



American Society of Hematology  
2021 L Street NW, Suite 900,  
Washington, DC 20036  
Phone: 202-776-0544 | Fax 202-776-0545  
editorial@hematology.org

## **Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset**

Tracking no: BLD-2020-010568R1

Christian Matek (Department of Internal Medicine III, University Hospital Munich, Ludwig-Maximilians-Universität, München - Campus Großhadern, Munich, Germany, Germany) Sebastian Krappe (Department of Computer Science, University of Koblenz-Landau, Koblenz, Germany, Germany) Christian Münzenmayer (Fraunhofer Institute for Integrated Circuits IIS, Germany) Torsten Haferlach (MLL Munich Leukemia Laboratory, Germany) Carsten Marr (Institute of Computational Biology Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Germany)

### **Abstract:**

Biomedical applications of deep learning algorithms rely on large, expert annotated data sets. The classification of bone marrow cell cytomorphology, an important cornerstone of hematological diagnosis, is still done manually thousands of times every day, due to a lack of datasets and trained models.

We apply convolutional neural networks (CNNs) to a large dataset of 171,374 microscopic cytological images taken from bone marrow smears of 945 patients diagnosed with a variety of hematological diseases. The dataset is the largest expert-annotated pool of bone marrow cytology images available in the literature so far. It allows us to train high-quality classifiers of leukocyte cytomorphology that identify a wide range of diagnostically relevant cell species at high precision and recall. Our CNNs outcompete previous feature based approaches and provide a proof-of-concept to the classification problem of single bone marrow cells.

Our work is a step towards automated evaluation of bone marrow cell morphology using state-of-the-art image classification algorithms. The underlying dataset represents both an educational resource as well as a reference for future AI-based approaches to bone marrow cytomorphology.

**Conflict of interest:** No COI declared

**COI notes:**

**Preprint server:** No;

**Author contributions and disclosures:** Christian Matek, Carsten Marr and Torsten Haferlach conceived of the study. Sebastian Krappe and Christian Münzenmayer digitised samples. Christian Matek trained and evaluated network algorithms and analysed results. All authors contributed to interpreting the results and writing the paper.

**Non-author contributions and disclosures:** No;

**Agreement to Share Publication-Related Data and Data Sharing Statement:** Public deposit.

**Clinical trial registration information (if any):**

# Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset

Christian Matek<sup>1,2</sup>, Sebastian Krappe<sup>3,4</sup>, Christian Münzenmayer<sup>3</sup>, Torsten Haferlach<sup>5</sup>, Carsten Marr<sup>1</sup>

1 Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

2 Department of Internal Medicine III, University Hospital Munich, Ludwig-Maximilians-Universität, München - Campus Großhadern, Munich, Germany

3 Image Processing and Medical Engineering Department, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

4 Department of Computer Science, University of Koblenz-Landau, Koblenz, Germany

5 MLL Munich Leukemia Laboratory, Munich, Germany

## Key Points

- A dataset of over 170,000 microscopic images allows training neural networks for identification of bone marrow cells with high accuracy
- Neural networks outperform a feature-based approach to bone marrow cell classification and can be analysed with explainability and feature embedding methods

## Abstract

Biomedical applications of deep learning algorithms rely on large, expert annotated data sets. The classification of bone marrow cell cytomorphology, an important cornerstone of hematological diagnosis, is still done manually thousands of times every day, due to a lack of datasets and trained models.

We apply convolutional neural networks (CNNs) to a large dataset of 171,374 microscopic cytological images taken from bone marrow smears of 945 patients diagnosed with a variety of hematological diseases. The dataset is the largest expert-annotated pool of bone marrow cytology images available in the literature so far. It allows us to train high-quality classifiers of leukocyte cytomorphology that identify a wide range of diagnostically relevant cell species at high precision and recall. Our CNNs outcompete previous feature-based approaches and provide a proof-of-concept to the classification problem of single bone marrow cells.

Our work is a step towards automated evaluation of bone marrow cell morphology using state-of-the-art image classification algorithms. The underlying dataset represents both an educational resource as well as a reference for future AI-based approaches to bone marrow cytomorphology.

## Introduction

Examination and differentiation of bone marrow cell morphologies is an important cornerstone in the diagnosis of malignant and non-malignant diseases affecting the hematopoietic system <sup>1–5</sup>. While a large number of sophisticated methods, including cytogenetics, immunophenotyping and increasingly molecular genetics, are now available, cytomorphological examination remains an important first step in the diagnostic workup of many intra- and extramedullary pathologies. Having been established already in the 19th century <sup>6</sup>, the role of bone marrow cytology is still central for its relatively quick and widespread technical availability <sup>7</sup>. The method has so far proven difficult to automatize, which is why in a clinical routine workflow, microscopic examination and classification of single-cell morphology is still primarily performed by human experts. However, manual evaluation of bone marrow smears can be tedious and time-consuming and highly depend on examiner skill and experience, especially in unclear cases <sup>8</sup>. Hence, the number of high-quality cytological examinations is limited by the availability and experience of trained experts, while examiner classifications have been found to be subject to substantial inter- and intra-rater variability <sup>9–11</sup>. Furthermore, examination of individual cell morphologies is inherently qualitative, which makes the method difficult to combine with other diagnostic methods that offer more quantitative data.

Few attempts to automatize cytomorphological classification of bone marrow cells have been undertaken. Most of them are based on extracting hand-crafted single cell features from digitized images and using those to classify the cell in question <sup>12,13</sup>. Additionally, the majority of previous work in automated cytomorphological classification has been focused either on the classification of physiological cell types or the classification of peripheral blood smears <sup>14–16</sup>, hence limiting its usability for classification of leukocytes in the bone marrow for the diagnosis of hematological malignancies. Deep-learning approaches to bone marrow cell classification have so far either focussed on relatively low numbers of samples or disease classes, or not made the corresponding data available publicly <sup>17–20</sup>.

Classification of natural images has undergone significant improvements in accuracy over the past few years, aided by the increasingly widespread use of convolutional neural networks (CNNs) <sup>21,22</sup>. In the meantime, this technology has also been applied to a variety of medical image interpretation tasks, including mitosis detection in histological sections of breast cancer <sup>23</sup>, skin cancer detection <sup>24</sup>, mammogram evaluation <sup>25</sup> and cytological classification in peripheral blood <sup>11</sup>. However, successful use of CNNs for image classification typically relies on the availability of sufficiently many, high-quality image data and high-quality annotation, which can both be difficult to access due to the expense involved in obtaining labels by medical experts <sup>26,27</sup>. This is true in particular in situations like the cytomorphological examination of bone marrow, where there is no underlying technical gold standard, and human examiners are needed to provide the ground truth labels for network training and evaluation.

Here, we present a large dataset of 171,374 expert-annotated single cell images from 945 patients diagnosed with a variety of hematological diseases. To the authors' knowledge, it is the largest image dataset of bone marrow cytomorphology available in the literature so far in terms of the number of diagnoses, patients and cell images included. It is therefore a resource to be used both for educational purposes and future approaches to automated image-based bone marrow cell classification. We use the dataset to train two CNN-based classifiers for single-cell images of bone-marrow leukocytes, one using ResNeXt, a recent model that proved successful in natural image classification, as well as a simpler sequential network architecture. We test and compare the classifiers and find that they outperform previous methods while achieving excellent levels of accuracy for many cytomorphologic cell classes with direct clinical relevance. The fact that both classifiers based on different models attain good results increases confidence in the robustness of our findings.

## Methods

### Dataset selection and digitization

Bone-marrow cytologic preparations were included from 961 patients diagnosed with a variety of hematological diseases at MLL between 2011 and 2013. All patients had given written informed consent to the use of clinical data according to the Declaration of Helsinki. Here images of single cells do not allow any patient specific tracking. The study was approved by the MLL internal institutional review board. Age range of included patients was 18.1 - 92.2 years, with a median of 69.3 years and a mean of 65.6 years. The cohort included 575 (59.8%) male and 385 (40.1%) female patients, as well as 1 (0.1%) patient of unknown gender.

All bone marrow smears were stained according to standard protocols as used in daily routine. May-Grünwald-Giemsa/Pappenheim staining was used as published elsewhere <sup>2</sup>.

Bone marrow smears are digitized with an automated microscope (Zeiss Axio Imager Z2) in several steps. At first the whole slide is captured in low optical magnification (1-fold magnification) to obtain an overview image. For an automatic system and to minimize the scanning duration, automated detection of the bone marrow smear on the microscopic slide is necessary. The contour of the smear is identified by a combination of thresholding and k-means clustering methods <sup>28</sup>. Then the bone marrow smear region is determined and digitized by meandering the slide with a mid-level (5-fold) magnification objective. Relevant regions are selected by human experts and scanned automatically in high magnification (40x oil immersion) for the morphological cell analysis. All images used for this work have been captured with a CCD-camera mounted on a brightfield microscope (Zeiss Axio Imager Z2). The dimensions of the original images are 2452 × 2056 pixel and the physical size of a camera pixel is 3.45 × 3.45 μm. For the localization of single cells, a method considering the foreground ratio of the high resolution bone marrow images is used <sup>29</sup>. A quadratic region around each found cell center is presented to experienced

cytologists of the Munich Leukemia Laboratory (MLL) to determine the ground truth classifications for single-cell images. A total of 945 patients was included in the final analysis dataset. Diagnoses represented in the cohort include a variety of myeloid and lymphoblastic malignancies, lymphomas as well as non-malignant and reactive alterations, reflecting the sample entry of a large laboratory specialised in hematology (see Fig. S1 of the Supplementary Material).

From the examined regions, diagnostically relevant cell images were annotated into 22 different classes according to the morphological scheme shown in Fig. 1A. When annotating individual samples, morphologists were asked to annotate 200 cells per slide in accordance with routine practice. In order to avoid biasing the annotation for easily classifiable cell images, separate classes for artifacts, cells that could not be identified, and other cells belonging to morphological classes not represented in the scheme, were included. From the annotated regions, images of a size of 250x250 pixels were extracted containing the respective annotated cell as a main content in the patch center (see Fig. 1A). No further cropping, filtering or segmentation between foreground and background took place, leaving the algorithm with the task to identify the main image content relevant for the respective annotation. To exclude correlations between different images in the dataset, we screened for overlaps between images using the SIFT algorithm<sup>30</sup>, and discarded images for which overlapping local features were detected. The final number of images contained in each of the 22 morphological classes is shown in Table 1. Overall, the cleaned dataset comprises 171,374 single-cell images.

Cell types represented in our morphological classification scheme appear in very different frequencies on bone marrow smears, resulting in a highly imbalanced distribution of training images (Fig. 1B). Class imbalance is a challenging feature of many medical datasets<sup>31</sup>, and in the present case arises from both the uneven prevalence of different disease entities included, and the different intrinsic prevalence of specific cell classes in a given sample. In order to counteract the class imbalance in the training process, we used dataset augmentation<sup>21</sup>, and upsampled the training data to approximately 25,000 images per class by performing a set of augmentation transformations. Firstly, we used clockwise rotations by a random continuous angle in the range of 0°-180°, as well as vertical and horizontal flips, shifts by up to 25% of the image weight and height, and shears by up to 20% of the image size. In addition to these geometric transformations, we also included stain-color augmentation transformations, which has been shown to improve robustness and generalisability of the resulting classifier<sup>32</sup>. Following the strategy proposed in ref.<sup>33</sup>, we first separated the eosin-like and the hematoxylin-like component according to the PCA-based method by Macenko *et al.*<sup>34</sup>. These two stain components were then perturbed using the method and default parameters of ref.<sup>33</sup>, thus simulating the variability of stain intensity.

## Network structure and training

We used the ResNeXt-50 architecture developed by Xie *et al.* <sup>35</sup>, a successful image classification network obtaining a second rank in image classification of the ImageNet ILSVRC 2016 competition <sup>36</sup>. The network topology was employed before in the classification of peripheral blood smears <sup>14</sup>, making it a natural choice for the morphologic classification of bone marrow cells. One advantage of the ResNeXt architecture is its low number of hyper-parameters. We kept the cardinality hyper-parameter C at the value of 32, as was done in the original work <sup>35</sup>. We furthermore modify the network input to accept images of the size 250 x 250 pixels, and adjusted the number of output nodes to 22, corresponding to the number of morphological classes defined in our annotation scheme. Overall, the resulting network possessed 23,059,094 trainable parameters. When generating class predictions on images, the output node with the highest activation determined the cell class prediction.

Networks were trained on Nvidia Tesla V100 graphics processing units, where training of the ResNeXt model took approximately 48 hours of computing time. For the training of individual networks reported in this paper, we used 80% of the available images of each class, whereas 20% were used for evaluation of the trained network. This stratified train-test split was performed in a random fashion. Data augmentation was performed after the test-train-split. For 5-fold cross-validation, we performed a stratified split of the dataset into 5 mutually disjoint folds, each containing approximately 20% of images in the respective cell class. We then trained 5 different networks for 13 epochs, where each network used a different fold for testing and the remaining 4 folds for training of the network. Results were then averaged across the 5 different networks. To evaluate the robustness of our results with respect to network structure, we also trained a sequential model with a simpler architecture that has been used before to train a classifier for leukocytes in peripheral blood <sup>11</sup>. The precise network architecture used is shown in Fig. S5 of the Supplementary Material. With input and output channels adjusted to match those of the ResNeXt model, it contained a total of 303,694 trainable parameters. Distribution of data into test and training sets for the different folds was kept identical to the one used in training the ResNeXt model. The feature-based approach of Krappe *et al.* <sup>13</sup> used minimum redundancy selection <sup>37</sup> of over 6000 features per cell to train a Support Vector Machine (SVM). Furthermore, a slightly different training strategy was employed. While the split into test and training data was kept identical for training the ResNeXt and the sequential CNN models, the feature-based approach used 70% of the data for training and 30% for evaluation <sup>13</sup>. To ensure that our results are robust with respect to this slight difference in split strategy, we trained a ResNeXt-50 model using a stratified split of the data into 70% training and 30% test data. The results show only minor deviations from the 5-fold cross-validation results (cf. Fig. S6 of the Supplementary Material).



## Results

The trained deep neural ResNeXt shows accurate prediction performance for most morphological classes in our scheme (Fig. 2). As might be expected for a data-driven learning algorithm such as a neural network, classification performance tends to increase with a higher number of available training sample images. For quantitative evaluation of our training algorithm, we employ the common measures of precision and recall, defined as

$$Precision = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad Recall = \frac{\text{true positive}}{\text{true positive} + \text{false negative}},$$

where „true positive“ and „true negative“ are defined as the number of images that are classified or not classified into a given class in agreement with the ground truth. Similarly, „false positive“ and „false negative“ signify the number of images that are classified or not classified into a given class in disagreement with the ground truth.

As has been noted before, precise differentiation of individual morphological classes can be difficult, in particular when they are closely related in the leukocyte differentiation lineage<sup>13</sup>. As a result of this intrinsic uncertainty of morphological classification, some predictions of the network can be considered tolerable even though they differ from the ground-truth label provided by the human annotator. As an example, a confusion between segmented and band neutrophils, which are consecutive morphological classes in the continuous process of myelopoiesis, can be considered tolerable. This consideration led Krappe *et al.*<sup>13</sup> to introduce so-called tolerance classes for the evaluation of their feature-based classifier on a related single-cell dataset. In this work, similar tolerance classes are employed and shown in detail in Fig. 3A.

The values of precision and recall attained by the ResNeXt network for individual morphological classes are given in Table 1 as mean  $\pm$  standard deviation across the five networks evaluated on the mutually disjoint folds (see Methods). Both a strict and a tolerant evaluation strategy are used, where the former compares the network prediction to the ground truth label only and the latter takes tolerance classes into account. The analogous analysis was also performed for the sequential model trained for comparison on the identical data. A confusion matrix analogous to Fig. 2 as well as class-wise precision and recall values are given in Fig. S3 and Tab. S1 of the Supplementary Material. Overall, the sequential network attains similar but somewhat inferior performance values, in agreement with the comparative evaluation of both network architectures in the classification of peripheral blood cells from a dataset of 15 malignant and non-malignant cell classes relevant in AML<sup>11</sup>.

In a direct performance comparison of ResNeXt, the sequential model, and the feature based approach of ref.<sup>13</sup> the neural networks outperform the feature-based classifier in all classes (Fig. 3B) but segmented neutrophils and lymphocytes, where the average tolerance recall of our 5-fold cross-validation lies slightly below the tolerance recall of the

feature-based method. This effect may be due to the distinctive signal of segmented neutrophils and lymphocytes in the feature space used in the classifier of ref.<sup>13</sup> which explicitly includes parameters of nuclear shape. In contrast, the neural network classifiers used in this work do not rely on the extraction of handcrafted morphological parameters, but extract the relevant features from the training dataset.

As is expected for a data-driven method, the classifier performs less favourably on classes for which only few training images are available, such as faggott cells or pathological eosinophils. For image classification tasks focussed on recognition of these specific cell types, more training data would be required. Furthermore, training a binary instead of a full multi-class classifier might yield better prediction performance<sup>38</sup>.

## External validation

To test the generalisability of our model, evaluation of the network's predictions on an external dataset not used during training is required. At present, very few publicly available datasets that include single bone-marrow cells in sufficient number, imaging and annotation quality exist, rendering evaluation of generalizability of our network's predictions challenging. We evaluated our model on an annotated dataset by Choi *et al.*<sup>20</sup>, for which 627 single-cell images from 30 slides of 10 patients are available with annotations for different stages of the erythroid and myeloid lineage. The dataset includes images with different illuminations and image resolutions. As no information on the physical pixel size was available, we scaled all single-cell images up to the size of 250x250 pixels, and generated predictions from the scaled image. Note that this may lead to input images of systematically different sizes compared to the images of our own dataset.

Of the 627 annotated images, 247 were classified into the "Artifact" or "Not identifiable" categories, indicating that the network is not able to predict the class of these images. Predictions on the remaining 380 images are shown in Fig. 5, indicating fair performance of the classifier on the external dataset. In particular, most cells are classified correctly into their respective lineage. Given the different imaging and annotation strategies followed in the compilation of both datasets, a considerable amount of tolerable confusion between individual lineage steps is expected.

It has to be noted that when compared to the internal dataset, the external evaluation dataset is relatively small and heterogeneous in terms of staining and background lighting. Furthermore, considerable differences in terms of imaging and annotation strategies exist. For example, the lymphoid lineage is not covered, and the annotation classes differ from those used in our dataset. Nevertheless, the performance of the classifier on the external dataset indicates that the model is able to generalize, and recognise cases in which no confident prediction can be made. It might be expected that including further information on the external dataset, e.g. matching the patch size or background brightness to the one



used during training, or matching the stain color distribution, would further increase the performance of our classifier.

On the sequential model, a qualitatively similar performance is observed on the external dataset (cf. Fig. S4 of the Supplementary Material), suggesting that generalisation is robust against different network architectures.

### **Classification analysis and explainability**

As they are developed based on the training set in a data-driven way, the classification decisions of neural networks do not lend themselves to direct human interpretation. In order to nevertheless gain insight into the classification decisions taken by these algorithms, a variety of explainability methods has recently been developed <sup>39</sup>. To determine which regions of the input images are important for the network's classification decisions, we analysed the ResNeXt model using two different methods. Specifically, we used SmoothGrad <sup>40</sup> and GradCAM <sup>41</sup>, two recently developed explainability algorithms which have been shown to fulfill basic requirements for explainability methods, namely sensitivity to data and model parameter randomization <sup>42</sup>. Results for key cell classes are shown in Fig. 4, suggesting that the model has learned to focus on the relevant input of a single-cell patch, namely the main leukocyte shown in it, while ignoring background features like erythrocytes, cell debris or parts of other cells visible in the patch. Furthermore, specific defining structures known to be relevant to human examiners when classifying cells appear to also play a role in the network's attention, such as the cytoplasm of eosinophils and the cell membrane of hairy cells. While as post-hoc classification explanations, these analyses do not in themselves guarantee the correctness of a particular classification decision, they may increase confidence that the network has learned to focus on relevant features of the single-cell images, and predictions are based on reasonable features.

As a second test that the classifier has learned relevant and consistent information, we embedded the extracted features represented in the flattened final convolutional layer of the network with 2048 dimensions into 2 dimensions for each member of the test set using the UMAP algorithm<sup>43</sup>. The result of the embedding is shown in Fig. 6, suggesting that the classifier has extracted features that generally separate individual classes well, although some classes such as Monocytes can be challenging to distinguish from unrelated others, as can be seen by their proximity to unrelated classes in the embedding. Additionally, the embedding shows that classes representing consecutive steps of cell development, e.g. proerythroblasts, erythroblasts and normoblasts, are mapped to neighboring parts of feature space. This indicates that the network extracts relevant features indicative of the continuous nature of development between these classes.

## **Discussion**

Neural networks have been shown to be successful in a variety of image classification problems in the past. In this work, we present a large, annotated, high-quality dataset of

microscopic images taken from bone-marrow smears of a large patient cohort, which can be used as a reference for developing machine learning approaches to morphological classification of diagnostically relevant leukocytes. To our knowledge, this image database is the most extensive one available in the literature so far, both in terms of patients, diagnoses and single-cell images included.

We used the dataset to train and test a state-of-the-art CNN for morphological classification. Overall, results are encouraging, with high precision and recall values obtained for most diagnostically relevant classes. In direct comparison of recall values, our network clearly outperforms a feature-based classifier<sup>13,44</sup> recently developed on the same dataset for most morphological cell classes. Our findings are in line with experiences from other areas of medical imaging, where deep-learning-based image classification tasks have achieved higher benchmarks than methods that require extraction of handcrafted features<sup>27,45</sup>. The key ingredient to a successful application of CNNs is a large enough and high-quality training dataset<sup>21</sup>.

While CNNs have outperformed classifiers relying on handcrafted features across a wide range of tasks, the structure of their output usually does not lend itself to straightforward human interpretation. To address this drawback, a variety of explainability methods have been developed. In this work, we used the SmoothGrad and the Grad-CAM algorithms, and found that the algorithm has indeed learned to focus on relevant regions of the single-cell image, as well as to pay attention to features known to be characteristic of specific cell classes. By analysing the features extracted by the network using the UMAP embedding, we could furthermore confirm that the network has learned to stably separate morphological classes and map cells with morphological similarities into neighboring regions of feature space. The features learned by the network to classify single-cell images therefore appear robust and tolerant with respect to some label noise which cannot be avoided in a data-driven method relying on expert annotations.

In the present study, we primarily followed a single-center approach, with all bone-marrow smears included for training prepared in the same laboratory, and digitized using the same scanning equipment. Within that setting, the network described in the work shows very encouraging performance. External validation, though challenging due to the limited amount of available data, indicates that the method is generalizable to data obtained in other settings. Applicability to other laboratories and scanners may be further increased using larger and more diverse datasets, and including specific information on imaging and data handling into the image analysis pipeline<sup>21,46</sup>. Further expansion of the morphological database, ideally in a multi-centric study and including a range of scanner hardware, would likely further increase the performance and robustness of the network, in particular for classes containing few samples in our dataset. However, due to the number of cases and diagnoses included, we expect our dataset to reasonably reflect the morphological variety for most cell classes. In order to evaluate performance of our network in a real-world diagnostic setting, further work is needed. Given the variety of diagnostic modalities used in hematology, we anticipate that inclusion of complementary data, e. g. from flow

cytometry or molecular genetics would further increase the quality of predictions that can be obtained by neural networks.

## **Data availability**

Data is available for review via a process to be defined by the Blood editorial board, and will be published upon acceptance via The Cancer Imaging Archive (TCIA) under (insert DOI here).

## **Acknowledgements**

We thank Matthias Hehr for feedback on our manuscript. Christian Matek and Carsten Marr acknowledge support from the German National Research foundation (DFG) through grant SFB 1243. Carsten Marr has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 866411).

## **Authorship Contributions**

Christian Matek, Carsten Marr and Torsten Haferlach conceived of the study. Sebastian Krappe and Christian Münzenmayer digitised samples. Christian Matek trained and evaluated network algorithms and analysed results. All authors contributed to interpreting the results and writing the paper.

## **Disclosure of Conflicts of Interest**

The authors state they have no conflicts of interest.

## References

1. Theml HK, Diem H, Haferlach T. Color Atlas of Hematology. 2004;
2. Löffler H, Rastetter J, Haferlach T (eds ). Atlas of Clinical Hematology, 6th Ed. Springer-Verlag Berlin Heidelberg; 2005.
3. Haferlach T. Hämatologische Erkrankungen: Atlas und diagnostisches Handbuch. Springer-Verlag; 2020.
4. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375–2390.
5. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424–447.
6. Thomas X. First contributors in the history of leukemia. *World Journal of Hematology*. 2013;2(3):62.
7. Tkachuk DC, Hirschmann JV, Wintrobe MM. Wintrobe's Atlas of Clinical Hematology. Lippincott Williams & Wilkins; 2007.
8. Briggs C, Longair I, Slavik M, et al. Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system. *International Journal of Laboratory Hematology*. 2009;31(1):48–60.
9. Fuentes-Arderiu X, Dot-Bach D. Measurement uncertainty in manual differential leukocyte counting. *Clin. Chem. Lab. Med.* 2009;47(1):112–115.
10. Font P, Loscertales J, Soto C, et al. Interobserver variance in myelodysplastic syndromes with less than 5 % bone marrow blasts: unilineage vs. multilineage dysplasia and reproducibility of the threshold of 2 % blasts. *Ann. Hematol.* 2015;94(4):565–573.
11. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*. 2019;1(11):538–544.
12. Krappe S, Benz M, Wittenberg T, Haferlach T, Münzenmayer C. Automated classification of bone marrow cells in microscopic images for diagnosis of leukemia: a comparison of two classification schemes with respect to the segmentation quality. *Medical Imaging 2015: Computer-Aided Diagnosis*. 2015;
13. Krappe S, Wittenberg T, Haferlach T, Münzenmayer C. Automated morphological analysis of bone marrow cells in microscopic images for diagnosis of leukemia: nucleus-plasma separation and cell classification using a hierarchical tree model of hematopoiesis. *Medical Imaging 2016: Computer-Aided Diagnosis*. 2016;
14. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks.
15. Scotti F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. *CIMSA. 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2005*. .
16. Kimura K, Tabe Y, Ai T, et al. A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Scientific Reports*. 2019;9(1.):
17. Mori J, Kaji S, Kawai H, et al. Assessment of dysplasia in bone marrow smear with convolutional neural network. *Sci. Rep.* 2020;10(1):14734.
18. Anilkumar KK, Manoj VJ, Sagi TM. A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of Leukemia. *Biocybernetics and Biomedical Engineering*. 2020;40(4):1406–1420.
19. Jin H, Fu X, Cao X, et al. Developing and Preliminary Validating an Automatic Cell

- Classification System for Bone Marrow Smears: a Pilot Study. *J. Med. Syst.* 2020;44(10):184.
20. Choi JW, Ku Y, Yoo BW, et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. *PLoS One*. 2017;12(12):e0189259.
  21. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
  22. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*. 2017;29(9):2352–2449.
  23. Albarqouni S, Baur C, Achilles F, et al. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Trans. Med. Imaging*. 2016;35(5):1313–1321.
  24. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.
  25. McKinney SM, Sieniek M, Godbole V, et al. Addendum: International evaluation of an AI system for breast cancer screening. *Nature*. 2020;586(7829):E19.
  26. Greenspan H, van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*. 2016;35(5):1153–1159.
  27. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*. 2017;19(1):221–248.
  28. Krappe S., Evert, A., Haferlach, T. Held, C., Wittenberg, T., Münzenmayer, C. Smear detection for the automated image-based morphological analysis of bone marrow samples for leukemia diagnosis. *Forum Life Science*. 2015;
  29. Krappe S, Maciejewski K, Eismann E, et al. Lokalisierung von Knochenmarkzellen für die automatisierte morphologische Analyse von Knochenmarkpräparaten. *Informatik aktuell*. 2014;403–408.
  30. Lowe DG. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*. 1999;
  31. Rahman MM, Mostafizur Rahman M, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*. 2013;224–228.
  32. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* 2019;58:101544.
  33. Tellez D, Balkenhol M, Otte-Holler I, et al. Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. *IEEE Transactions on Medical Imaging*. 2018;37(9):2126–2136.
  34. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009;
  35. Xie S, Girshick R, Dollar P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017;
  36. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. 2015;115(3):211–252.
  37. Schowe B, Morik K. Fast-Ensembles of Minimum Redundancy Feature Selection. *Ensembles in Machine Learning Applications*. 2011;75–95.
  38. Schouten JPE, Matek C, Jacobs LFP, et al. Tens of images can suffice to train neural networks for malignant leukocyte detection. *Sci. Rep.* 2021;11(1):7995.
  39. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. *Lecture Notes in Computer Science*. 2019;
  40. Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825*. 2017;
  41. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks

via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017;

42. Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* 2018;31:9505–9515.
43. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 2018;3(29):861.
44. Krappe S. Automatische Klassifikation von hämatopoetischen Zellen für ein computer-assistiertes Mikroskopiesystem. 2018.
45. Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Z. Med. Phys.* 2019;29(2):86–101.
46. Kothari S, Phan JH, Stokes TH, et al. Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J Biomed Health Inform.* 2014;18(3):765–772.

## Tables

Class	Precision <sub>tolerant</sub>	Recall <sub>tolerant</sub>	Precision <sub>strict</sub>	Recall <sub>strict</sub>	Images
Band neutrophils	0.91 ± 0.02	0.91 ± 0.01	0.54 ± 0.03	0.65 ± 0.04	9968
Segmented neutrophils	0.95 ± 0.01	0.85 ± 0.03	0.92 ± 0.02	0.71 ± 0.05	29424
Lymphocytes	0.90 ± 0.03	0.72 ± 0.03	0.89 ± 0.03	0.70 ± 0.03	26242
Monocytes	0.57 ± 0.05	0.70 ± 0.03	0.57 ± 0.05	0.70 ± 0.03	4040
Eosinophils	0.85 ± 0.05	0.91 ± 0.03	0.85 ± 0.05	0.91 ± 0.03	5883
Basophils	0.14 ± 0.04	0.64 ± 0.07	0.14 ± 0.04	0.64 ± 0.07	441
Metamyelocytes	0.68 ± 0.04	0.87 ± 0.03	0.30 ± 0.05	0.64 ± 0.08	3055
Myelocytes	0.78 ± 0.03	0.91 ± 0.01	0.52 ± 0.05	0.59 ± 0.06	6557
Promyelocytes	0.91 ± 0.02	0.89 ± 0.03	0.76 ± 0.05	0.72 ± 0.08	11994
Blasts	0.79 ± 0.03	0.69 ± 0.03	0.75 ± 0.03	0.65 ± 0.03	11973
Plasma cells	0.81 ± 0.06	0.84 ± 0.04	0.81 ± 0.06	0.84 ± 0.04	7629
Smudge cells	0.28 ± 0.09	0.90 ± 0.10	0.28 ± 0.09	0.90 ± 0.10	42
Other cells	0.22 ± 0.06	0.84 ± 0.06	0.22 ± 0.06	0.84 ± 0.06	294
Artefacts	0.82 ± 0.05	0.74 ± 0.06	0.82 ± 0.05	0.74 ± 0.06	19630
Not identifiable	0.27 ± 0.04	0.63 ± 0.04	0.27 ± 0.04	0.63 ± 0.04	3538
Proerythroblasts	0.67 ± 0.10	0.85 ± 0.04	0.57 ± 0.09	0.63 ± 0.13	2740
Erythroblasts	0.79 ± 0.04	0.90 ± 0.04	0.33 ± 0.05	0.70 ± 0.08	2521
Normoblasts	0.93 ± 0.01	0.82 ± 0.02	0.92 ± 0.01	0.74 ± 0.04	24874
Hairy cells	0.80 ± 0.03	0.88 ± 0.02	0.35 ± 0.08	0.80 ± 0.06	409
Pathological eosinophils	0.02 ± 0.03	0.20 ± 0.40	0.02 ± 0.03	0.20 ± 0.40	8
Immature lymphocytes	0.35 ± 0.11	0.57 ± 0.13	0.08 ± 0.03	0.53 ± 0.15	65
Fagott cells	0.17 ± 0.05	0.63 ± 0.27	0.17 ± 0.05	0.63 ± 0.27	47

Table 1. **Class-wise precision and recall of the neural network classifier, as obtained by 5-fold cross-validation and denoted as mean ± standard deviation.** Results are shown both for the tolerant evaluation, allowing for mixups between classes that are difficult to distinguish, and the strict evaluation. The overall number of cell images contained in each class of the dataset is given in the right column.



## Figure Legends

Figure 1. **Structure of the 22 morphological classes of bone-marrow cells used in this work.** (A) Ordering of the classes into hematopoietic lineages. In agreement with routine practice, major physiological classes of myelopoiesis and lymphopoiesis are included, as well as characteristic pathological classes and classes for artifacts and unclear objects. (B) Distribution of the 171,374 non-overlapping images of our dataset into the classes used.

Figure 2. **Accurate ResNeXt prediction for most morphological classes.** Confusion matrix of the predictions obtained by the ResNeXt classifier on the test database annotated by gold-standard labels provided by human experts. Plotted values were obtained by 5-fold cross-validation, and are normalized row-wise in order to account for class imbalance. The number of single-cell images included in each category is indicated in the logarithmic plot on the right. Note enhanced confusion probability between consecutive steps of granulopoiesis and erythropoiesis, as might be expected due to unsharp delineations between individual morphological classes. Separate confusion maps of individual folds are shown in Fig. S2.

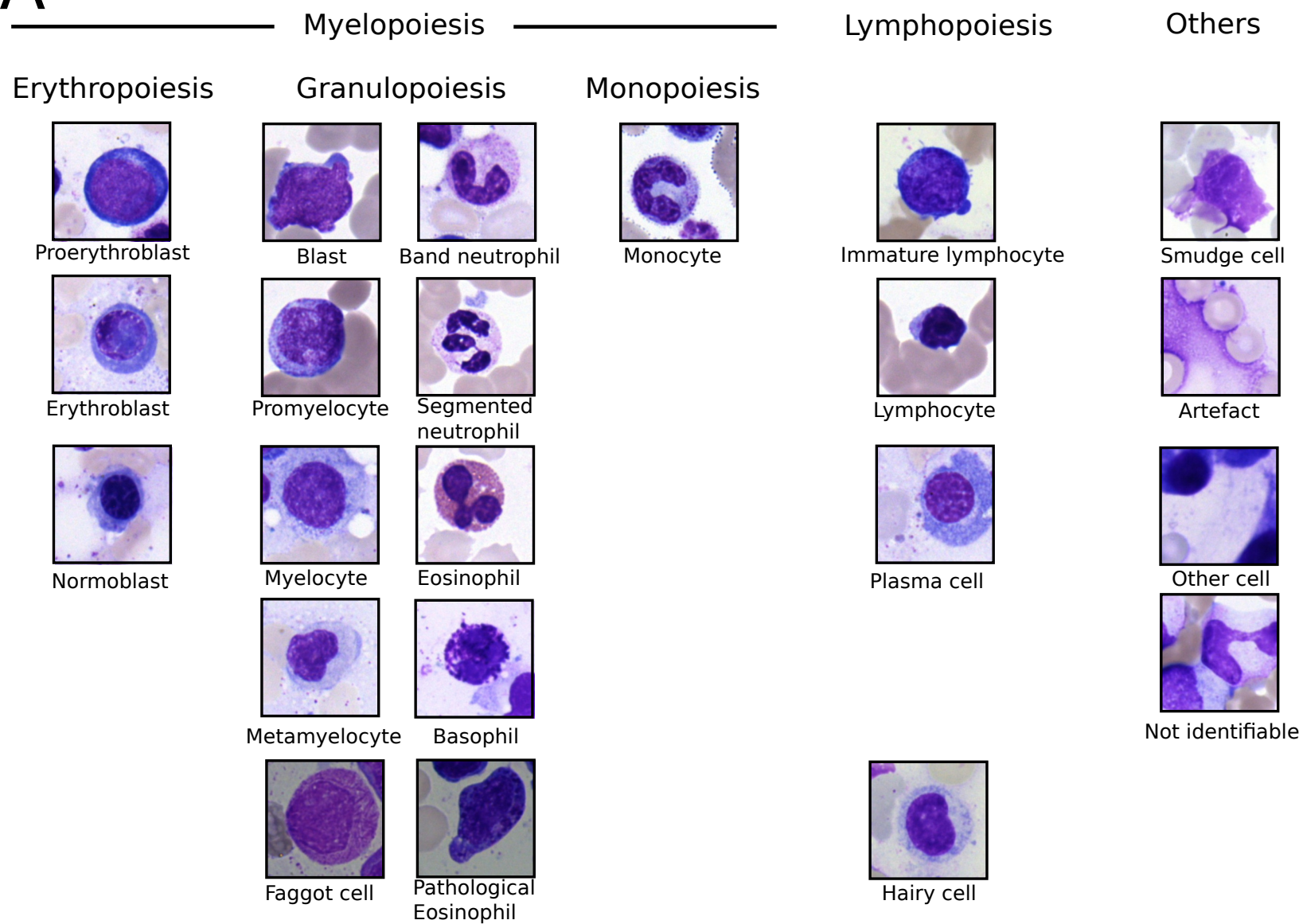
Figure 3. **Both CNN models outperform the feature-based classifier in terms of tolerant class-wise recall.** (A) Some morphological classes can be difficult to distinguish, so that a misclassification can be considered tolerable. Strict classification evaluation accepts only precise agreement of ground truth and network prediction, as shown in the black diagonal entries of the matrix. Mixups that are considered tolerable are colored blue. (B) Tolerance improvement for key classes. Error bars indicate standard deviation across 5 cross-validation folds. For segmented neutrophils and lymphocytes, performance of the feature-based classifier is slightly higher than of the neural networks. In all other classes, both CNNs consistently outperform the feature-based classifier of ref. <sup>13</sup>. This might be due to the distinctive signal of the nuclear shape of segmented neutrophils and lymphocytes in feature space. Additionally, ResNeXt outperforms the sequential network in several key classes, reflecting the greater complexity of the network used.

Figure 4. **Network prediction analysis shows focus on relevant image regions.** Original images classified correctly by the network are shown in the first row. The second row shows analysis using the SmoothGrad algorithm. The lighter a pixel appears the more it contributes to the classification decision taken by the network. Results of a second network analysis method, the Grad-CAM algorithm, are shown in the third row as a heatmap overlaid to the input image. Image regions containing relevant features are colored red. Both analysis methods suggest that the network has learned to focus on the leukocyte while ignoring background structure. Note the attention of the network to features known to be relevant for particular classes, such as the cytoplasmic structure in eosinophils, or the nuclear configuration in plasma cells.

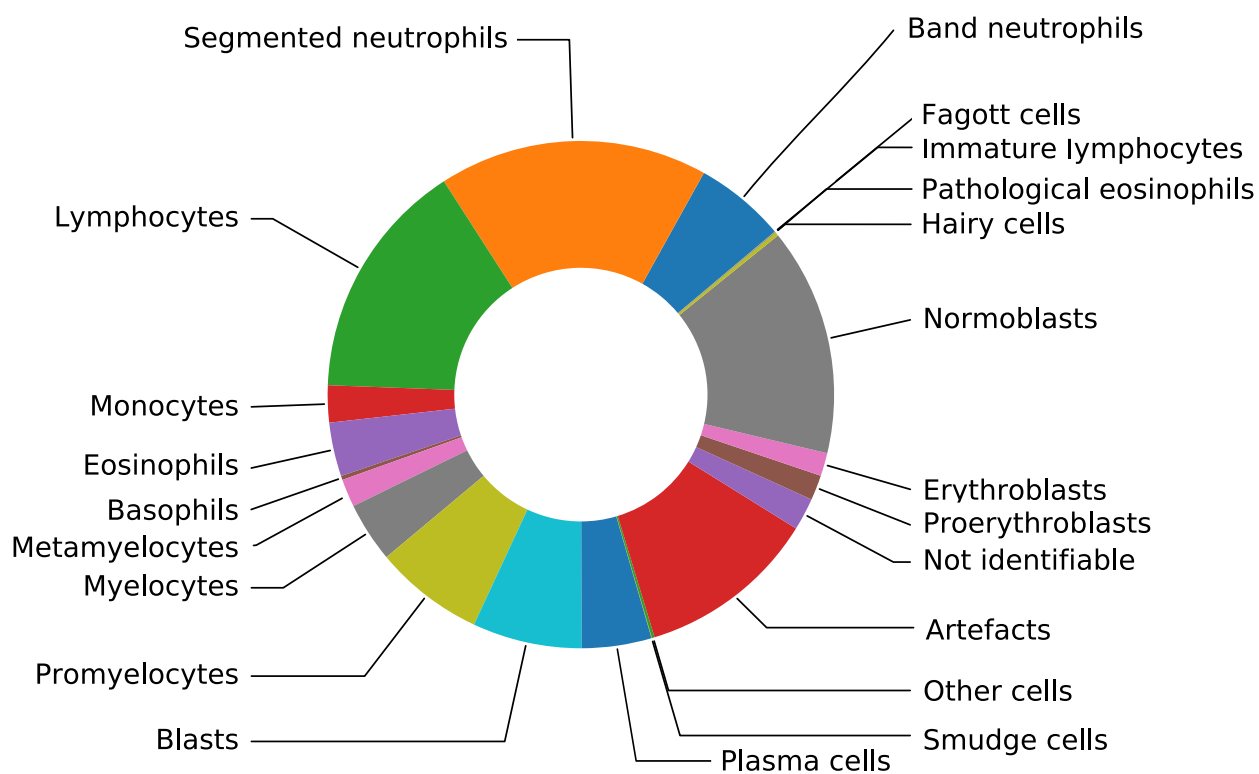
Figure 5. **The network exhibits fair performance on an external dataset.** Out of a total of 627 single-cell images from an external dataset, the network predicted classes on 380 images (61% of the dataset). The confusion matrix shows fair agreement between annotations of ref.<sup>20</sup> and the network predictions, which possess slightly different but compatible classification schemes. Good agreement is generally observed apart from confusions within the myeloid and erythroid lineages.

Figure 6. **UMAP embedding of extracted features.** UMAP embedding of the test set using the algorithm of ref.<sup>43</sup>. The flattened final convolutional layer of the ResNeXt-50 network containing 2048 features is embedded into two dimensions. Each point represents an individual single-cell patch and is colored according to its ground-truth annotation. All annotated classes are separated well in feature space. Cell types belonging to consecutive development steps tend to be mapped to neighbouring positions, reflecting the continuous transition between the corresponding classes.

# A

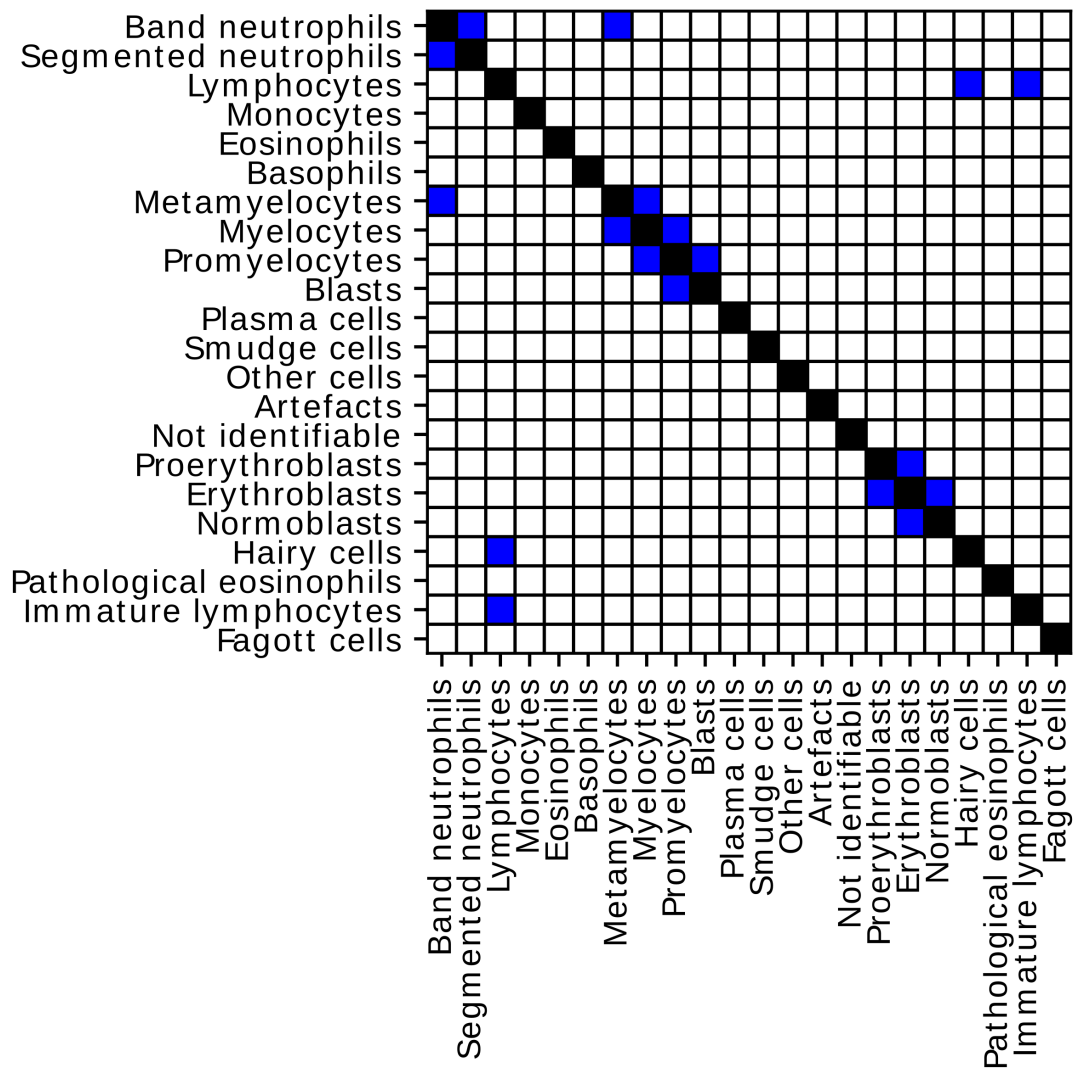


# B





A



B

