:** Investigation. **P. Valent:** Investigation. **U. Jäger:** Investigation. **I. Simonitsch-Klupp:** Investigation. **G. Superti-Furga:** Investigation. **P.B. Staber:** Resources, data curation, supervision, funding acquisition, investigation, writing–original draft. **B. Snijder:** Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing.

Acknowledgments

We thank the patients and their families for their trust in taking part in this study. The study was academically funded. T. Heinemann reports grants from Deutsche Forschungsgemeinschaft (DFG; German Research Foundation project number 389640585) and grants from Swiss National Science Foundation (PP00P3_163961) during the conduct of the study. B. Snijder acknowledges funding from the ETH Zurich, the Swiss National Science Foundation (PP00P3_163961, PP00P3_194809, and CRSII5_193832), and the European Research Council (SCIPER; 803063).

Note

Supplementary data for this article are available at Blood Cancer Discovery Online (https://bloodcancerdiscov.aacrjournals.org/).

Received November 24, 2021; revised June 8, 2022; accepted September 2, 2022; published first September 13, 2022.

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793–5.
2. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001;344:1031–7.
3. Perl AE, Martinelli G, Cortes JE, Neubauer A, Berman E, Paolini S, et al. Gilteritinib or chemotherapy for relapsed or refractory FLT3-mutated AML. N Engl J Med 2019;381:1728–40.
4. Kornauth C, Pemovska T, Vladimer GI, Bayer G, Bergmann M, Eder S, et al. Functional precision medicine provides clinical benefit in advanced aggressive hematological cancers and identifies exceptional responders. Cancer Discov 2022;12:372–87.
5. Pemovska T, Kontro M, Yadav B, Edgren H, Eldfors S, Szwajda A, et al. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. Cancer Discov 2013;3:1416–29.
6. Snijder B, Vladimer GI, Krall N, Miura K, Schmolke AS, Kornauth C, et al. Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. Lancet Haematol 2017;4:e595–606.
7. Dietrich S, Oleś M, Lu J, Sellner L, Anders S, Velten B, et al. Drug-perturbation-based stratification of blood cancer. J Clin Invest 2018;128:427–45.
8. Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer 2015;15:747–56.
9. Frismantas V, Dobay MP, Rinaldi A, Tchinda J, Dunn SH, Kunz J, et al. Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. Blood 2017;129:e26–37.
10. Letai A. Functional precision medicine: Putting drugs on patient cancer cells and seeing what happens. Cancer Discov 2022;12:290–2.
11. Malani D, Kumar A, Brück O, Kontro M, Yadav B, Hellesøy M, et al. Implementing a functional precision medicine tumor board for acute myeloid leukemia. Cancer Discov 2022;12:388–401.
12. Irmisch A, Bonilla X, Chevrier S, Lehmann K-V, Singer F, Toussaint NC, et al. The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support. Cancer Cell 2021;39:288–93.

13. Vladimer GI, Snijder B, Krall N, Bigenzahn JW, Huber KVM, Lardeau CH, et al. Global survey of the immunomodulatory potential of common drugs. Nat Chem Biol 2017;13:681–90.
14. Severin Y, Hale BD, Mena J, Goslings D, Frey BM, Snijder B. Multiplexed high-throughput immune cell imaging reveals molecular health-associated phenotypes. bioRxiv. 2021:2021.12.03.471105. Available from: https://www.biorxiv.org/content/10.1101/2021.12.03.471105.full.
15. Shilts J, Severin Y, Galaway F, Müller-Sienerth N, Chong Z-S, Pritchard S, et al. A physical wiring diagram for the human immune system. Nature 2022;608:397–404.
16. Bourquin J-P. A precision medicine approach to haematological malignancies. Lancet Haematol 2017;4:e567–8.
17. Wheeler DA, Takebe N, Hinoue T, Hoadley KA, Cardenas MF, Hamilton AM, et al. Molecular features of cancers exhibiting exceptional responses to treatment. Cancer Cell 2021;39:38–53.
18. Bibbo M, Wilbur D. Comprehensive cytopathology e-book. Amsterdam: Elsevier Health Sciences; 2014.
19. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
20. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
21. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.
22. Buggenthin F, Buettner F, Hoppe PS, Endele M, Kroiss M, Strasser M, et al. Prospective identification of hematopoietic lineage choice by deep learning. Nat Methods 2017;14:403–6.
23. Brück OE, Lallukka-Brück SE, Hohtari HR, Ianevski A, Ebeling FT, Kovanen PE, et al. Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with MDS. Blood Cancer Discov 2021;2:238–49.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. page 770–8.
25. Schmidl C, Vladimer GI, Rendeiro AF, Schnabl S, Krausgruber T, Taubert C, et al. Combined chemosensitivity and chromatin profiling prioritizes drug combinations in CLL. Nat Chem Biol 2019;15:232–40.
26. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review 2020;53:5455–516.
27. Kurilov R, Haibe-Kains B, Brors B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. Sci Rep 2020;10:2849.
28. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 2016;127:2375–90.
29. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 2016;127:2391–405.
30. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 2016;374:2209–21.
31. Lenz G, Wright GW, Emre NCT, Kohlhammer H, Dave SS, Davis RE, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proc Natl Acad Sci U S A 2008;105:13520–5.
32. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018;562:526–31.
33. Lee J-K, Liu Z, Sa JK, Shin S, Wang J, Bordyuh M, et al. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. Nat Genet 2018;50:1399–411.
34. van de Wetering M, Francies HE, Francis JM, Bounova G, Iorio F, Pronk A, et al. Prospective derivation of a living organoid biobank of colorectal cancer patients. Cell 2015;161:933–45.
35. Valent P, Orfao A, Kubicek S, Staber P, Haferlach T, Deininger M, et al. Precision medicine in hematology 2021: definitions, tools, perspectives, and open questions. Hemasphere 2021;5:e536.
36. Letai A. Functional precision cancer medicine—moving beyond pure genomics. Nat Med 2017;23:1028.
37. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.
38. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3:199–217.
39. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7:R100.
40. Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Valen D. Deep learning for cellular image analysis. Nat Methods 2019;16:1233–46.
41. Alom MZ, Yakopcic C, Taha TM, Asari VK. Microscopic blood cell classification using inception recurrent residual convolutional neural networks. NAECON 2018—IEEE National Aerospace and Electronics Conference. 2018:222–7.
42. Shu X, Sansare S, Jin D, Tong K-Y, Pandey R, Zhou R. White blood cell classification using quantitative phase microscopy based deep learning. Biophotonics Congress: Optics in the Life Sciences Congress 2019 (BODA, BRAIN, NTM, OMA, OMP). Optical Society of America; 2019:DT3B.3.