Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2024

Supporting Information

Composite Machine Learning Strategy for Natural Products Taxonomical Classification and Structural Insights

Qisong Xu[a], Alan K.X. Tan[a], Liangfeng Guo[a], Yee Hwee Lim[a,b], Dillon W. P. Tay[a]*, Shi Jun Ang[a,c,d]*

[a] *Institute of Sustainability for Chemicals, Energy and Environment (ISCE^2), Agency for Science, Technology and Research (A*STAR), 1 Pesek Road, Jurong Island, Singapore 627833, Republic of Singapore*

[b] *Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Republic of Singapore*

[c] *Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore*

[d] *Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Republic of Singapore*

*Corresponding Authors: Dillon W. P. TAY (dillon_tay@isce2.a-star.edu.sg) and Shi Jun ANG (ang_shi_jun@ihpc.a-star.edu.sg)

Table of Contents

1. Molecular structures of natural products

A small subset of NPs belonging to six or more kingdoms (structures shown in **Figure S1**) contained molecules such as long-chain fatty acids and steroids that are common across various kingdoms.[1, 2]

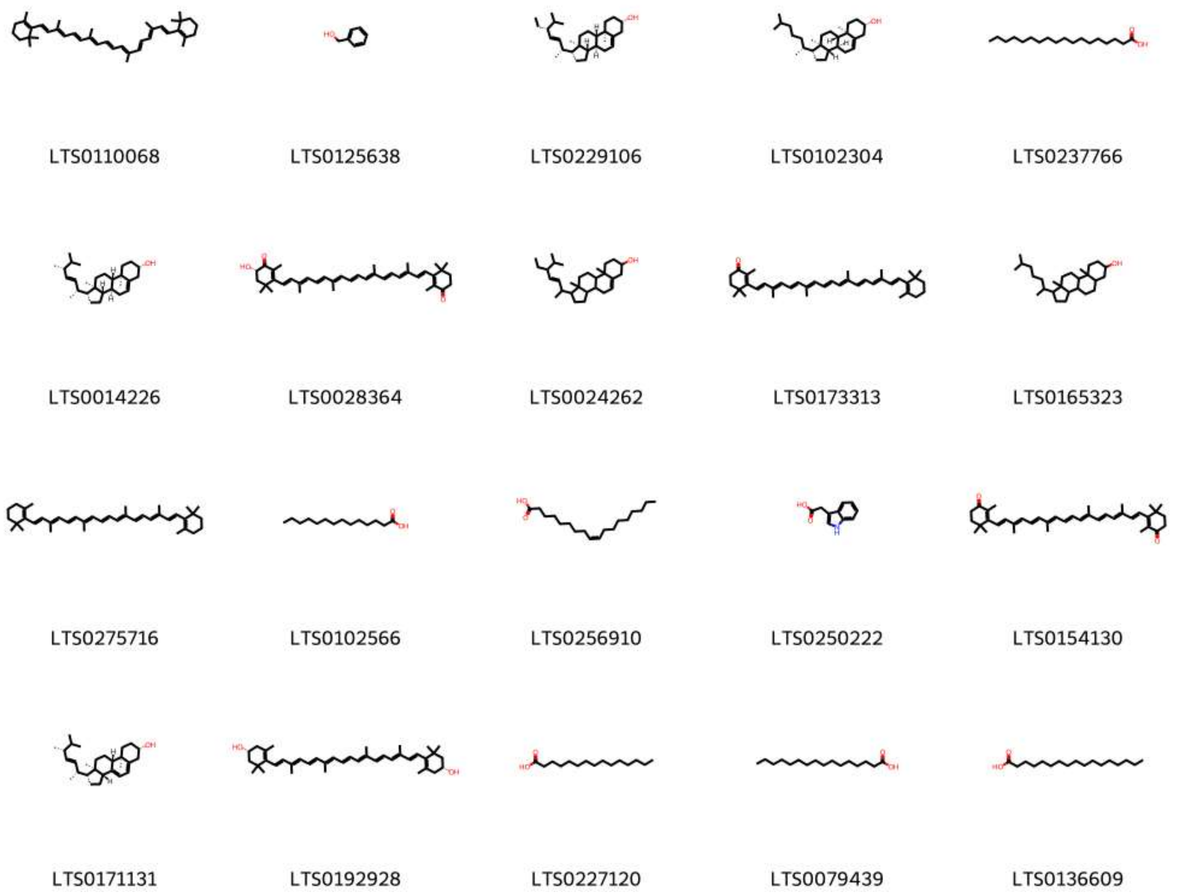

Figure S1. Molecular structures and LOTUS IDs of twenty natural products with origins from six or more kingdoms.

However, not all isomeric SMILES stemming from the same non-isomeric SMILES may share the same kingdom label. For example, in **Figure S2** a cucurbitane-type triterpenoid has eight different isomeric SMILES that stem from the same non-isomeric SMILES, all from the plant kingdom.[3] However, in **Figure S3**, while maltotriose also has eight different isomeric SMILES from the same non-isomeric SMILES, its constituent isomeric SMILES are divided across different kingdoms.[4]

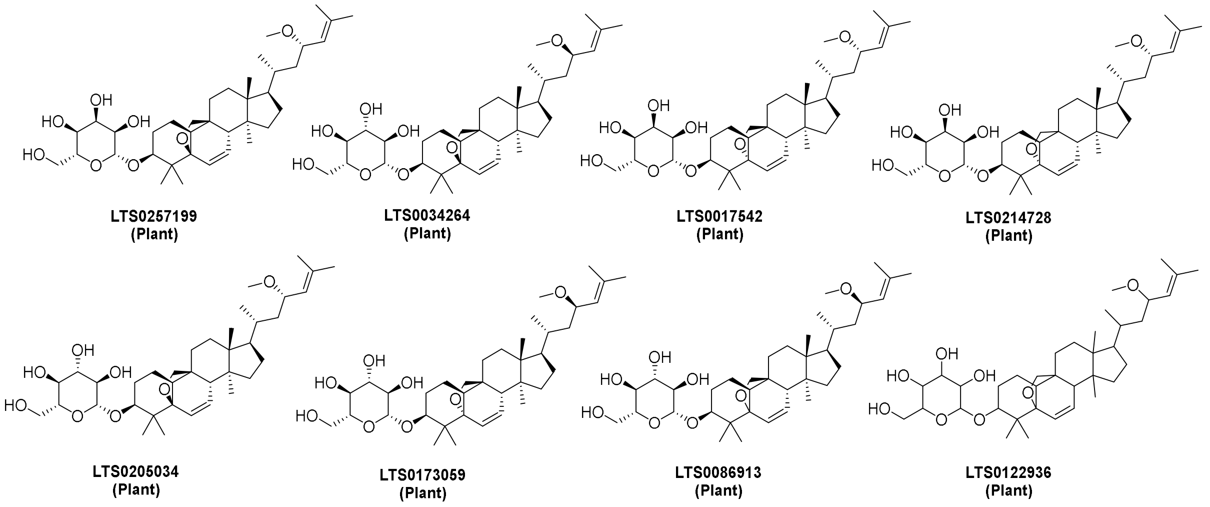

Figure S2. Molecular structures and LOTUS ID of cucurbitane-type triterpenoid stereoisomers, all of which belong to the plant kingdom.

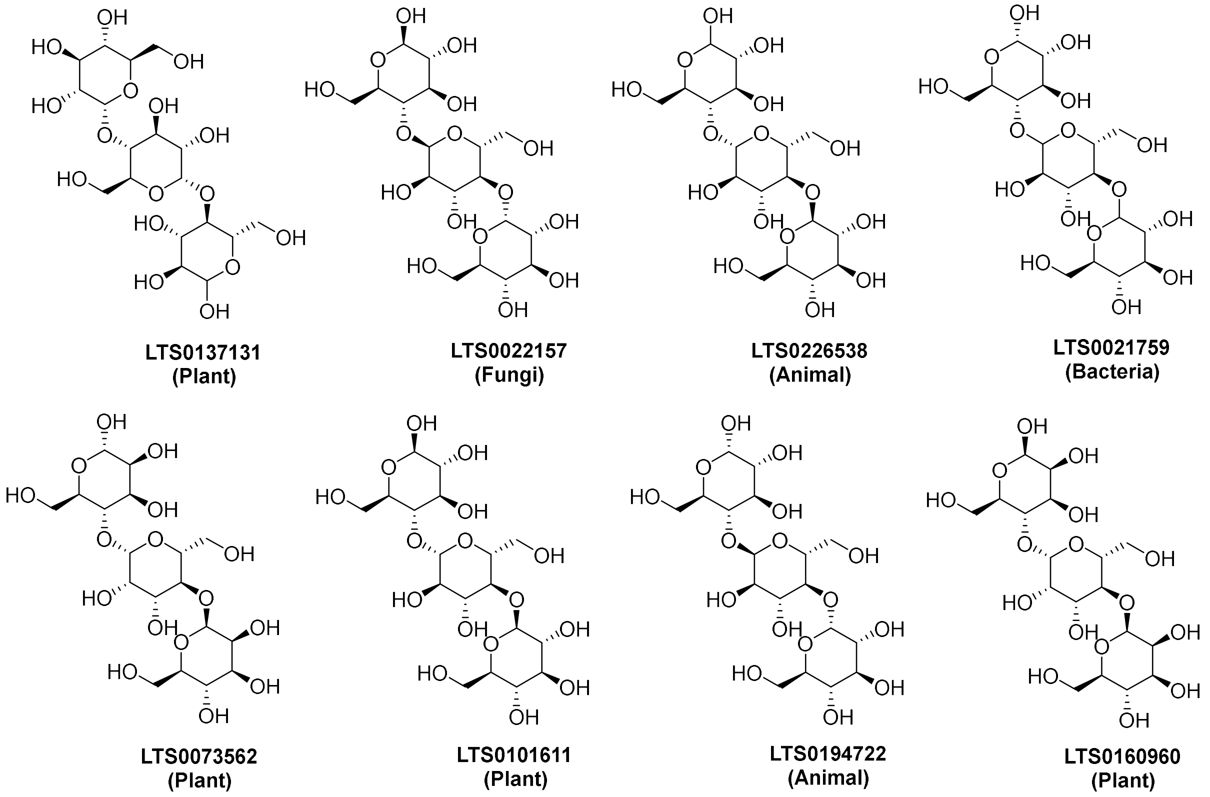

Figure S3. Molecular structures and LOTUS ID of maltotriose stereoisomers, which belong to four different kingdoms (plant, fungi, animal, and bacteria).

2. Multiclass classification performance

Table S1. Comparison of Balanced Accuracy (BA) and MCC from various molecular fingerprints and machine learning algorithms to predict five different kingdoms (Plant, Bacteria, Fungi, Animal, and Chromista). Values reported are training results from stratified 5-fold cross validation.

Algorithm	MAP4		MPN		last_FFN	
	BA	MCC	BA	MCC	BA	MCC
GCNN	-	-	-	-	94.3 ± 1.1	95.1 ± 0.7
NB	20.0 ± 0.0	0.0 ± 0.0	30.9 ± 0.5	39.1 ± 0.9	67.5 ± 0.8	83.9 ± 0.4
QDA	54.4 ± 0.2	22.7 ± 0.3	63.1 ± 0.1	79.7 ± 0.1	96.6 ± 0.1	96.5 ± 0.0
KNN	55.9 ± 5.0	61.6 ± 4.7	83.0 ± 0.7	87.4 ± 0.5	94.0 ± 0.1	96.9 ± 0.0
RF	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.7 ± 0.5	99.8 ± 0.3
LGBM	84.3 ± 0.1	81.7 ± 0.1	98.6 ± 0.0	98.3 ± 0.1	100.0 ± 0.0	100.0 ± 0.0
XGB	93.8 ± 0.1	92.8 ± 0.1	99.7 ± 0.0	99.6 ± 0.0	99.9 ± 0.0	99.9 ± 0.0
SVM	93.2 ± 0.3	95.9 ± 0.0	92.9 ± 0.1	94.0 ± 0.0	97.6 ± 0.1	98.2 ± 0.0

BA = Balanced Accuracy. MCC = Matthews Correlation Coefficient. GCNN = Graph Convolutional Neural Network. NB = Gaussian Naïve Bayesian. QDA = Quadratic Discriminant Analysis. RF = Random Forest. SVM = Support Vector Machine. LGBM = Light Gradient-Boosting Machine. XGB = Extreme Gradient Boosting. Error of 1 standard deviation shown.

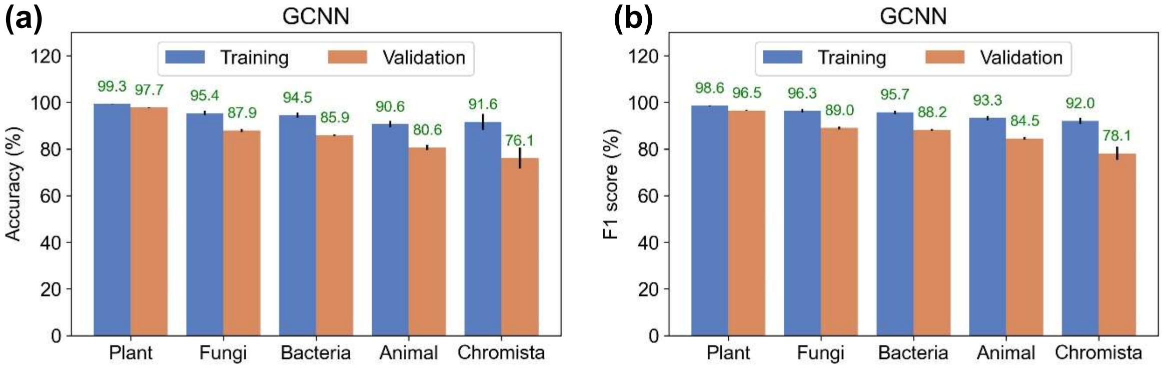

Figure S4. Stratified 5-fold cross validated performance of GCNN model across 5 different kingdoms in terms of (a) accuracy and (b) F1 score.

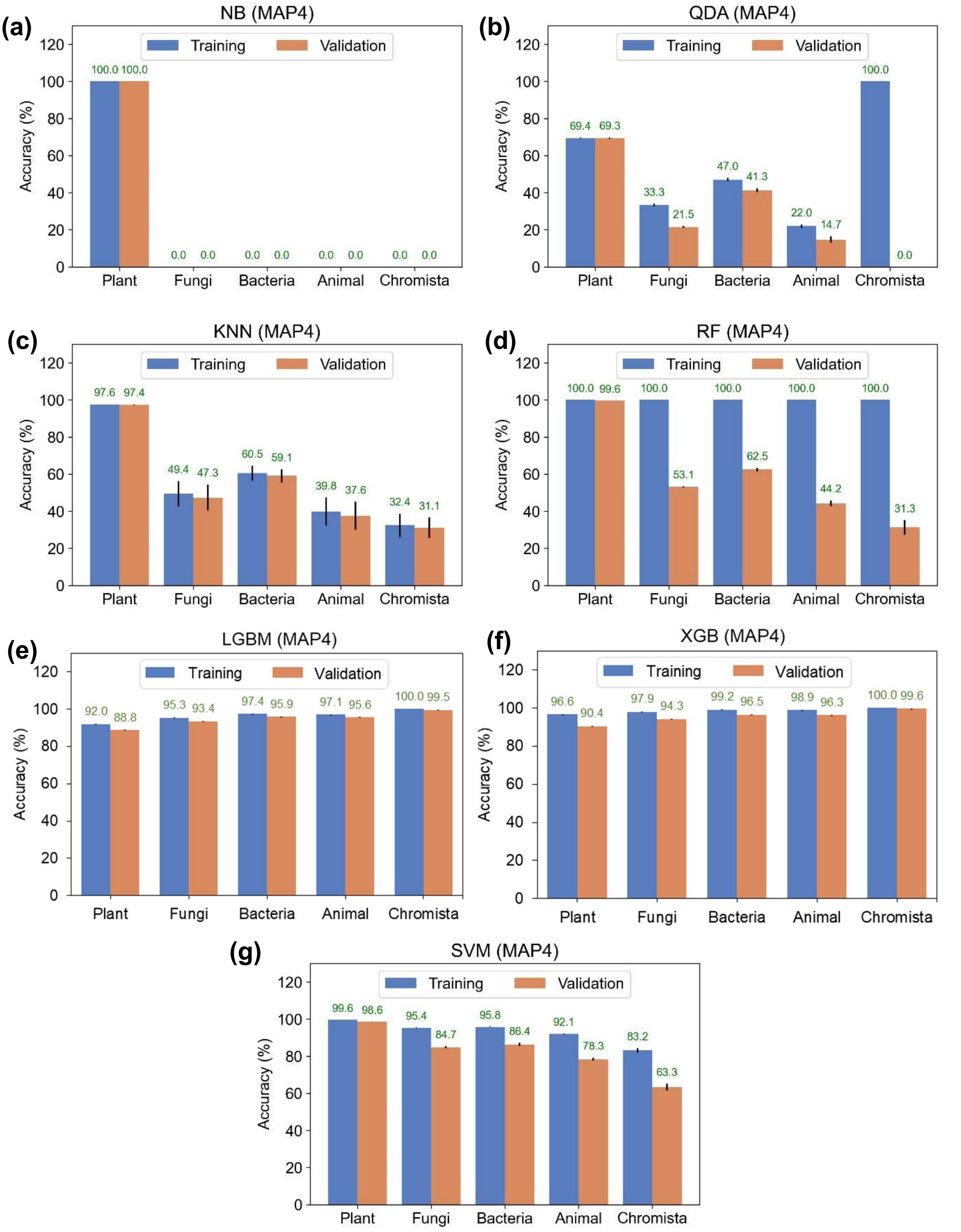

Figure S5. Stratified 5-fold cross validated classification accuracies across 5 different kingdoms using MAP4 fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

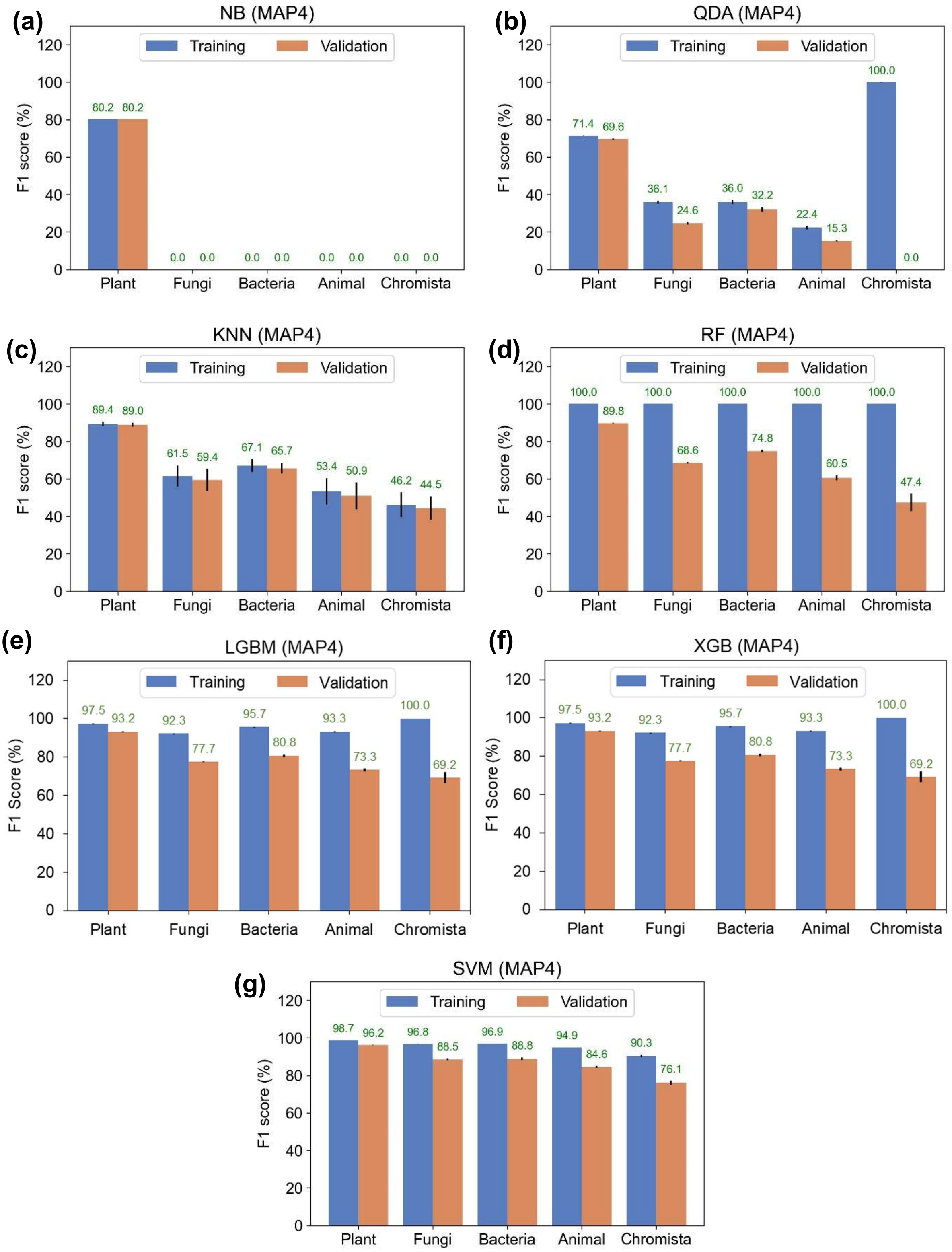

Figure S6. Stratified 5-fold cross validated F1 scores across 5 different kingdoms using MAP4 fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

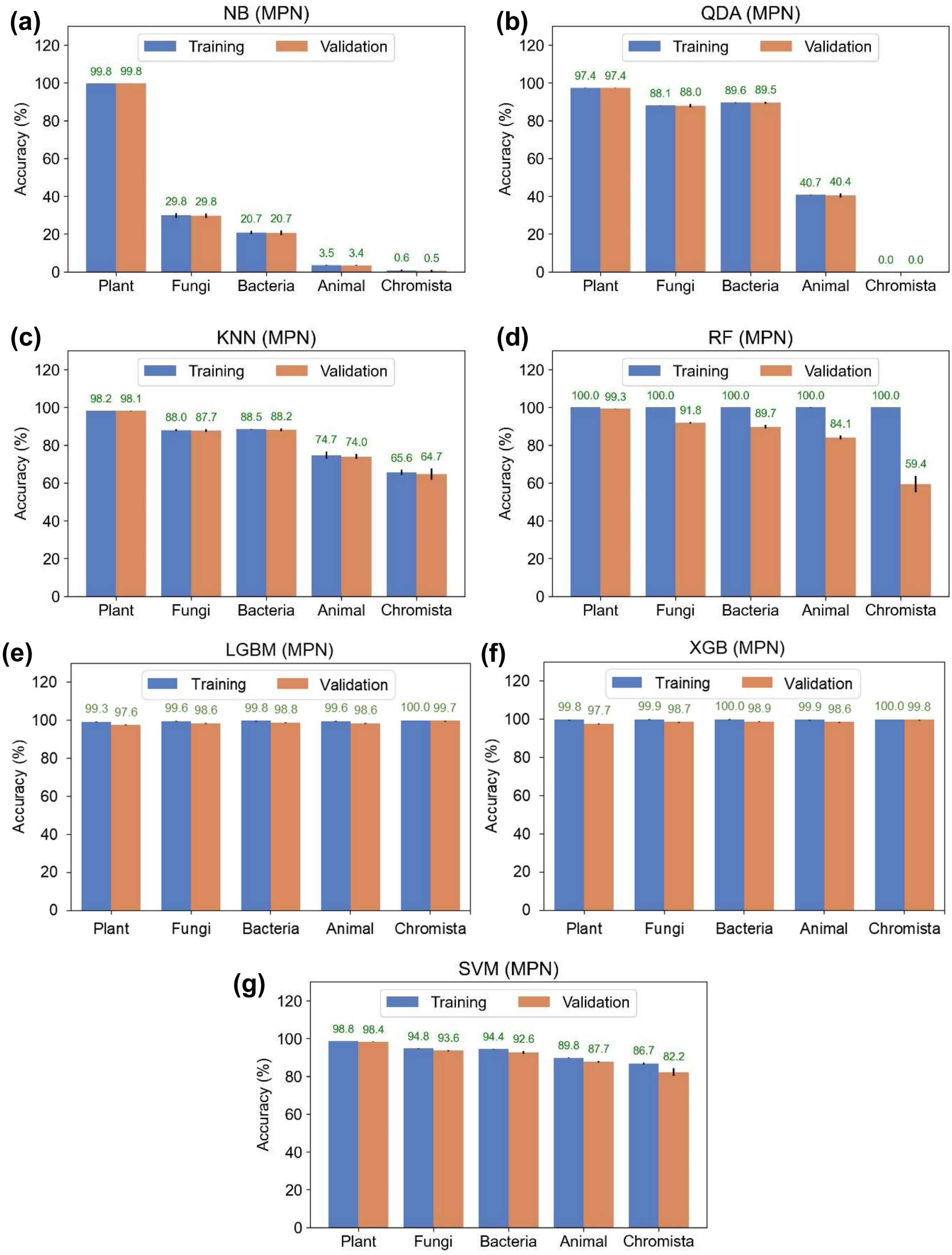

Figure S7. Stratified 5-fold cross validated classification accuracies across 5 different kingdoms using MPN fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

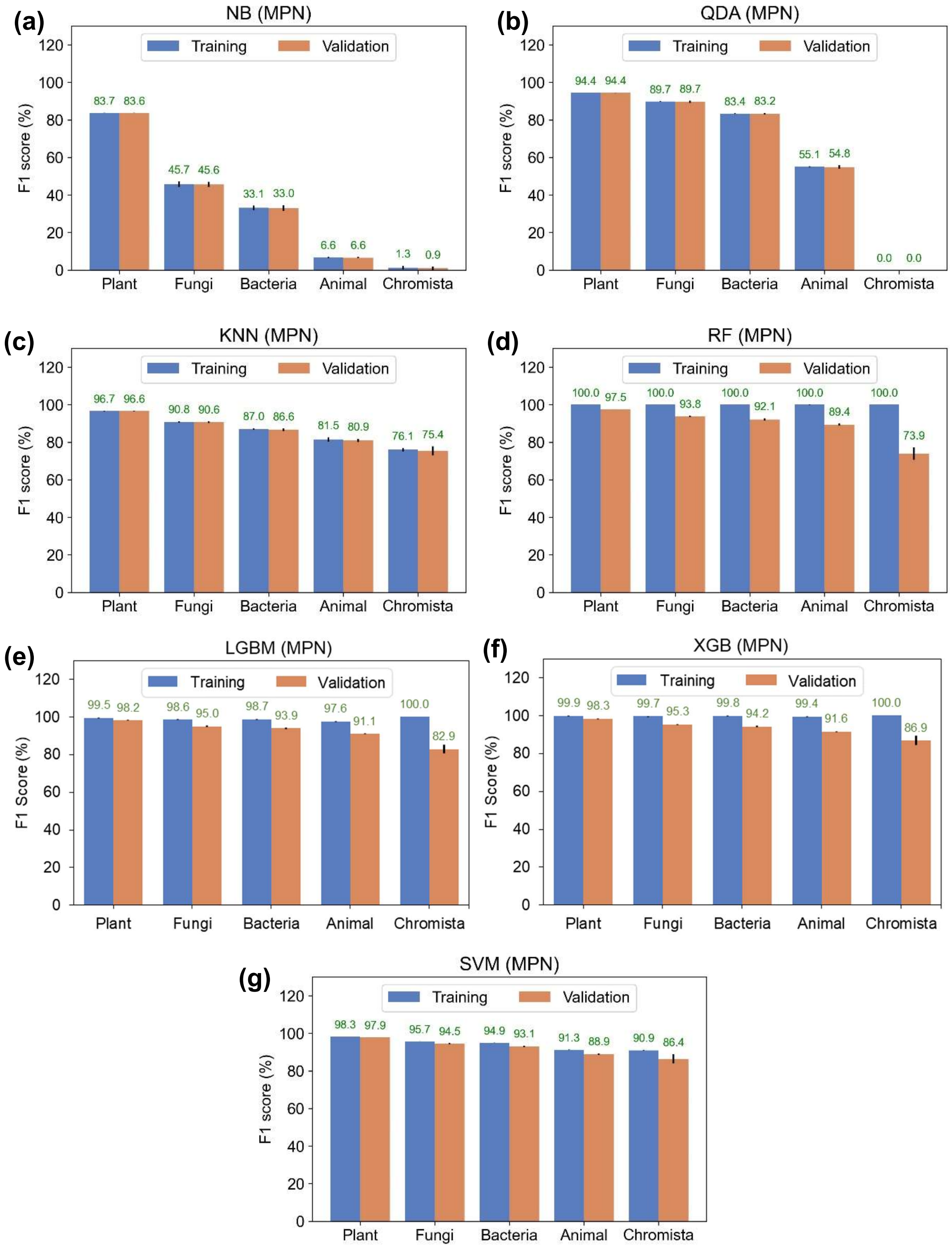

Figure S8. Stratified 5-fold cross validated F1 scores across 5 different kingdoms using MPN fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

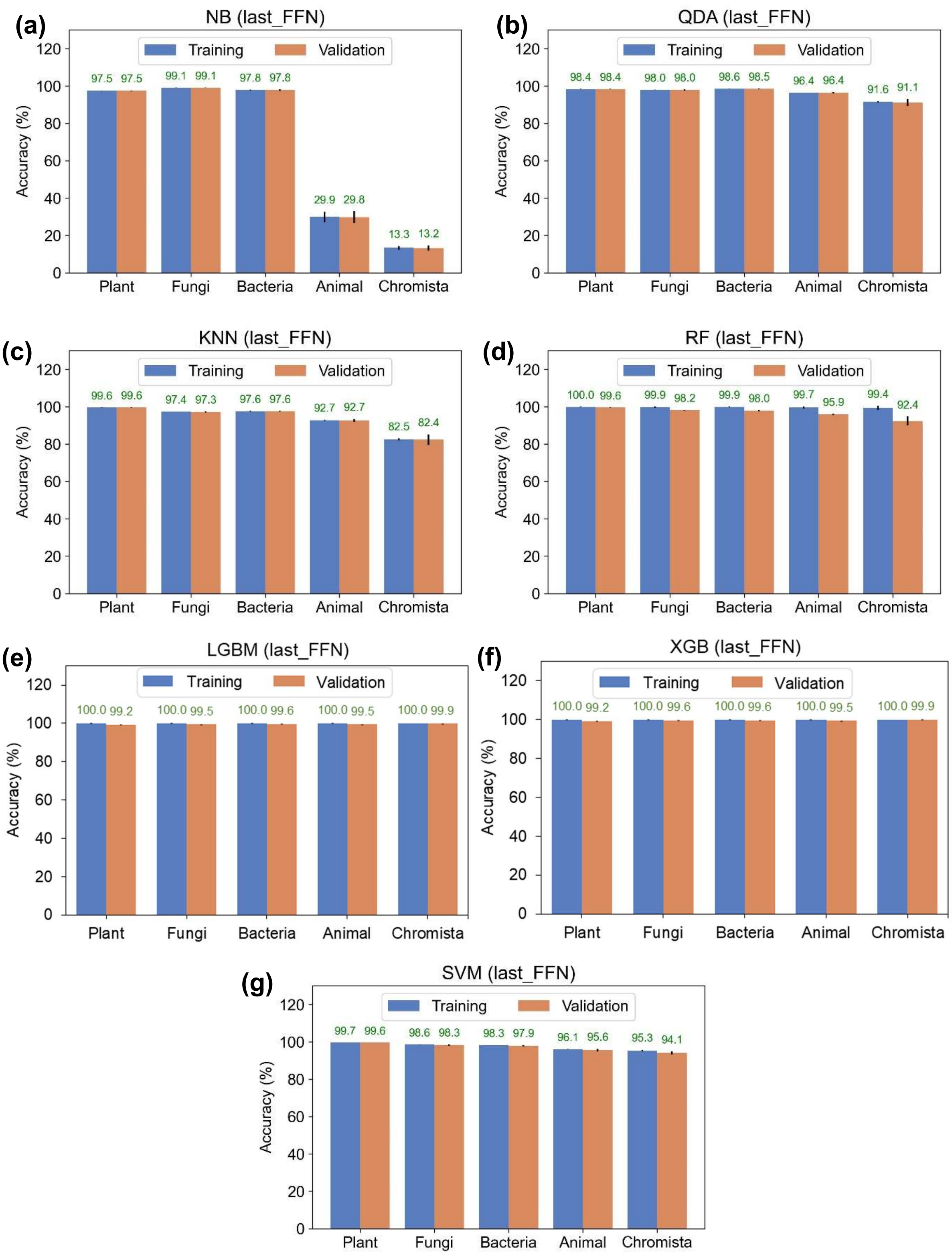

Figure S9. Stratified 5-fold cross validated classification accuracies across 5 different kingdoms using last_FFN fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

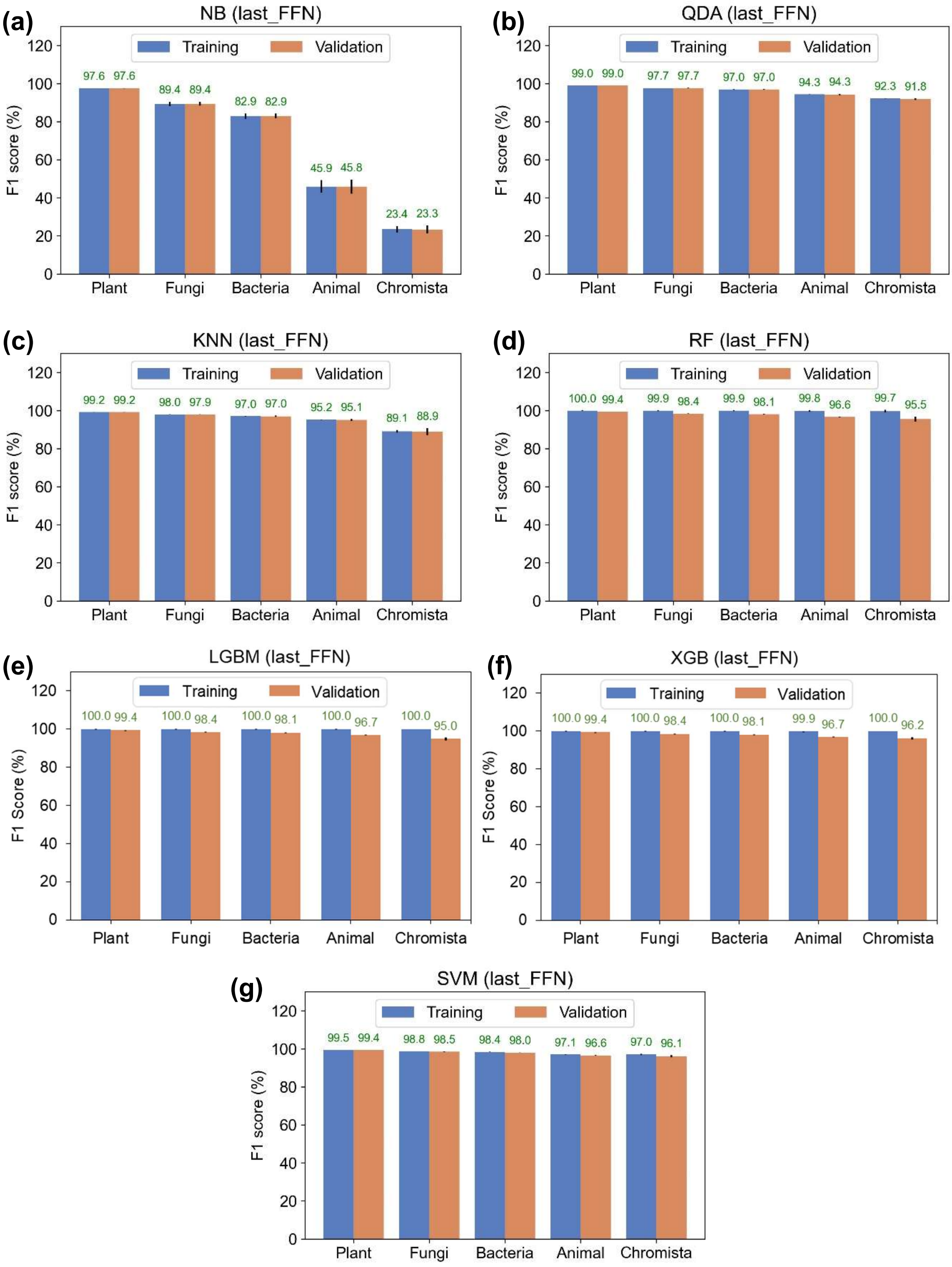

Figure S10. Stratified 5-fold cross validated F1 scores across 5 different kingdoms using last_FFN fingerprints through various machine learning algorithms: (a) NB, (b) QDA, (c) KNN, (d) RF, (e) LGBM, (f) XGB, and (g) SVM. Error bars of 1 standard deviation shown.

3. *t*-SNE results

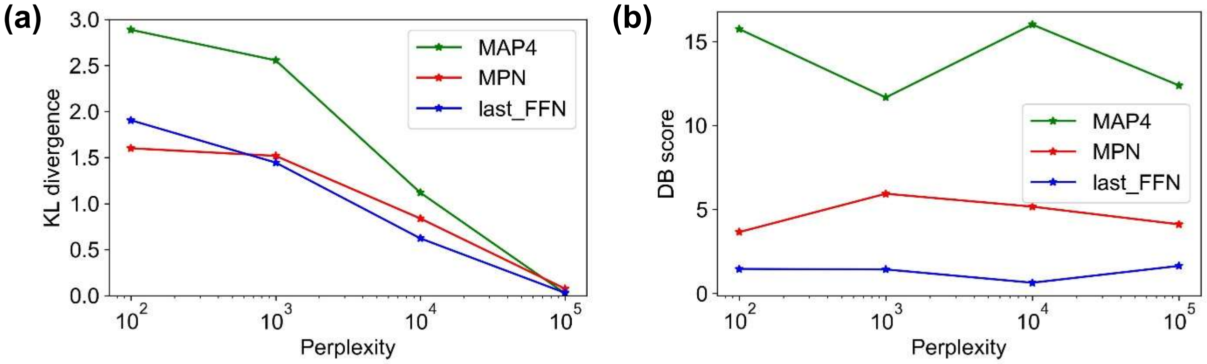

Figure S11. (a) Kullback–Leibler (KL) divergence and (b) Davies-Bouldin (DB) score as a function of perplexity for MAP4, MPN and last_FFN fingerprints.

References

1. C. C. C. R. De Carvalho and M. J. Caramujo, *Molecules*, 2018, **23**, 2583.
2. J. Volkman, *Appl. Microbiol. Biotechnol.*, 2003, **60**, 495-506.
3. T. Akihisa, N. Higo, H. Tokuda, M. Ukiya, H. Akazawa, Y. Tochigi, Y. Kimura, T. Suzuki and H. Nishino, *J. Nat. Prod.*, 2007, **70**, 1233-1239.
4. R. S. Singh, G. K. Saini and J. F. Kennedy, *Carbohydr. Polym.*, 2008, **73**, 515-531.