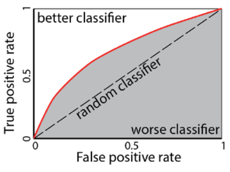
Table 3. Evaluation metrics used in surveyed DL algorithms

Evaluation Method	Equations	Explanation
Pseudobulk RNA-seq		Average of normalized (log2-transformed) scRNA-seq counts across cells is calculated and then correlation coefficient between the pseudobulk and the actual bulk RNA-seq profile of the same cell type is evaluated.
Mean squared error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$	MSE assesses the quality of a predictor, or an estimator, from a collection of observed data x, with $\hat{x}$ being the predicted values.
Pearson correlation	$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	where cov() is the covariance, σ_X and σ_Y are the standard deviation of *X* and *Y*, respectively.
Spearman correlation	$\rho_s = \rho_{r_X,r_Y} = \frac{cov(r_X, r_Y)}{\sigma_{r_X}\sigma_{r_Y}}$	The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables, where r_X is the rank of X.
Entropy of accuracy, H_{acc} [20]	$H_{acc} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_i} p_i(x_j)\log\left(p_i(x_j)\right)$	Measures the diversity of the ground-truth labels within each predicted cluster group. $p_i(x_j)$ (or $q_i(x_j)$) are the proportions of cells in the *j*th ground-truth cluster (or predicted cluster) relative to the total number of cells in the *i*th predicted cluster (or ground-truth clusters), respectively.
Entropy of purity, H_{pur} [20]	$H_{pur} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} q_i(x_j)\log\left(q_i(x_j)\right)$	Measures the diversity of the predicted cluster labels within each ground-truth group
Entropy of mixing [31]	$E = \sum_{i=1}^{C} p_i \log(p_i)$	This metric evaluates the mixing of cells from different batches in the neighborhood of each cell. C is the number of batches, and p_i is the proportion of cells from batch i among *N* nearest cells.
Mutual Information (MI) [157]	$MI(U,V) = \sum_{i=1}^{\vert U\vert}\sum_{j=1}^{\vert V\vert} P_{UV}(i,j)\log\left(\frac{P_{UV}(i,j)}{P_U(i)P_V(j)}\right)$	where $P_U(i) = \frac{\vert U_i\vert}{N}$ and $P_V(j) = \frac{\vert V_j\vert}{N}$. Also, define the joint distribution probability is $P_{UV}(i,j) = \frac{\vert U_i \cap V_j\vert}{N}$. The MI is a measure of mutual dependency between two cluster assignments *U* and *V*.
Normalized Mutual Information (NMI) [158]	$NMI(U,V) = \frac{2 \times MI(U,V)}{[H(U) + H(V)]}$	where $H(U) = \sum P_U(i)\log\left(P_U(i)\right), H(V) = \sum P_V(i)\log\left(P_V(i)\right)$. The NMI is a normalization of the MI score between 0 and 1.
Kullback–Leibler (KL) divergence [159]	$D_{KL}(P\Vert Q) = \sum_{x\in\chi} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$	where discrete probability distributions *P* and *Q* are defined on the same probability space χ. This relative entropy is the measure for directed divergence between two distributions.
Jaccard Index	$J(U,V) = \frac{\lfloor U \cap V\rfloor}{\lfloor U \cup V\rfloor}$	0 ≤ *J*(*U*,*V*) ≤ 1. J = 1 if clusters *U* and *V* are the same. If *U* are *V* are empty, J is defined as 1.
Fowlkes-Mallows Index for two clustering algorithms (FM)	$FM = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}$	TP as the number of pairs of points that are present in the same cluster in both *U* and *V*; FP as the number of pairs of points that are present in the same cluster in *U* but not in *V;* FN as the number of pairs of points that are present in the same cluster in V but not in *U;* and TN as the number of pairs of points that are in different clusters in both *U* and *V*.
Rand index (RI)	$RI = (a+b)/\binom{n}{2}$	Measure of constancy between two clustering outcomes, where *a* (or *b*) is the count of number of pairs of cells in one cluster (or different clusters) from one clustering algorithm but also fall in the same cluster (or different clusters) from the other clustering algorithm.
Adjusted Rand index (ARI) [160]	$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$	ARI is a corrected-for-chance version of *RI,* where *E*[*RI*] is the expected Rand Index.
Silhouette index	$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$	where *a*(*i*) is the average dissimilarity of *i*th cell to all other cells in the same cluster, and *b*(*i*) is the average dissimilarity of *i*th cell to all cells in the closest cluster. The range of *s*(*i*) is [−1,1], with 1 to be well-clustered and -1 to be completely misclassified.
Maximum Mean Discrepancy (MMD) [58]	$MMD(F,p,q) = \sup_{f\in F}\Vert \mu_p - \mu_q\Vert_f$	MMD is a non-parametric distance between distributions based on the reproducing kernel Hilbert space, or, a distance-based measure between two distribution p and q based on the mean embeddings μ_p and μ_q in a reproducing kernel Hilbert space F.
k-Nearest neighbor batch-effect test (kBET) [161]	$a_n^k = \sum_{l=1}^{L}\frac{(N_{nl}^k - k \cdot f_l)^2}{k \cdot f_l} \sim X_{L-1}^2$	Given a dataset of N cells from L batches with N_l denoting the number of cells in batch l, N_{nl}^k is the number of cells from batch l in the *k*-nearest neighbors of cell n, f_l is the global fraction of cells in batch l, or $f_l = \frac{N_l}{N}$, and X_{L-1}^2 denotes the X^2 distribution with $L-1$ degrees of freedom. It uses a X^2-based test for random neighborhoods of fixed size to determine the significance (“well mixed”).
Local Inverse Simpson’s Index (LISI) [33]	$\frac{1}{\lambda(n)} = \frac{1}{\Sigma_{l=1}^{L}\left(p(l)\right)^2}$	This is the inverse Simpson’s Index in the *k*-nearest neighbors of cell n for all batches, where $p(l)$ denotes the proportion of batch l in the *k*-nearest neighbors. The score reports the effective number of batches in the *k*-nearest neighbors of cell n.
Homogeneity	$HS = 1 - \frac{H(P(U \mid V))}{H\left(P(U)\right)}$	where *H*() is the entropy, and *U* is the ground-truth assignment and *V* is the predicted assignment. The *HS* range from 0 to 1, where 1 indicates perfectly homogeneous labelling.
Completeness	$CS = 1 - \frac{H\left(P(V \mid U)\right)}{H\left(P(V)\right)}$	Its values range from 0 to 1, where 1 indicates all member from a ground-truth label are assigned to a single cluster.
V-Measure [162]	$V_\beta = \frac{(1+\beta)HS \times CS}{\beta HC + CS}$	where β indicates the weight of *HS*. V-Measure is symmetric, *i.e.* switching the true and predicted cluster labels does not change V-Measure.
Precision, recall	$Precision = \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN}$	TP: true positive, FP: false positive, FN, false negative.
Accuracy	$Accuracy = \frac{TP+TN}{N}$	N: all samples tested, TN: true negative
F_1-score	$F_1 = \frac{2\ Precision \cdot Recall}{Precision + Recall}$	A harmonic mean of precision and recall. It can be extended to F_β where β is a weight between precision and recall (similar to V-measure).
AUC, RUROC		Area Under Curve (grey area). Receiver operating characteristic (ROC) curve (red line). The similar measure can be performed on Precision-Recall curve (PRC), or AUPRC. Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model (mostly for imbalanced dataset).