OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf239
https://doi.org/10.1093/bib/bbaf239
Problem Solving Protocol

Evaluation of out-of-distribution detection methods for data shifts in single-cell transcriptomics

Lauren Theunissen[1,2,3,*], Thomas Mortier[2,4], Yvan Saeys[1,3], Willem Waegeman[2]

[1]Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research and VIB Center for AI and Computational Biology (VIB.AI), 9000 Ghent, Belgium
[2]Department of Data-analysis and Mathematical Modeling, Ghent University Faculty of Bioscience Engineering, 9000 Ghent, Belgium
[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University Faculty of Sciences, 9000 Ghent, Belgium
[4]Department of Environment, Ghent University Faculty of Bioscience Engineering, 9000 Ghent, Belgium

*Corresponding author. E-mail: lauren.theunissen@ugent.be

Abstract

Automatic cell-type annotation methods assign cell-type labels to new, unlabeled datasets by leveraging relationships from a reference RNA-seq atlas. However, new datasets may include labels absent from the reference dataset or exhibit feature distributions that diverge from it. These scenarios can significantly affect the reliability of cell type predictions, a factor often overlooked in current automatic annotation methods. The field of out-of-distribution detection (OOD), primarily focused on computer vision, addresses the identification of instances that differ from the training distribution. Therefore, the implementation of OOD methods in the context of novel cell type annotation and data shift detection for single-cell transcriptomics may enhance annotation accuracy and trustworthiness. We evaluate six OOD detection methods: LogitNorm, MC dropout, Deep Ensembles, Energy-based OOD, Deep NN, and Posterior networks, for their annotation and OOD detection performance in both synthetical and real-life application settings. We show that OOD detection methods can accurately identify novel cell types and demonstrate potential to detect significant data shifts in non-integrated datasets. Moreover, we find that integration of the OOD datasets does not interfere with OOD detection of novel cell types.

Keywords: uncertainty; OOD detection; single-cell RNA-seq; novel cell-type detection; data shifts

Introduction

The development of automatic cell-type annotation tools has been a popular area of research in recent years. Large-scale atlases [1–3] have significantly enhanced the value of automatic cell-type annotation, as they can be used to build pre-trained annotation tools [4]. End-users can use these pre-trained tools [5–8] to assign cell-type labels to their own, often smaller and more specific, datasets and leverage the rich information present in the atlases. However, these smaller datasets can contain biological and/or technical variation that is not present in the training atlases, due to the presence of new labels, patients, disease states, tissues, or the use of different protocols, etc. This often results in a distribution or data shift between the reference data, here the atlas, and the test data, the smaller dataset. Data shifts severely impact the performance and reliability of the tools as generally multi-class classification models—such as cell-type annotation tools—assume that training and test datasets are independent and identically distributed (i.i.d.) according to an unknown distribution. In practice, these distribution shifts can lead to a much higher number of incorrectly allocated cell type labels than the end user might anticipate, given the tools' reported performance [9]. Ideally, any annotation tool should be able to accurately detect the presence of novel cell types in the test dataset, but also detect data shifts that severely influence the tool's annotation performance, so that the data shift can be mitigated with the help of integration techniques [10] or studied in more depth.

Out-of-distribution (OOD) detection is an interesting machine learning field that tries to handle the effects of data shifts during classification. OOD detection methods intend to identify and flag samples in the test data that are affected by data shifts, ensuring accurate and reliable classification [11]. Numerous OOD detection methods have been developed, particularly in the field of computer vision [12–17]. These methods in essence try to perform uncertainty quantification and threshold their uncertainty estimates to identify and flag aberrant samples, i.e. samples deviating from the training data distribution. Incorporation of OOD detection methods in automatic annotation tools to mitigate data shifts could help with the the detection of novel cell types, low quality samples or large data shifts. Current annotation tools in the single-cell field most often do not address any of these scenario's. A comprehensive comparison of different OOD methodologies for cell type annotation in single-cell transcriptomics, across various real-life biological settings, is currently lacking.

In this paper, we evaluate six established OOD methodologies from the computer vision field for detecting data shifts and novel cell types in single-cell transcriptomics across three datasets. We assess these methodologies based on their in-distribution (ID) annotation performance and their OOD detection capabilities. For OOD detection, we synthetically remove the least common cell type populations during training, as is commonly done in the literature, but also examine various applications with naturally occurring data shifts, such as annotating data from a new patient,

Received: January 28, 2025. **Revised:** April 01, 2025. **Accepted:** May 05, 2025

tissue or disease or data generated with a different protocol. We also explore how these biological shifts impact cell type annotation. Finally, we investigate the impact of integration on cell type annotation and OOD detection.

Materials and methods

Formal problem definition

Cell-type annotation is a multi-class classification problem, where the goal is to predict cell type labels from the label space $\mathcal{Y} = \{1, ..., K\}$, which contains K classes, based on inputs from some input space $\mathcal{X}$. The training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$ contains N data points, sampled i.i.d. from some unknown joint distribution $\mathcal{P}_{\text{train}}(\mathbf{x}, y)$ on $\mathcal{X} \times \mathcal{Y}$. Given $\mathcal{D}_{\text{train}}$, we will learn with most methods a classifier that returns probabilities. More formally, this is a classifier $f : \mathbf{x} \rightarrow \Delta^K$ with $\Delta^K = \{p : \sum_{i=1}^K p_i = 1,\ 0 \leq p_i \leq 1\}$. The classifier f will be optimized with the help of a loss function $\mathcal{L}(\cdot)$.

Dataset shifts occur when the training and test joint distributions are different $\mathcal{P}_{\text{train}}(\mathbf{x}, y) \neq \mathcal{P}_{\text{test}}(\mathbf{x}, y)$. Dataset shifts can be categorized into three different categories: covariate shifts, prior probability shifts and concept shifts. Covariate shift, also called population drift, refers to a change in the distribution of the input variable $\mathbf{x}$: $P_{\text{train}}(y|\mathbf{x}) = P_{\text{test}}(y|\mathbf{x})$, but $P_{\text{train}}(\mathbf{x}) \neq P_{\text{test}}(\mathbf{x})$. A prior probability shift refers to changes in the distribution of the label space $\mathcal{Y}$: $P_{\text{train}}(\mathbf{x}|y) = P_{\text{test}}(\mathbf{x}|y)$ but $P_{\text{train}}(y) \neq P_{\text{test}}(y)$. A concept shift or drift occurs when there is a change between the relationship of $\mathcal{X}$ and $\mathcal{Y}$. Formally, this can be defined as $P_{\text{train}}(y|\mathbf{x}) \neq P_{\text{test}}(y|\mathbf{x})$ but $P_{\text{train}}(\mathbf{x}) = P_{\text{test}}(\mathbf{x})$ [18]. In single-cell transcriptomics, covariance shifts can happen due to various biological factors. Occasionally, these shifts are accompanied by a prior probability shift because of new cell type labels in the test data. A concept shift is not naturally present in the setting of cell type annotation for single-cell transcriptomics and will thus not be considered in this evaluation (see Datasets).

Out-of-distribution detection is a binary classification problem, where the goal is to label an input $\mathbf{x} \in \mathcal{X}$ as ID if it belongs to $\mathcal{P}_{train}$ and OOD if it does not. In practice, OOD detection is often implemented with the help of a scoring function $S(\mathbf{x})$ so that:

$$\hat{y}^{ood} = \begin{cases} ID & \text{if } S(\mathbf{x}) \geq \tau, \\ OOD & \text{if } S(\mathbf{x}) < \tau \end{cases} \qquad (1)$$

where τ is a user-defined threshold and $\hat{y}^{ood}$ the OOD prediction. In this paper we evaluate the methods' ability to both correctly perform ID classification, as well as to correctly perform OOD classification. This setting is also referred to as open-set recognition, open category detection or open set learning in the machine learning literature, but not all authors adhere to this nomenclature [11].

OOD methods

In this paper, we evaluate six OOD methods: LogitNorm [19], MC dropout [20], Deep Ensembles [21], Energy-based OOD (EBO) [22], Deep nearest neighbors (Deep NN) [23], and Posterior Networks [24]. We chose these methods based on their performance in the recent out-of-distribution benchmark by Yang *et al.* [12] and their popularity.

LogitNorm: The authors of the LogitNorm method start from the assumption that the model should produce predictions with lower confidence scores for OOD data, compared to ID data [19]. However, in reality, the confidence scores are often overconfident, i.e. high scores are assigned to all predictions regardless of their correctness. This problem mainly occurs in deep neural networks [25]. We will use $h(\mathbf{x})$ to denote the outputs of the penultimate layer of the neural network $f(\mathbf{x})$. These outputs are commonly referred to as the logits. The authors consider the cross-entropy loss $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L}_{\text{CE}}(f(\mathbf{x}), y) = -\log \frac{e^{h_y(\mathbf{x})}}{\sum_{j=1}^{K} e^{h_j(\mathbf{x})}}, \qquad (2)$$

to optimize the NN model f and show that the overconfidence of the model is caused by the cross-entropy loss that keeps increasing the magnitude of the logit vectors, even when a sample is already correctly classified. To alleviate this problem the logit normalization loss, dubbed the LogitNorm loss $\mathcal{L}_{\text{LogitNorm}}$, is proposed as an alternative to the cross-entropy loss:

$$\mathcal{L}_{\text{LogitNorm}}(f(\mathbf{x}), y) = -\log \frac{e^{h_y(\mathbf{x})/(t||h(\mathbf{x})||)}}{\sum_{j=1}^{K} e^{h_j(\mathbf{x})/(t||h(\mathbf{x})||)}}, \qquad (3)$$

where the temperature t will modulate the magnitude of the logits $h(\mathbf{x})$. $h(\mathbf{x})$ can be decomposed into two components without loss of generality: the euclidean norm $||h(\mathbf{x})||$, which represents the magnitude of the logit vector, and the unit vector $\widehat{h}(\mathbf{x})$, which depicts the direction. The LogitNorm loss $\mathcal{L}_{\text{LogitNorm}}$ (Equation 3) decouples the influence of the logits' magnitude from the training process [19]. The scoring function used for OOD detection is derived from the probabilities returned by the neural network f, which was trained with the logitnorm loss $\mathcal{L}_{\text{LogitNorm}}$: $S(\mathbf{x}) = \mathbf{max}_{k=1,...,K} f_k(\mathbf{x})$.

MC dropout is a well-known randomization technique that was initially introduced to improve predictive performance, but nowadays it is also commonly used for uncertainty quantification and OOD detection. With the Dropout mechanism, each node of each layer is dropped or excluded from the NN during training with a dropout probability of 0.5. In this way, for each training sample a different thinned network is sampled and trained [26]. Gal *et al.* [20] show that an NN, with dropout applied before every weight layer, can be interpreted as a Bayesian approximation of the Deep Gaussian Process model. With MC dropout, T stochastic forward passes are made through the dropout network. So T thinned networks are sampled and evaluated. The final prediction scoring function, used for OOD detection and cell type annotation, is the softmax of the averaged logits over the T forward passes or evaluated networks: $f(\mathbf{x}) = \text{Softmax}\left(\frac{\sum_{m=1}^{T} h(\mathbf{x})_m}{T}\right)$ and $S(\mathbf{x}) = \mathbf{max}_{k=1,...,K} f_k(\mathbf{x})$[20].

Deep Ensembles improves model uncertainty estimation by using an ensemble of NN-models, trained with a proper scoring rule (A proper scoring rule S_{proper} is a scoring rule such that $S_{proper}(P_\theta, P) \leq S_{proper}(P, P)$ if and only if $P_\theta(y|\mathbf{x}) = P(y|\mathbf{x})$.) such as the cross entropy loss $\mathcal{L}_{CE}$ (Equation 2) [21]. With Deep Ensembles T NN-models are randomly initialized, independently trained, and the softmax averages of the T logits $h(\mathbf{x})$ are used for OOD detection and cell type annotation: $f(\mathbf{x}) = \text{Softmax}\left(\frac{\sum_{m=1}^{T} h(\mathbf{x})_m}{T}\right)$ and $S(\mathbf{x}) = \mathbf{max}_{k=1,...,K} f_k(\mathbf{x})$.

Posterior Networks explicitly model an (epistemic) uncertainty distribution $\boldsymbol{q}$ characterized by the family of Dirichlet distributions, over the categorical class distribution $\mathbf{p} = [p_1, ..., p_K]$ that is estimated by $f(\mathbf{x})$ for every $\mathbf{x}$. Deep Ensembles and MC dropout represent $\boldsymbol{q}$ implicitly, and can only sample from it and estimate statistics on it. The explicit parametrization of this uncertainty distribution $\boldsymbol{q}$ over the categorical class distribution $\boldsymbol{p}$ allows to compute and distinguish between epistemic uncertainty, aleatoric uncertainty and class predictions in one pass. Posterior Networks map $\mathbf{x}$ to a latent space $\mathcal{Z}$ with the help of an encoder network $f_{Encoder}$. They also train a Normalizing flow, parameterized by ϕ, to learn flexible density functions $P(\mathbf{z}|\phi)$ for every class, that are evaluated at the positions of the latent vectors $\mathbf{z} \in \mathcal{Z}$. The resulting densities are used for the parametrization of the Dirichlet distribution $\boldsymbol{q}$ for each data point $\mathbf{x}$. Higher densities correspond to higher confidence in the Dirichlet distribution. When $\boldsymbol{x}$ corresponds to a low density region, a very high epistemic uncertainty will be predicted by the Dirichlet distribution. Both the encoder network and the normalizing flow are jointly optimized with the help of the uncertainty-aware loss $\mathcal{L}_{\mathrm{UCA}}$ [24]:

$$\mathcal{L}_{\mathrm{UCA}}(f_{Encoder}, \phi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{p}_i)}[\mathcal{L}_{\mathrm{CE}}(\mathbf{p}_i, \mathbf{y}_i)] - H(\mathbf{q}_i), \tag{4}$$

where N indicates the number of observations, $\mathbf{y}$ is the one-hot-encoded vector representation of y and $H(\mathbf{q})$, an entropy regularizer that promotes smooth distributions of $\mathbf{q}$. The normalized densities for every sample are used to perform OOD detection and cell type annotation. Class probabilities are obtained from the densities by applying Bayes' rule, while the OOD detection score $S(\mathbf{x})$ is proportional to the highest class density.

Energy-based OOD is a *post-hoc* method that consists of calculating an energy-score S_E after training and inference based on the logits $h(\mathbf{x})$ returned by the NN model f:

$$S_E(f(\mathbf{x}), t) = -t\sum_{j=1}^{K} e^{h_j(\mathbf{x})/t}, \tag{5}$$

where t is a temperature parameter that rescales the logits $h(\mathbf{x})$. Samples with higher energy scores will be seen as more likely OOD, thus the negative values of the energy scores will be used for OOD detection [22].

Deep nearest-neighbors: The Deep NN method performs non-parametric density estimation with the help of k-nearest neighbors. The method calculates the Euclidean distance in the space formed by the normalized penultimate layer embedding: $h_{norm}(\mathbf{x}) = h(\mathbf{x})/||h(\mathbf{x})||_2$. This is done for every observation of the test data $h_{norm}(\mathbf{x}_i^{test})$ and its k-th nearest neighbor $\mathbf{x}_k$ in the normalized-penultimate embedded training data $h_{norm}(\mathbf{x}_k^{train})$. This distance is then used as a score for OOD detection [23].

Datasets and model construction

We evaluated the OOD methods across three different datasets: the Lung, Immune, and COPD dataset [10, 27]. Information on the preprocessing of the datasets can be found in Supplementary Appendix 1.1. Characteristics of the datasets are reported in Table 1. The Lung dataset consists of three healthy 10X transplant datasets, one Drop-seq transplant dataset and one 10X lung biopsy dataset. The Immune dataset contains peripheral blood mononuclear cell (PBMC) data, sequenced with SMART-seq and 10X, together with Bone marrow data sequenced with 10X. The COPD dataset contains samples from healthy (smoker and non-smoker) patients, patients with idiopathic pulmonary fibrosis (IPF) and patients with chronic obstructive pulmonary disease (COPD), all sequenced with 10X. Inspection of the COPD dataset showed no clear disease (batch) effects between the disease and control data parts, resulting in our believe that the dataset is most likely integrated (Fig. 1A), though this is not clearly specified by the authors in the manuscript or data file [27]. Each OOD method was implemented with multiple NN architectures for f. The results of the best overall performing networks are reported. The effect of integration was further evaluated with the help of the overall best performing networks on the Lung and Immune dataset, for more information on data integration and model construction, we refer the reader to Supplementary Appendix 1.2.

OOD scenarios

With the help of these datasets, we mimicked several application settings, where the test data could naturally contain a data shift. For the Immune and Lung dataset, we considered the annotation of a new patient dataset, a dataset generated by a different protocol and a dataset originating from a different tissue. For the COPD dataset, again annotation of a new patient was considered, but also annotation of two dataset parts with a disease effect. Figure 1A visualizes the application settings and their corresponding ID and OOD data parts for the three datasets. To evaluate the effect of a patient shift across all the three datasets, a grouped cross-validation scheme was implemented so that each patient in the dataset was once considered as test data. The resulting metrics were averaged over all the patients (see Fig. 1A). Patients with less than 500 sequenced cells in total were excluded from the analysis.

We evaluated annotation performance on the ID data and OOD data separately and evaluated OOD detection. In order to do this, a part of the ID data were included in the test data to evaluate the ID classification performance of the model (see Fig. 1B). Before we could evaluate OOD detection performance, a ground truth OOD label needed to be determined for all the samples in the test data, indicating whether each sample is ID or OOD. We evaluated OOD detection for two scenarios: (i) a scenario where a severe data shift is present, so the goal is to detect the entire OOD data part as OOD and (ii) a scenario where the OOD detection goal is to detect novel cell types (cell types not present in the training data). This latter scenario is not applicable for all application settings as not all the test datasets (naturally) contain novel cell types (Fig. 1C). Next to these biological OOD settings, we also tested the OOD performance of the methods in an artificial OOD setting. For the 10X healthy control data of the COPD dataset, we excluded the 2%,5%, or 10 % least occurring cell types during training (see Fig. 1D). This resulted in respectively the exclusion of 22, 34, and 42 out of 60 cell types, due to the severe imbalanced nature of the data. We will refer to this experiment as minor novel cell type detection in the rest of this manuscript. For all the training and test datasets (incl. the OOD data parts), with exception of the biological patient OOD setting, cell type populations with less than 10 observations were filtered out. For the biological patient OOD setting this filtering was performed on the entire dataset across all patients.

Evaluation metrics

To evaluate the cell-type annotation performance, the ID accuracy and OOD accuracy metric are calculated based on the cell type labels assigned to the cells. The former is calculated on the part of the in-distribution data amended to the test dataset, the latter is calculated on the OOD data (see Fig. 1B). For the OOD detection evaluation, the AUROC metric and the false positive rate (FPR) for a true positive rate (TPR) of 95% are reported based on the predicted OOD labels. To calculate these metrics, ground truth OOD labels needed to be assigned. To do so, we assigned a positive value (1) to OOD samples and a negative value (–1) to ID samples. We considered two scenarios (i) all samples from the OOD data part are OOD (AUROCd, FPRd) and (ii) only the samples with cell type labels unseen during training, i.e. novel cell types (if present in the test data) are OOD (AUROCc, FPRc).

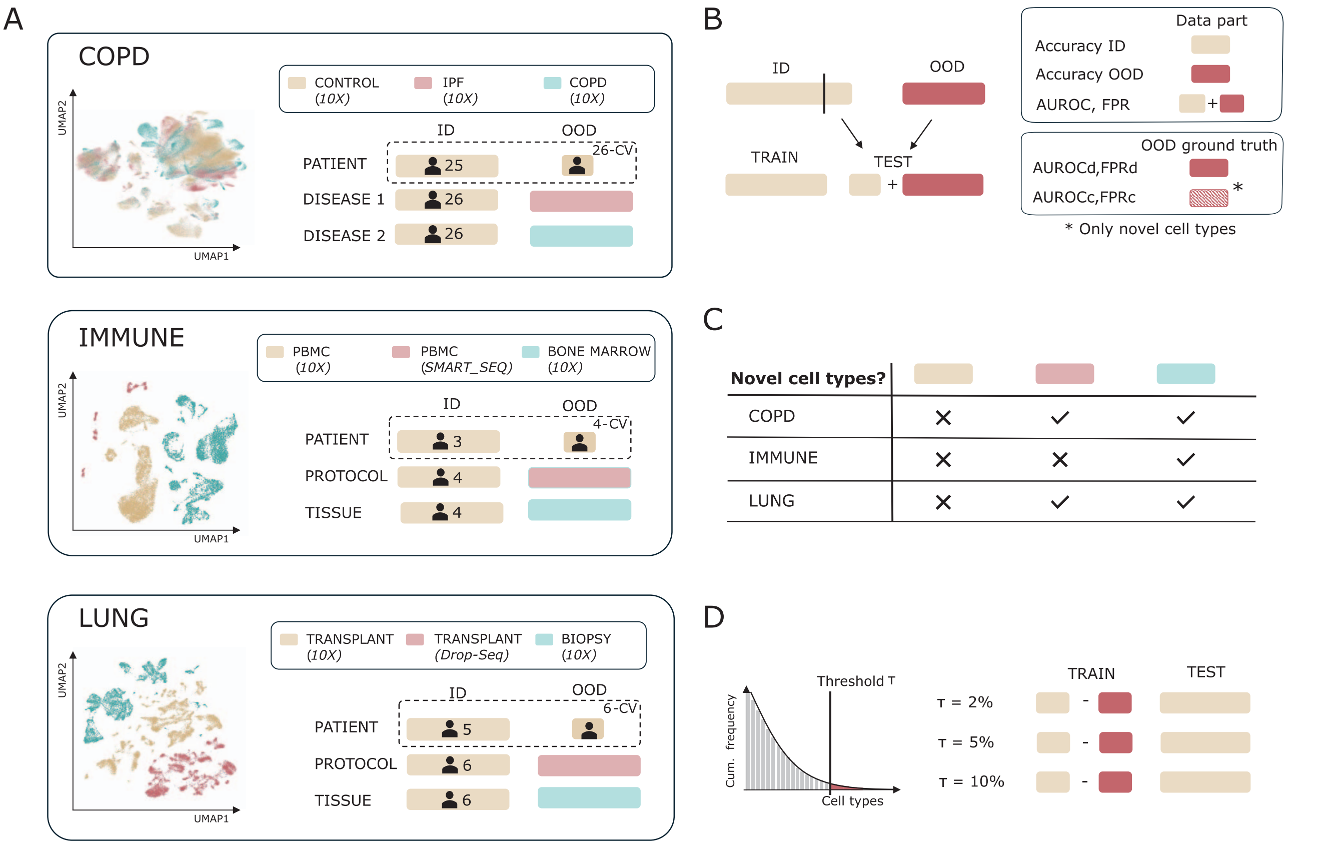

Figure 1. (A) A graphical representation of the biological application settings where the six OOD methods are tested on, simulated with the Lung, Immune, and COPD dataset [10, 27]. (B) A train-test split scheme for all the application settings that indicates which data parts are used to measure in-distribution and out-of-distribution performance (ID and OOD accuracy) and OOD detection evaluation (AUROC, FPR). Panel B also visualizes the parts of the OOD data that are truly considered OOD (= the ground truth) for calculation of the OOD detection metrics, distinguishing between AUROCc/FPRc and AUROCd/FPRd. (C) Indicates the presence of naturally occurring novel cell types in the OOD data across all biological application settings. The colors correspond to the colors used in panel A to indicate the biological application splits. (D) A visual representation of the minor novel cell type analyses, where for the 10X healthy control data of the COPD dataset [27], a percentage of cell types, ranked based on their abundance in the dataset, was excluded from the training data.

Results and discussion

OOD detection does not influence ID annotation

The results of the six OOD methods on the Immune and Lung datasets are presented in Supplementary Tables A1 and A2, and Fig. 2. The ID performance is high and similar across all methods, except for the Posterior Network. This was expected, as the OOD methods, with the exception of the Posterior Network, retain the main training objective to correctly classify ID cell types, similar to a general cell type annotation tool, and slightly adapt the learning loss or use the output of this classification task to perform OOD detection.

The Posterior Network, however, performs significantly worse for ID classification in comparison to the other methods and gave unstable results across multiple analysis runs. To mitigate this instability, we ran all the Posterior Network analyses three times and reported the average. A possible explanation for this instability could be due to the loss function, as it has been reported that loss functions of deep evidential methods such as the posterior networks are unstable [28, 29]. The authors reported comparable results with other OOD methods for computer vision data, but single-cell data differs significantly, making it difficult to pinpoint the cause of the inferior performance. Notably, the authors themselves state that improving (ID) accuracy is not the goal of their method [24].

Our hyperparameter tuning showed that non-linear networks consistently performed slightly better than linear ones, contradicting the literature [30–32]. A recent paper suggested that non-linear methods outperform linear ones in a cross-tissue context with large-scale datasets [33]. However, our OOD patient splits occur within one tissue and with training datasets of a relative small size. Given the small performance differences and the reasonably non-complex datasets, we conclude that optimal ID classification performance may have been achieved for these datasets.

OOD methods are able to accurately detect novel cell types and possibly large data shifts

As mentioned before, we considered two possible OOD scenarios (i) the presence of a severe data shift resulting in (desired) OOD detection of the entire OOD data part and (ii) detection of novel cell types in the absence of a large data shift. In order to conclude which OOD scenario is desired for the different biological annotation applications, we calculated the Wasserstein's distance across the first 100 principal components of the ID and OOD data parts. For more information on this calculation, we refer the reader to Supplementary Appendix 1.3. Supplementary Figure A2.A and B shows these distances for each annotation application of the Lung and Immune datasets. The largest difference between the training

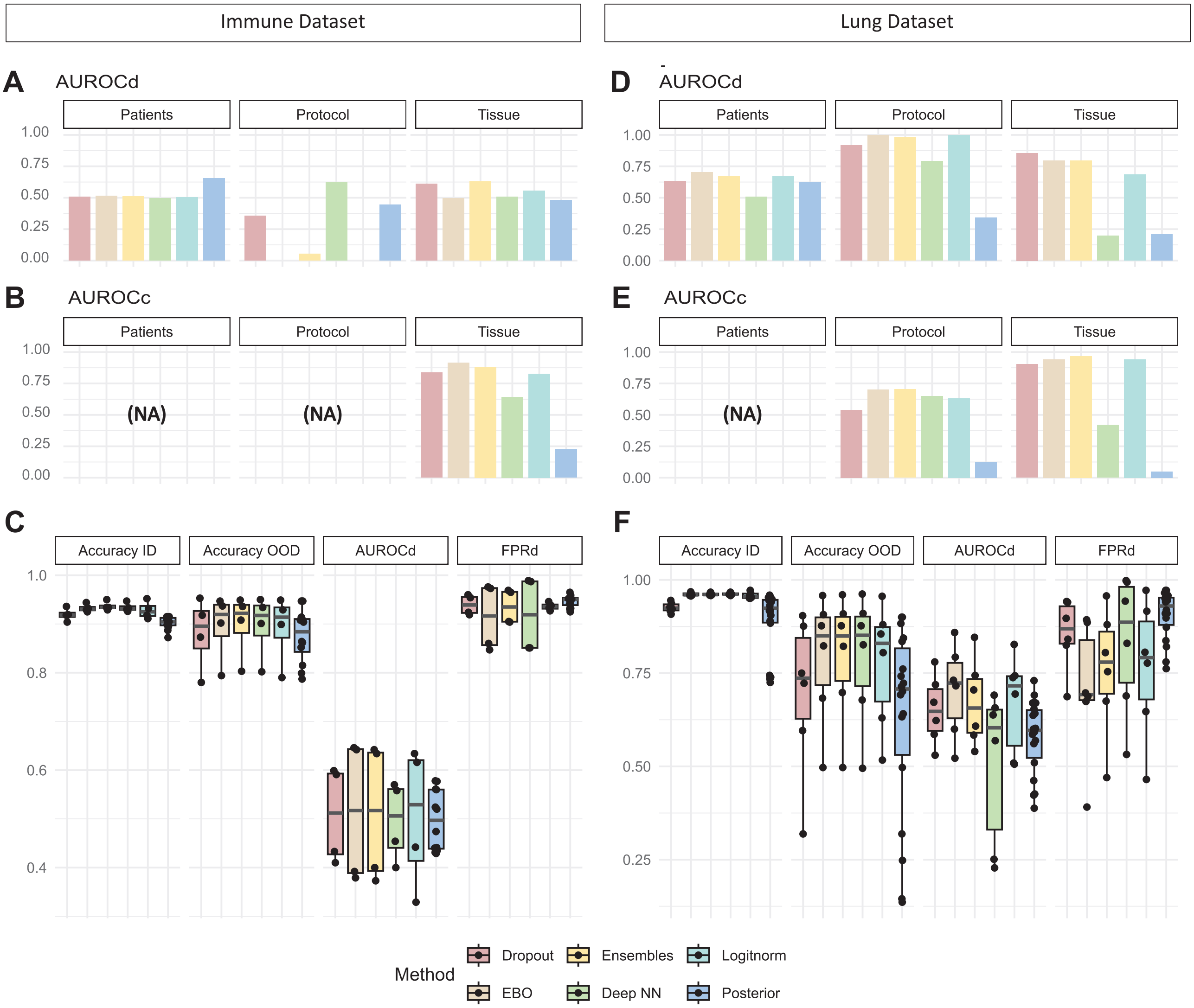

Figure 2. Overview of the OOD detection results of the six methods on the Immune (A–C) and the Lung dataset (D–F) across the biological application settings. Figures A and D display the AUROCd or AUROC dataset metric, where the AUROC is calculated when the entire OOD data part of the test dataset is desired to be detected as OOD, for all the six methods. Figures B and E display (if applicable) the AUROCc or AUROC cell type metric, where the AUROC metric is calculated when the goal is to detect naturally occurring novel cell types in the OOD data part of the test dataset, for all the 6 methods. The box plots in figures C and F show the ID accuracy, OOD accuracy, AUROCd, and FPRd when the TPR is 95% for all the patients separately for the patient OOD splits. Every point represents one analysis, which for all methods besides the posterior network, corresponds with one patient.

and OOD distributions occurs in the protocol OOD scenario, indicating that OOD detection of the entire OOD protocol data part might be desired for both datasets.

Figure 2 shows the OOD detection performance of different methods for the Lung and Immune datasets. For the Immune dataset (Fig. 2A–C), novel cell types naturally occurred only in the tissue OOD split, where the Energy-based OOD detection method performed best with an AUROC of 91.5%. In the protocol split, no method accurately detected the entire dataset as OOD, despite the large Wasserstein distances (Supplementary Fig. A2.B). For the Lung dataset (Fig. 2D–F), most methods accurately detected the entire protocol OOD data part, with Energy-based OOD detection achieving an AUROC of 100%. The naturally occurring novel cell types in the tissue split were best recognized by Deep ensembles (AUROC 96.%), followed closely by the Energy-based OOD (AUROC 94.2%) and LogitNorm (AUROC 94.1%) methods. Overall, the methods performed well in recognizing novel cell types in the tissue OOD split, with the Energy-based OOD method being the top performer.

Based on the mixed results across the two datasets for the protocol splits, it is unclear if the methods can accurately flag entire data parts affected by large batch effects due to protocol differences or other variations leading to covariate shifts. The Immune dataset, which has the largest distance between protocol distributions, is not flagged, while the Lung OOD protocol is. Given the clear distinctions in the UMAPs in Fig. 1, we believe separation should be possible, and these protocols should be able to detect the splits based on their mode of operation. A potential bottleneck is that the OOD methods operate in the embedded space of the ID annotation task, which may be suboptimal for OOD detection, especially for detecting covariate shifts.

We also visualized for the patient splits the relation between the OOD accuracy and AUROCd in Supplementary Fig. A3, to see if severe patients shifts are being picked up by the best performing methods: Deep ensembles and EBO. It seems that most of the patients with a severe data shift are picked up as being OOD by the detection methods across the datasets. So overall, the OOD methods are able to pick up severe data shifts, though not consistently across all datasets.

Patient, protocol and tissue each have a clear effect on annotation with increasing severity

A key question is how the biological settings affect the ID and OOD annotation performance. Based on the results reported in Supplementary Tables A1 and A2, a conclusion could be made that, on average, patient effects are negligible as the accuracy drop from ID to OOD annotation is relatively small, especially for the Immune dataset. However, as visualized in Fig. 2C and F, for some patients a clear data shift is occurring, indicating the importance of checking or accounting for possible patient effects during cell type annotation.

To find out the influence of tissue and protocol shifts on cell type annotation, we re-visualized the results in Supplementary Figs A4 and A5 to clearly see the influence of the different OOD splits. These results show that both a protocol and tissue shift have a severe influence on annotation performance and that the influence of a new tissue is significantly larger than that of a new protocol. These results contradict the results of the Wasserstein distance measures. An explanation for this lies in the calculation of the Wasserstein distance and the nature of the data generated by different protocols. Protocols like 10X, SMART-Seq, and Drop-seq have varying sensitivities, i.e. detected genes per cell, and capture efficiency, leading to noticeable global data shifts captured by the Wasserstein distance metric in the reduced Principal Component Analysis (PCA) space [34, 35]. However, cell-type annotation mainly relies on a limited set of marker genes. If the signal in the individual marker genes is conserved, cell type annotation will not be significantly affected by the large data shifts. Data distributions of different tissues sequenced with the same protocol will show biological variation and data shifts, which may be present in marker genes. This can hinder cell type annotation without greatly impacting the Wasserstein metric. Therefore, a large Wasserstein distance will indicate a severe data shift, that ideally would be flagged by the OOD methods. But the Wasserstein distance does not necessarily reflect annotation performance on the OOD dataset.

Integration does not interfere with novel cell-type detection

The results for the integrated COPD dataset are presented in Supplementary Table A3 and Supplementary Figs A1 and A6. Integration significantly improves the annotation performance of the CODP OOD data. The performance drops between ID and OOD annotation performance seen on the Lung and Immune dataset are not visible on this dataset. For this dataset, the COPD disease effect seems the most compensated, the ID and OOD accuracy difference is the smallest for this set-up. The drop increases slightly for the IPF data and between the patients. However, the detection performance of naturally occurring novel cell types also drops. For the IPF OOD data part, Ensembles perform the best with an AUROC of 75.6%. For the COPD OOD data part, Energy-based OOD performs the best with an AUROC of 70.5%.

To check what the influence is of integration on OOD performance and whether it plays a part in the slightly diminished OOD detection performance for the COPD dataset, we integrated the Lung and Immune dataset and performed OOD detection with the best performing setups on the non-integrated datasets. The influence of integration on the data is visualized in Supplementary Figs A8 and A9 for the Immune and Lung data, respectively, and the OOD detection results are visualized in Supplementary Figs A10 and A11. As visible in Supplementary Figs A8 and A9, integration was not perfect, which makes it hard to formulate concrete conclusions in regards to the OOD detection of the full data part or OOD accuracy. Our results do indicate that integration does not seem to influence novel cell type detection. This indicates that the slightly diminished OOD performance on the COPD data most likely has to do with the biological setting of the dataset or the larger number of patients present in the dataset, leading to more observed variation.

Note that for the COPD dataset, we did not report the results of Posterior Networks, since the method failed to learn anything during training across all our setups (OOD scenarios, network configurations, etc.), i.e. the uncertainty-aware loss value did not change during training. To address this, we tried varying the learning rates (1e-1 to 1e-8), we increased the number of epochs and relaxed the early stopping criteria, but no improvement was observed. Since Posterior Networks train a normalizing flow and embedding network in one pass, identifying the bottleneck is challenging. Given the COPD dataset's larger size, more features, and more classes compared to the other datasets (see Table 1), there could be multiple reasons why this method did work on other datasets but not on the COPD dataset. We hypothesize that the reason for the network's inability to learn from this dataset lies in challenges associated with density estimation; however, further investigation regarding the obtained results are required and is left as future work.

Synthethic OOD evaluation is not representative for real world performance

We also performed a synthetic minor novel cell type split on the COPD 10X control data (Supplementary Figs A1B and A7). Here, not Energy-based OOD, but Deep NN was the best performer, underlining the importance of evaluating OOD methods in real-life scenarios. For the 2%, 5%, and 10% hold-out schemes the AUROC values for the novel cell type detection were 82.8%, 86.5%, and 80.1%, respectively. This demonstrates the method's insensitivity to the number of novel cell types, as is the case for all the other methods, with exception of the Ensembles, where detection performance slightly improved as more cell types were rejected.

Practical guidelines

Based on all our results, we would recommend users and developers that are interested in more robust and accurate cell type annotation to use Energy-based OOD after integration of the dataset with the reference dataset. Because Energy-based OOD detection performed the best across our analyses and is easy to implement on top of existing annotators. Energy-based OOD calculates a post-hoc score that has a linear time complexity for the number of cell types, meaning that it does not involve any change to the existing annotator algorithm and it does not have a large computational burden. A general downside to OOD detection is that it does require thresholding of an OOD score (see Materials and methods). The determination of this threshold will balance the amount of correctly flagged samples as OOD with the amount of wrongly flagged samples, as the OOD detectors are not 100% accurate. The implementation of OOD detection will always result in some incorrectly flagged samples. So, it is up to the user or developer to decide if this is an issue.

Table 1. Characteristics of the biological application settings of the three datasets (Lung, Immune, and COPD) used to evaluate OOD detection and details of the experimental setups of the biological application settings

Dataset	N° cells	N° genes	N° cell populations	Biological application		
	ID/OOD		ID/OOD (NC)[a]	Split	Protocol	Tissue/Disease
Immune	8829[b]	12 303	10[b]	Patient	10X	PBMC
	8829/1022	12 303	10/10 (-)	Protocol	Smart-seq	PBMC
	8829/9581	12 303	10/16 (6)	Tissue	10X	Bone marrow
Lung	11 828[b]	15 148	13[b]	Patient	10X	Transplant
	12 716/9701	15 148	13/11 (1)	Protocol	Drop-seq	Transplant
	12 716/10 033	14 148	13/10 (3)	Tissue	10X	Biopsy
COPD	95 506[b]	45 497	78[b]	Patient	10X	Control
	96 260/147 158	45 947	78 / 97 (19)	Disease 1	10X	IPF
	92 260/69 423	45 947	78 / 86 (8)	Disease 2	10X	COPD

[a]NC stands for new cell types, cell types not present in the ID data. [b]As for the patient splits a group cross-validation scheme is used, where each patient is separately used as test data, the total number of cells and cell-type populations across all patients is reported.

Conclusion

In this paper, we evaluated six established OOD detection methods from the computer vision field for cell-type annotation of single-cell transcriptomics data, specifically for the detection of data shifts and novel cell types to improve cell type annotation. We evaluated their performance on a synthetic use-case and on real-life biological data shifts such as the introduction of new patients, a new protocol, a new tissue or a new disease. We performed these analyses on three datasets and investigated the influence of integration on OOD detection performance. Based on our results, we recommend to use Energy-based OOD detection for novel cell-type detection, as it overall performed best in our evaluation on real-life biological data shifts. Energy-based OOD is also computationally very efficient and easy to implement as it uses a post-inference score for OOD detection. We illustrated the importance of including real-life biological scenario's in the OOD detection evaluation as they severely impacted the cell type annotation performance and gave different results for OOD detection in comparison to the synthetic evaluation. Our results showed promise for the OOD detection methods to also be used for detecting data shifts, but the results were inconsistent across dataset, so more research needs to be performed before a firm conclusion can be made. Lastly, we saw that integration does increase cell type annotation on data under the influence of a data shift and does not negatively influence novel cell-type detection.

Key Points

- Detecting novel cell-type detection works well using OOD methods, with Energy-based OOD detection performing best on all our datasets
- OOD methods can only be properly evaluated on real-life biological data shifts
- Introduction of new patients, protocols, tissues or diseases can have a significant impact on cell-type annotation.
- OOD detection methods can identify severe data shifts, but not reliably
- Integration of datasets does not hinder novel cell type detection.

Author contributions

Lauren Theunissen, Thomas Mortier, Yvan Saeys, and Willem Waegeman conceived the experiments; Lauren Theunissen conducted the experiments, analyzed the results, and wrote the manuscript; Lauren Theunissen, Thomas Mortier, Yvan Saeys, and Willem Waegeman reviewed the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the Flemish Government under the Flanders AI Research Program (Project 174K02325).

Data availability

The datasets were derived from sources in the public domain, they can be accessed through the repositories mentioned in the corresponding articles. Code to reproduce the analyses in this paper is freely available at https://github.com/Latheuni/OOD_eval.git.

References

1. Regev A, Teichmann SA, Lander ES. *et al*. The human cell atlas. *eLife* 2017;**6**. https://doi.org/10.7554/eLife.27041
2. Sikkema L, Ramírez-Suástegui C, Strobl DC. *et al*. An integrated cell atlas of the lung in health and disease. *Nat Med* 2023;**29**: 1563–77. https://doi.org/10.1038/s41591-023-02327-2
3. Deprez M, Zaragosi LE, Truchi M. *et al*. A single-cell atlas of the human healthy airways. *Am J Respir Crit Care Med* 2020;**202**: 1636–45. https://doi.org/10.1164/rccm.201911-2199OC
4. Hrovatin K, Sikkema L, Shitov VA. *et al*. Considerations for building and using integrated single-cell atlases. *Nat Methods* 2024;**22**: 41–57. https://doi.org/10.1038/s41592-024-02532-y

5. Li R, Zhang J, Li Z. Easycelltype: marker-based cell-type annotation by automatically querying multiple databases. *Bioinform Adv* 2023;**3**. https://doi.org/10.1093/bioadv/vbad029
6. Hou W, Ji Z. Assessing gpt-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5. https://doi.org/10.1038/s41592-024-02235-4
7. Shao X, Yang H, Zhuang X. *et al*. Scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res* 2021;**49**:e122. https://doi.org/10.1093/nar/gkab775
8. Yang L, Ng YE, Sun H. *et al*. Single-cell Mayo Map (scMayoMap): an easy-to-use tool for cell type annotation in single-cell RNA-sequencing data analysis. *BMC Biol* 2023;**21**:223. https://doi.org/10.1186/s12915-023-01728-6
9. Sun X, Lin X, Li Z. *et al*. A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. *Brief Bioinform* 2022;**23**:01. https://doi.org/10.1093/bib/bbab567
10. Luecken MD, Büttner M, Chaichoompu K. *et al*. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
11. Yang ɪ, Zhou K, Ling Y. *et al*. Generalized out-of-distribution detection: a survey. *Int J Comput Vis* 2024;**132**:5635–62. https://doi.org/10.1007/s11263-024-02117-4
12. Yang J, Wang P, Zou D. *et al*. Openood: benchmarking generalized out-of-distribution detection. In: *Advances in Neural Information Processing Systems*, Vol. **35**, 2022.
13. Yang J, Zhou K, Li Y. *et al*. Generalized Out-of-Distribution Detection: A Survey. Int J Comput Vis 2024;**132**:5635–62. https://doi.org/10.1007/s11263-024-02117-4
14. Liu J, Shen Z, He Y. *et al*. Towards out-of-distribution generalization: a survey. 2023. https://doi.org/10.48550/arXiv.2108.13624
15. Cui P, Wang J. Out-of-distribution (OOD) detection based on deep learning: a review. *Electronics* 2022;**11**:3500. https://doi.org/10.3390/electronics11213500
16. Lu S, Wang Y, Sheng L. *et al*. Recent advances in OOD detection: problems and approaches. 2024. https://doi.org/10.48550/arXiv.2409.11884
17. Salehi M, Mirzaei H, Hendrycks D. *et al*. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: solutions and future challenges. 2022. https://doi.org/10.48550/arXiv.2110.14051
18. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R. *et al*. A unifying view on dataset shift in classification. *Pattern Recognition* 2012;**45**: 521–30. https://doi.org/10.1016/j.patcog.2011.06.019
19. Wei H, Xie R, Cheng H. *et al*. Mitigating neural network overconfidence with logit normalization. In: *Proceedings of Machine Learning Research*, Vol. **162**, 2022.
20. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *33rd International Conference on Machine Learning, ICML 2016*, Vol. **3**, 2016.
21. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems*, Vol. **2017**, December, 2017.
22. Liu W, Wang X, Owens JD. *et al*. Energy-based out-of-distribution detection. In: *Advances in Neural Information Processing Systems*, Vol. **2020**, December, 2020.
23. Sun Y, Ming Y, Zhu X. *et al*. Out-of-distribution detection with deep nearest neighbors. In: *Proceedings of Machine Learning Research*, Vol. **162**, 2022.
24. Charpentier B, Zügner D, Günnemann S. Posterior network: uncertainty estimation without OOD samples via density-based pseudo-counts. In: *Advances in Neural Information Processing Systems*, Vol. **2020**, December, 2020.
25. Guo C, Pleiss G, Yu S. *et al*. On calibration of modern neural networks. In: *34th International Conference on Machine Learning, ICML 2017*, Vol. **3**, 2017.
26. Srivastava N, Hinton G, Krizhevsky A. *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
27. Adams TS, Schupp JC, Poli S. *et al*. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020;**6**. https://doi.org/10.1126/sciadv.aba1983
28. Bengs V, Hüllermeier E, Waegeman W. Pitfalls of epistemic uncertainty quantification through loss minimisation. In: *Advances in Neural Information Processing Systems*, Vol. **35**, 2022.
29. Juergens M, Meinert N, Bengs V. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *Proceedings of the 41st International Conference on Machine Learning, Vol. 235 of Proceedings of Machine Learning Research*. In: Salakhutdinov R, Kolter Z, Heller K. *et al*. (eds), pp. 22624–42. PMLR 21–27 July, 2024.
30. Abdelaal T, Michielsen L, Cats D. *et al*. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:194. https://doi.org/10.1186/s13059-019-1795-z
31. Köhler ND, Büttner M, Theis FJ. Deep learning does not outperform classical machine learning for cell-type annotation. *bioRiv* 2021. https://doi.org/10.1101/653907
32. Huang Y, Zhang P. Evaluation of machine learning approaches for cell-type identification from single-cell transcriptomics data. *Brief Bioinform* 2021;**22**. https://doi.org/10.1093/bib/bbab217
33. Fischer F, Fischer DS, Mukhin R. *et al*. Sctab: scaling cross-tissue single-cell annotation models. *Nat Commun* 2024;**15**:6611. https://doi.org/10.1038/s41467-024-51059-5
34. Mereu E, Lafzi A, Moutinho C. *et al*. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 2020;**38**:747–55. https://doi.org/10.1038/s41587-020-0469-4
35. Ziegenhain C, Vieth B, Parekh S. *et al*. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;**65**:631–643.e4. https://doi.org/10.1016/j.molcel.2017.01.023

Briefings in Bioinformatics, 2025, **26(3)**, bbaf239
https://doi.org/10.1093/bib/bbaf239
Problem Solving Protocol