Supplementary Information

Calibration and generalizability of probabilistic models on low-data chemical datasets with Dionysus

Gary Tom,[1,2,3] Riley J. Hickman,[1,2,3] Aniket Zinzuwadia,[4] Afshan Mohajeri,[5]
Benjamin Sanchez-Lengeling,[6] Alán Aspuru-Guzik[1,2,3,7,8,9,*]

[1] *Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, ON, Canada*
[2] *Department of Computer Science, University of Toronto, Toronto, ON, Canada*
[3] *Vector Institute for Artificial Intelligence, Toronto, ON, Canada*
[4] *Harvard Medical School, Harvard University, Boston, MA, USA*
[5] *Department of Chemistry, Shiraz University, Shiraz, Iran*
[6] *Google Research, Brain Team*
[7] *Department of Chemical Engineering & Applied Chemistry, University of Toronto, Toronto, ON, Canada*
[8] *Department of Materials Science & Engineering, University of Toronto, Toronto, ON, Canada*
[9] *Lebovic Fellow, Canadian Institute for Advanced Research, Toronto, ON, Canada*
[*] *alan@aspuru.com*

S.1. SUPPLEMENTARY INFORMATION

A. Crippen values in Mordred features

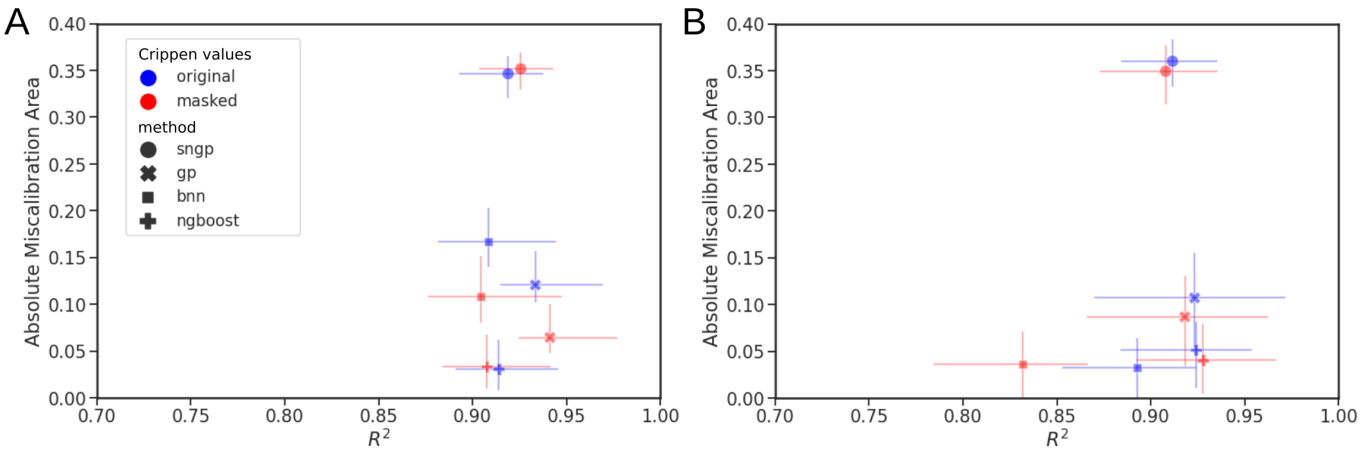

Figure S1: **Comparison of model performance on Mordred features with and without Crippen values**. The prediction and calibration tests are repeated for each model on the datasets related to solubility: A) Delaney and B) Freesolv. The featurizations are the original Mordred features presented in the main text, as well as masked Mordred features, which have the Crippen values masked, in order to remove any advantageous correlation with the targets.

B. Graph features

Node features	Categories
Atomic number	one-hot encoding from set of heavy atoms in dataset
Chirality	unspecified, CW, CCW, *UNK*
Atom degree	0, . . . , 10, *UNK*
Formal charge	-5, . . . , 5, *UNK*
Number of hydrogens	0, . . . , 8, *UNK*
Number of radical electrons	0, . . . , 4, *UNK*
Hybridization	sp, sp2, sp3, sp3d, sp3d2, *UNK*
Is aromatic	True/False
Part of ring	True/False
Edge features	**Categories**
Bond type	single, double, triple, aromatic, *UNK*
Stereo configuration	none, *Z*, *E*, *cis*, *trans*, any
Is conjugated	True/False

Table S5: Features for the vertex and edge features of the molecular graph. All cateogries are one-hot encoded and stacked to give a singular bit vector. *UNK* stands for "unknown", and is a catch-all category.

C. Performance and Calibration Metrics

BioHL	MFP	Mordred	Graph-based
SNGP	$-0.107^{+0.114}_{-0.432}$	$-0.136^{+0.021}_{-0.466}$	$-0.136^{+0.092}_{-0.437}$
GP	$0.383^{+0.308}_{-0.557}$	$0.817^{+0.131}_{-0.242}$	$0.750^{+0.160}_{-0.279}$
BNN	$0.011^{+0.197}_{-0.436}$	$-0.103^{+0.594}_{-1.378}$	$-0.066^{+0.296}_{-0.818}$
NGBoost	$0.320^{+0.373}_{-0.786}$	$0.843^{+0.096}_{-0.181}$	$0.803^{+0.142}_{-0.228}$
GNNGP			$-0.129^{+0.146}_{-0.458}$
Freesolv	MFP	Mordred	Graph-based
SNGP	$0.738^{+0.093}_{-0.081}$	$0.912^{+0.029}_{-0.026}$	$0.891^{+0.056}_{-0.039}$
GP	$0.716^{+0.130}_{-0.138}$	$0.924^{+0.031}_{-0.050}$	$0.875^{+0.039}_{-0.051}$
BNN	$0.601^{+0.097}_{-0.086}$	$0.892^{+0.030}_{-0.040}$	$0.845^{+0.044}_{-0.071}$
NGBoost	$0.556^{+0.131}_{-0.128}$	$0.925^{+0.027}_{-0.044}$	$0.887^{+0.038}_{-0.053}$
GNNGP			$0.903^{+0.039}_{-0.039}$
Delaney	MFP	Mordred	Graph-based
SNGP	$0.687^{+0.066}_{-0.073}$	$0.918^{+0.021}_{-0.025}$	$0.904^{+0.023}_{-0.027}$
GP	$0.724^{+0.052}_{-0.053}$	$0.934^{+0.016}_{-0.018}$	$0.897^{+0.025}_{-0.030}$
BNN	$0.687^{+0.055}_{-0.061}$	$0.908^{+0.019}_{-0.023}$	$0.905^{+0.029}_{-0.030}$
NGBoost	$0.486^{+0.070}_{-0.076}$	$0.915^{+0.019}_{-0.025}$	$0.897^{+0.029}_{-0.030}$
GNNGP			$0.911^{+0.027}_{-0.022}$

BioHL	MFP	Mordred	Graph-based
SNGP	$0.070^{+0.052}_{-0.031}$	$0.071^{+0.055}_{-0.032}$	$0.075^{+0.049}_{-0.035}$
GP	$0.061^{+0.060}_{-0.033}$	$0.141^{+0.098}_{-0.074}$	$0.200^{+0.097}_{-0.098}$
BNN	$0.363^{+0.069}_{-0.089}$	$0.103^{+0.088}_{-0.065}$	$0.217^{+0.088}_{-0.096}$
NGBoost	$0.109^{+0.098}_{-0.058}$	$0.109^{+0.057}_{-0.046}$	$0.086^{+0.088}_{-0.053}$
GNNGP			$0.211^{+0.101}_{-0.101}$
Freesolv	MFP	Mordred	Graph-based
SNGP	$0.268^{+0.033}_{-0.033}$	$0.359^{+0.024}_{-0.026}$	$0.345^{+0.028}_{-0.031}$
GP	$0.147^{+0.040}_{-0.046}$	$0.106^{+0.048}_{-0.048}$	$0.271^{+0.035}_{-0.037}$
BNN	$0.243^{+0.051}_{-0.051}$	$0.033^{+0.030}_{-0.017}$	$0.147^{+0.039}_{-0.038}$
NGBoost	$0.032^{+0.026}_{-0.016}$	$0.052^{+0.028}_{-0.023}$	$0.035^{+0.031}_{-0.018}$
GNNGP			$0.086^{+0.047}_{-0.044}$
Delaney	MFP	Mordred	Graph-based
SNGP	$0.202^{+0.029}_{-0.030}$	$0.347^{+0.017}_{-0.019}$	$0.326^{+0.018}_{-0.020}$
GP	$0.080^{+0.037}_{-0.032}$	$0.120^{+0.032}_{-0.036}$	$0.110^{+0.038}_{-0.038}$
BNN	$0.223^{+0.042}_{-0.039}$	$0.166^{+0.036}_{-0.040}$	$0.044^{+0.022}_{-0.019}$
NGBoost	$0.070^{+0.038}_{-0.036}$	$0.031^{+0.032}_{-0.017}$	$0.056^{+0.036}_{-0.028}$
GNNGP			$0.055^{+0.020}_{-0.019}$

Table S6: **Performance and calibration results on regression datasets**. (*left*) R^2 metric and (*right*) AMA for each feature and model pair. Graph inputs were used for GNNGP, while graph embeddings were used for all other models. The 95% confidence interval is reported.

BACE	MFP	Mordred	Graph-based
SNGP	$0.890^{+0.036}_{-0.037}$	$0.879^{+0.039}_{-0.041}$	$0.873^{+0.039}_{-0.040}$
GP	$0.917^{+0.029}_{-0.031}$	$0.915^{+0.028}_{-0.032}$	$0.865^{+0.041}_{-0.041}$
BNN	$0.918^{+0.028}_{-0.029}$	$0.893^{+0.034}_{-0.037}$	$0.869^{+0.040}_{-0.041}$
NGBoost	$0.895^{+0.031}_{-0.034}$	$0.890^{+0.032}_{-0.039}$	$0.857^{+0.039}_{-0.046}$
GNNGP			$0.845^{+0.041}_{-0.046}$
RBioDeg	MFP	Mordred	Graph-based
SNGP	$0.783^{+0.053}_{-0.055}$	$0.832^{+0.042}_{-0.048}$	$0.826^{+0.042}_{-0.048}$
GP	$0.837^{+0.040}_{-0.048}$	$0.858^{+0.039}_{-0.042}$	$0.835^{+0.043}_{-0.044}$
BNN	$0.826^{+0.042}_{-0.045}$	$0.852^{+0.038}_{-0.043}$	$0.833^{+0.040}_{-0.044}$
NGBoost	$0.791^{+0.050}_{-0.053}$	$0.846^{+0.042}_{-0.042}$	$0.829^{+0.046}_{-0.049}$
GNNGP			$0.840^{+0.044}_{-0.048}$
BBBP	MFP	Mordred	Graph-based
SNGP	$0.882^{+0.041}_{-0.050}$	$0.902^{+0.036}_{-0.039}$	$0.870^{+0.047}_{-0.053}$
GP	$0.910^{+0.035}_{-0.040}$	$0.922^{+0.0.35}_{-0.037}$	$0.894^{+0.038}_{-0.045}$
BNN	$0.886^{+0.043}_{-0.046}$	$0.900^{+0.043}_{-0.049}$	$0.883^{+0.043}_{-0.050}$
NGBoost	$0.842^{+0.051}_{-0.053}$	$0.896^{+0.042}_{-0.048}$	$0.834^{+0.050}_{-0.053}$
GNNGP			$0.879^{+0.043}_{-0.051}$

BACE	MFP	Mordred	Graph-based
SNGP	$0.184^{+0.075}_{-0.070}$	$0.191^{+0.068}_{-0.065}$	$0.155^{+0.051}_{-0.043}$
GP	$0.093^{+0.039}_{-0.032}$	$0.090^{+0.041}_{-0.035}$	$0.155^{+0.062}_{-0.052}$
BNN	$0.141^{+0.059}_{-0.049}$	$0.171^{+0.040}_{-0.040}$	$0.146^{+0.060}_{-0.056}$
NGBoost	$0.095^{+0.038}_{-0.033}$	$0.076^{+0.037}_{-0.032}$	$0.171^{+0.089}_{-0.071}$
GNNGP			$0.127^{+0.050}_{-0.045}$
RBioDeg	MFP	Mordred	Graph-based
SNGP	$0.147^{+0.056}_{-0.048}$	$0.249^{+0.078}_{-0.074}$	$0.244^{+0.086}_{-0.080}$
GP	$0.075^{+0.037}_{-0.030}$	$0.096^{+0.048}_{-0.038}$	$0.132^{+0.051}_{-0.046}$
BNN	$0.104^{+0.043}_{-0.040}$	$0.097^{+0.046}_{-0.039}$	$0.142^{+0.052}_{-0.050}$
NGBoost	$0.081^{+0.038}_{-0.034}$	$0.150^{+0.055}_{-0.055}$	$0.125^{+0.053}_{-0.048}$
GNNGP			$0.223^{+0.063}_{-0.058}$
BBBP	MFP	Mordred	Graph-based
SNGP	$0.160^{+0.064}_{-0.059}$	$0.263^{+0.086}_{-0.074}$	$0.290^{+0.119}_{-0.115}$
GP	$0.127^{+0.061}_{-0.046}$	$0.112^{+0.051}_{-0.043}$	$0.165^{+0.027}_{-0.026}$
BNN	$0.230^{+0.108}_{-0.095}$	$0.118^{+0.067}_{-0.053}$	$0.282^{+0.110}_{-0.100}$
NGBoost	$0.168^{+0.059}_{-0.058}$	$0.188^{+0.089}_{-0.083}$	$0.368^{+0.118}_{-0.150}$
GNNGP			$0.246^{+0.073}_{-0.063}$

Table S7: **Performance and calibration results on binary classification datasets**. (*left*) AUROC metric and (*right*) ECE for each feature and model pair. Graph inputs were used for GNNGP, while graph embeddings were used for all other models. The 95% confidence interval is reported.

D. Bayesian Optimization Traces

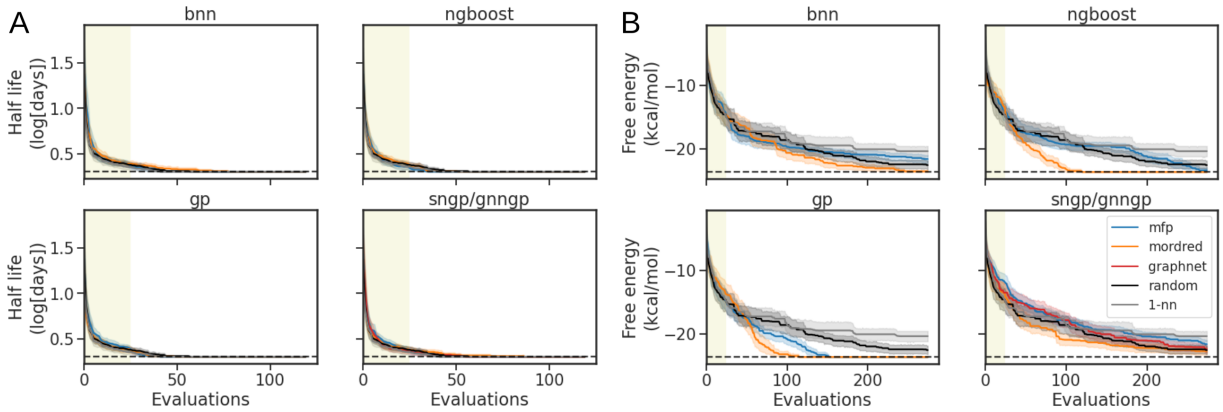

Figure S2: **BO traces for regression datasets**. Minimization traces for A) BioHL dataset, and B) Freesolv dataset. Traces show best molecule fitness as a function of evaluations, with 95% confidence interval from 30 independent runs. The shaded area are the 95% confidence intervals. The BO experiments start with randomly sampled 5% of the dataset (minimum of 25 molecules), indicated by the shaded region at the start of the optimization. The optimal molecule in the dataset is shown by the horizontal dashed lines. Random search and 1-nearest-neighbour traces are shown as baselines.

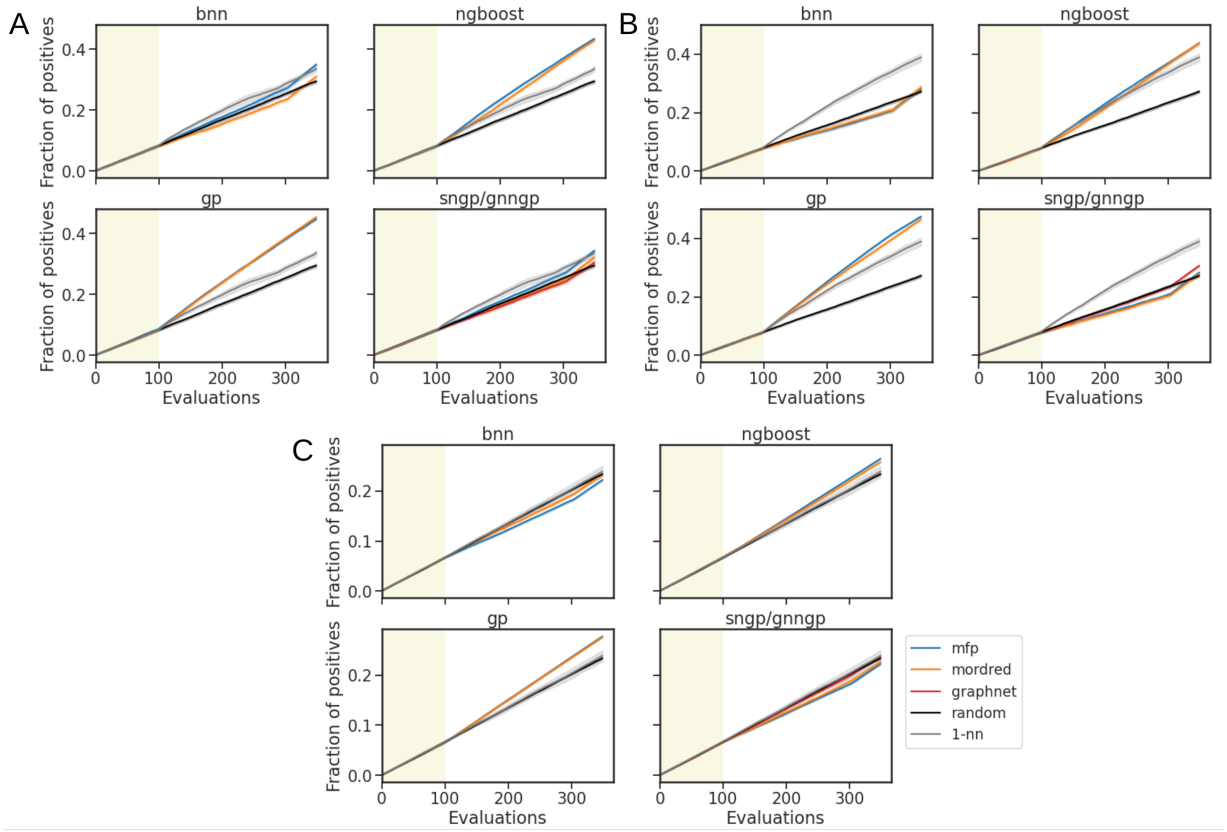

Figure S3: **BO traces for binary classification datasets**. Maximization traces for A) BACE dataset, B) RBioDeg dataset, and C) BBBP dataset. Traces show the fraction of positive hits as a function of evaluations, with 95% confidence interval from 30 independent runs. The BO experiments start with randomly sampled 10% of the dataset (maximum of 100 molecules), indicated by the shaded region at the start of the optimization. Random search and 1-nearest-neighbour traces are shown as baselines.

E. Generalizability

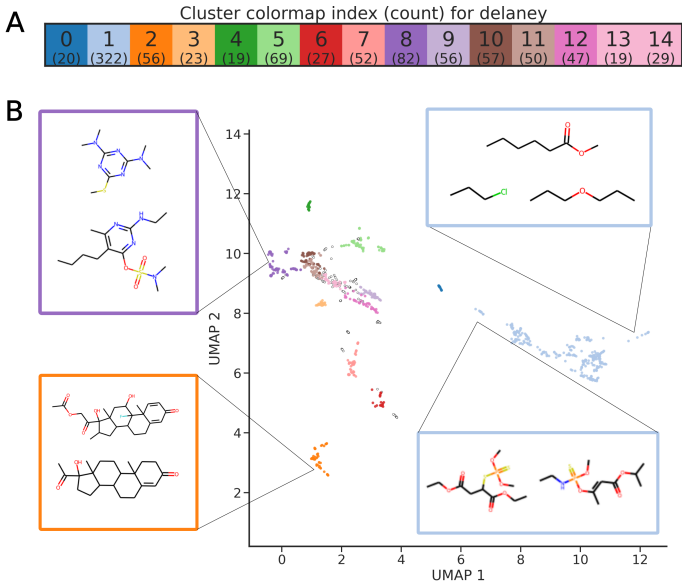

Figure S4: **Clusters generated for cluster splits on the Delaney dataset**. A) Clusters identified by HDBScan algorithm,[1] coloured and labelled, with the number of molecules per cluster listed. B) Visualization of UMAP[2] reduced chemical space, with samples of molecules from clusters shown. Similar clusters have similar structures. Molecules further in chemical space have more structural differences.

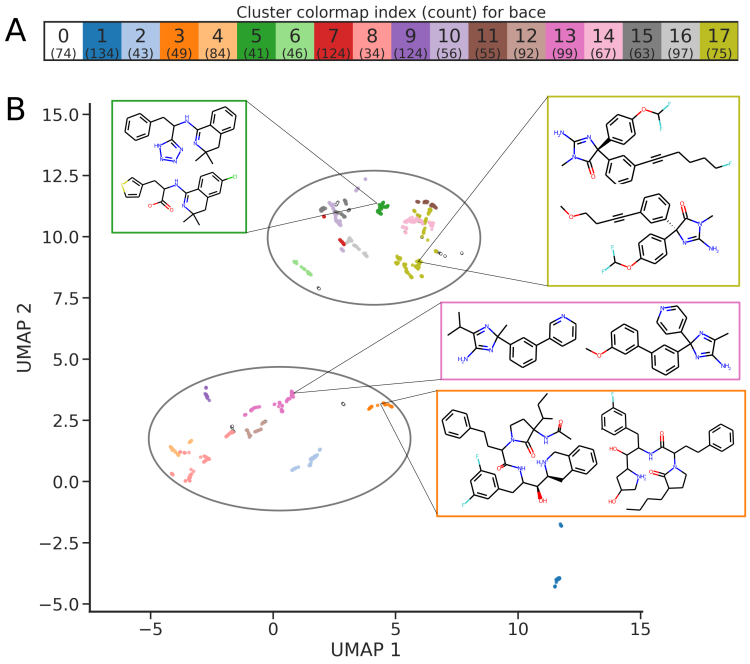

Figure S5: **Clusters generated for cluster splits on the BACE dataset**. A) Clusters identified by HDBScan algorithm, coloured and labelled, with the number of molecules per cluster listed. B) Visualization of UMAP reduced chemical space, with samples of molecules from clusters shown. The BACE chemical space is predominantly organized into two superclusters, indicated by the circles: the lower cluster (UMAP 2 < 6.0), and the upper cluster (UMAP 2 > 6.0).

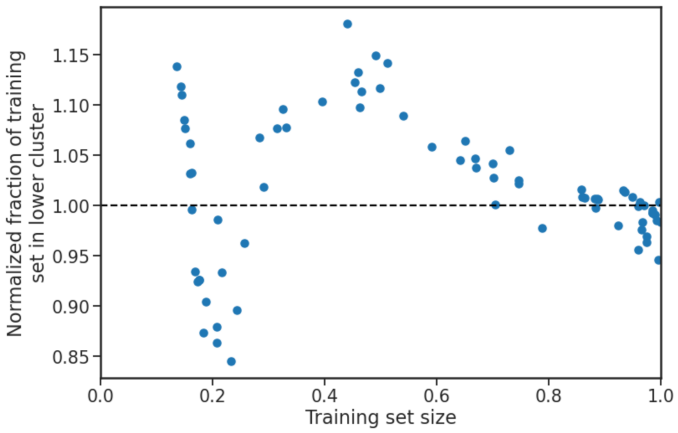

Figure S6: **Normalized fraction of training set molecules in the lower cluster for BACE dataset**. The fraction of training set molecules in the lower cluster, as observed in the BACE cluster feature space (Figure S5), normalized by the fraction of test set molecules in the lower cluster. When the normalized fraction is 1.0, the training set has the same fraction of molecules in each supercluster as the held out test set, which occurs near the maximal size of the training set. At around 50% training set size, most molecules in the training set are in the lower cluster, corresponding to the dip in performance for all models on the BACE ablated cluster splits.

[1] Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pacific-Asia conference on knowledge discovery and data mining; 2013. p. 160–172.
[2] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.