Supplementary Data

Table S1. Cross-validation accuracy across SNN and PID bins.

SNN Bin	PID Bin	Correct	Total	Percent Accuracy
0-1	<60	1458	1841	0.79
0-1	60-70	228	251	0.91
0-1	<60	16	34	0.47
0-1	60-70	812	905	0.90
0-1	70-80	670	753	0.89
0-1	80-90	454	548	0.83
0-1	>90	1428	1771	0.81
1-2	60-70	1	1	1.00
1-2	70-80	176	183	0.96
1-2	80-90	317	349	0.91
1-2	>90	648	721	0.90
2-3	70-80	22	22	1.00
2-3	80-90	105	111	0.95
2-3	>90	602	654	0.92
3-4	70-80	2	2	1.00
3-4	80-90	57	57	1.00
3-4	>90	349	359	0.97
>4	80-90	25	25	1.00
>4	>90	195	198	0.98

Table S2. Summary and accession information for all A-domains used in this study.

See attached file: SANDPUMA_TableS2.xlsx

Table S3. A-domain prediction methods used.

Method	Predictor Type	Year of Training
Bachmann-Ravel	Active-site Motifs	2009
Khayatt	Profile Hidden Markov Models	2013
prediCAT	Tree-based	2016
Minowa	Profile Hidden Markov Models	2007
NRPSPredictor2 ASM	Active-site Motifs	2011
NRPSPredictor2 SVM	Support Vector Machines	2011
NRPSSP	Profile Hidden Markov Models	2012
SEQL-NRPS	Greedy Coordinate-descent	2015

Table S4. McNemar p-values < 0.05 for shared coverage A-domains denote significant differences between correct and incorrect classifications between algorithms. Absent pairings were not significant.

Method A	Method B	Shared Coverage Count	McNemar p-value
NRPSPredictor2 - ASM	Minowa	336	2.41E-13
NRPSPredictor2 - ASM	Khayatt	429	2.39E-09
NRPSPredictor2 - SVM	Minowa	262	5.39E-09
NRPSPredictor2 - ASM	SEQL-NRPS	429	4.59E-08
NRPSPredictor2 - ASM	prediCAT - Monophyly & NN	410	9.56E-07
NRPSPredictor2 - ASM	NRPSSP	429	1.21E-05
NRPSPredictor2 - SVM	Khayatt	303	1.81E-05
NRPSSP	Minowa	337	5.23E-05
NRPSPredictor2 - ASM	Bachmann-Ravel	304	5.23E-05
NRPSPredictor2 - SVM	SEQL-NRPS	303	1.78E-04
NRPSPredictor2 - SVM	prediCAT - Monophyly & NN	290	2.00E-04
Khayatt	Minowa	337	1.43E-03
prediCAT - Monophyly & NN	Minowa	321	3.57E-03
Bachmann-Ravel	Minowa	272	3.57E-03
prediCAT - SNN Score >= 0.5	Minowa	58	4.43E-03
NRPSPredictor2 - SVM	Bachmann-Ravel	247	1.04E-02
prediCAT - SNN Score >= 0.5	SEQL-NRPS	63	1.33E-02
Bachmann-Ravel	SEQL-NRPS	306	1.61E-02
prediCAT - SNN Score >= 0.5	NRPSSP	63	4.12E-02
NRPSPredictor2 - SVM	NRPSSP	303	4.25E-02

Table S5. Classification statistics of SANDPUMA (max depth 40, minimum leaf support 10) on cross-validated (CV) dataset by substrate specificity. Cross-validation followed the scheme in Figure S2. The full dataset was randomized and broken into ~10% subsets (either 93 or 92 A-domain sequences). For each subset, SANDPUMA's individual algorithms were trained on the remaining 90% of the data. Method accuracy was assessed by querying the subset against these models. 100 query sets and training sets were assessed (10 randomizations each with 10 subsets) totaling 9280 individual queries.

Substrate	# CV Queries	Precision	Recall	F-score
(4R)-4[(E)-2-butenyl]-4-methyl- L-threonine	29	0.476	0.345	0.4
2,3-dehydroaminobutyric acid	40	0.719	0.575	0.639
2,3-dihydroxy-benzoic acid	10	1	1	1
2,4-diamino-butyric acid	128	0.847	0.906	0.875
2-aminobutyric acid	20	1	0.5	0.667
3,5-dihydroxy-phenyl-glycine	150	1	1	1
3-aminoadipic acid	110	1	1	1
3-hydroxy-3-methyl-proline	20	1	0.85	0.919
4-hydoxy-phenyl-glycine	88	1	0.989	0.994
4-hydroxy-proline	20	1	0.9	0.947
Alanine	545	0.863	0.783	0.821
Allo-isolucine	10	0.769	1	0.87
Allo-threonine	30	0.667	0.333	0.444
Alpha-hydroxy-glycine	10	0.435	1	0.606
Alpha-methyl-serine	10	1	0.7	0.824
Arginine	90	0.673	0.367	0.475
Asparagine	297	0.993	0.953	0.973
Aspartic acid	250	0.942	0.916	0.929
Beta-alanine	40	0.682	0.375	0.484
Beta-hydroxy-tyrosine	95	0.989	0.926	0.957
Cysteine	250	1	0.864	0.927
D- 2,4-diamino-butyric acid	10	1	0.8	0.889
D-alanine	40	0.833	0.25	0.385
D-allo-threonine	20	0.619	0.65	0.634
D-asparagine	10	0.714	1	0.833
D-cyclo- hydroxyornithine	10	0.2	0.2	0.2
Deoxy-threonine	20	1	0.9	0.947
D-glutamine	30	1	0.4	0.571
D-leu	50	0.634	0.52	0.571
D-proline	19	0.783	0.947	0.857
D-serine	30	0.897	0.867	0.881
D-tryptophan	10	0.833	0.5	0.625
D-valine	80	1	0.675	0.806
Gamma-hydroxy-valine	10	0.889	0.8	0.842
Glutamic acid	210	0.864	0.881	0.873
Glutamine	200	0.925	0.62	0.743

Glycine	380	1	0.887	0.94
Histidine	30	1	0.267	0.421
Hydroxyornithine	110	0.875	0.636	0.737
Isoleucine	210	0.925	0.824	0.872
Kynurenic acid	20	1	0.8	0.889
Leucine	760	0.91	0.854	0.881
Lysine	190	0.812	0.589	0.683
Methionine	10	0.286	0.8	0.421
Methyl-aspartic acid	40	0.976	1	0.988
N-hydroxy-alanine	10	1	0.8	0.889
N-hydroxy-valine	20	0.95	0.95	0.95
N-methyl-hydroxyornithine	20	0.15	0.3	0.2
Ornithine	210	0.774	0.867	0.818
Phenylacetate	30	0.903	0.933	0.918
Phenylalanine	209	0.904	0.722	0.803
Pipecolic acid	80	0.982	0.7	0.818
Proline	300	0.86	0.837	0.848
Serine	630	0.979	0.897	0.936
Threonine	524	0.905	0.889	0.897
Tryptophan	200	0.882	0.485	0.626
Tyrosine	260	0.942	0.746	0.833
Valine	725	0.92	0.83	0.873

Table S6: Comparison across methods of substrate specificity F-scores assessed by a cross-validation scheme as described in Figure S2 and Table S5.

Substrate	# CV Queries	prediCAT_MP	prediCAT_SNN	ASM	SVM	pHMM	SANDPUMA
3oh-3me-pro	20	0.947	0.947	0.947	NA	NA	0.919
3oh-leu	20	0.947	0.923	1.000	NA	NA	NA
4oh-pro	20	NA	NA	NA	NA	NA	0.947
aad	110	0.830	0.682	1.000	0.887	0.900	1.000
abu	20	NA	NA	NA	NA	NA	0.667
aeo	20	NA	0.909	1.000	NA	NA	NA
aile	10	NA	NA	NA	NA	NA	0.870
ala	570	0.627	0.435	0.774	0.731	0.667	0.821
alpha-me-ser	10	NA	NA	NA	NA	NA	0.824
aoh-gly	10	NA	NA	NA	NA	NA	0.606
arg	90	0.357	0.361	0.857	0.562	0.459	0.475
asn	300	0.762	0.625	0.948	0.963	0.843	0.973
asp	250	0.802	0.652	0.920	0.930	0.777	0.929
athr	30	NA	NA	NA	NA	NA	0.444
b-ala	40	NA	NA	0.750	NA	NA	0.484
bht	100	0.734	0.988	1.000	0.959	0.876	0.957
blys	30	NA	NA	1.000	NA	0.154	NA
bmt	29	NA	NA	NA	NA	NA	0.400
cys	250	0.804	0.345	0.948	0.947	0.892	0.927
dab	130	0.817	0.706	0.932	NA	0.843	0.875
d-ala	40	NA	NA	NA	NA	NA	0.385
d-asn	10	NA	NA	NA	NA	NA	0.833
d-athr	20	NA	0.433	NA	NA	NA	0.634
d-cyclo-horn	10	NA	NA	NA	NA	NA	0.200
d-dab	10	NA	NA	NA	NA	NA	0.889
deoxy-thr	20	NA	0.690	NA	NA	NA	0.947
d-gln	30	NA	0.286	NA	NA	NA	0.571
dhabu	40	0.196	0.629	NA	NA	0.475	0.639
dhb	20	NA	NA	NA	1.000	NA	1.000
d-leu	50	NA	0.020	NA	NA	NA	0.571
dpg	150	0.976	1.000	1.000	NA	1.000	1.000
d-pro	20	NA	0.667	NA	NA	NA	0.857
d-ser	30	0.667	0.952	NA	NA	0.714	0.881
d-trp	10	NA	NA	NA	NA	NA	0.625
d-val	80	0.508	0.546	NA	NA	0.727	0.806
gln	200	0.421	0.372	0.737	0.640	0.299	0.743
glu	210	0.514	0.517	0.909	0.873	0.665	0.873
gly	380	0.786	0.389	0.951	0.930	0.650	0.940
goh-val	10	NA	NA	NA	NA	NA	0.842
his	30	NA	NA	NA	NA	NA	0.421
horn	110	0.589	0.467	0.735	NA	0.593	0.737
hpg	90	0.379	0.928	1.000	0.942	0.722	0.994
hse	20	NA	0.947	0.947	NA	NA	NA
ile	210	0.581	0.644	0.848	0.844	0.542	0.872
kyn	20	1.000	0.909	1.000	NA	NA	0.889
leu	760	0.681	0.604	0.885	0.813	0.703	0.881
lys	190	0.565	0.531	0.680	0.413	0.534	0.683
me-asp	40	0.857	0.988	1.000	NA	1.000	0.988
met	10	NA	NA	NA	NA	NA	0.421
nme-fhorn	20	NA	NA	NA	NA	NA	0.200
noh-ala	10	NA	NA	NA	NA	NA	0.889
noh-val	20	1.000	0.952	0.952	NA	NA	0.950
orn	210	0.772	0.676	0.833	0.554	0.721	0.818
phe	210	0.449	0.168	0.626	0.760	0.407	0.803
phe-ac	30	0.667	0.829	0.696	NA	0.710	0.918
pip	80	0.457	0.521	0.580	0.855	0.640	0.818
pro	300	0.633	0.544	0.896	0.838	0.796	0.848
ser	630	0.796	0.638	0.893	0.880	0.707	0.936
thr	540	0.548	0.510	0.376	0.799	0.763	0.897
trp	200	0.604	0.448	0.669	0.605	0.367	0.626
tyr	260	0.647	0.630	0.870	0.796	0.660	0.833
val	740	0.683	0.595	0.776	0.742	0.750	0.873

Table S7. Classification space of SANDPUMA and individual methods.

Method	Total Substrates in Classification Space
ASM	82
pHMM	52
prediCAT (SNN >=0.5)	73
prediCAT Monophyly	67
SVM	26
SANDPUMA	104

Fig. S1. Frequency of nearest neighbor distance across cross-validation analysis. Empirical cutoff of 2.5 shown as dotted line.

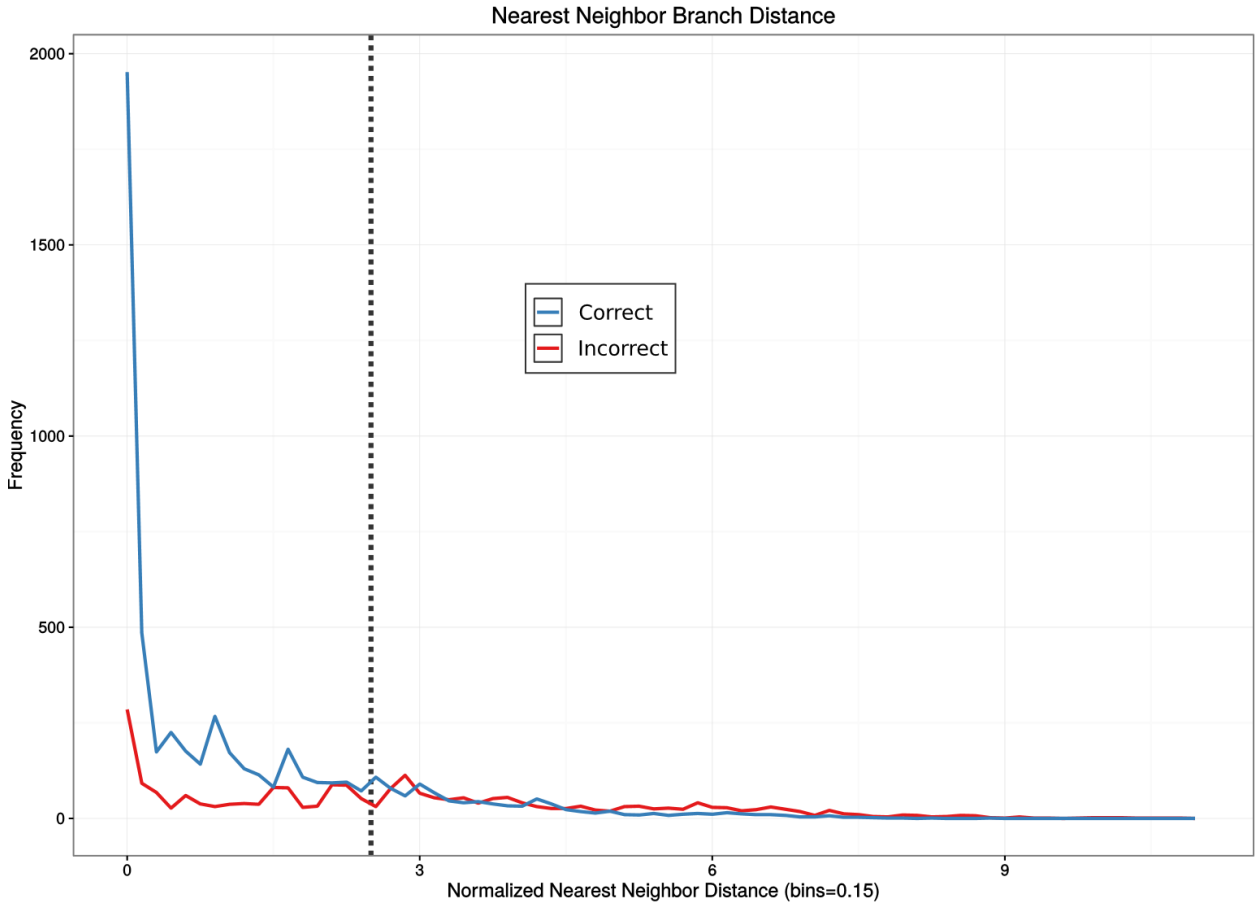

Fig. S2. Schematic of cross-validation resampling analysis.

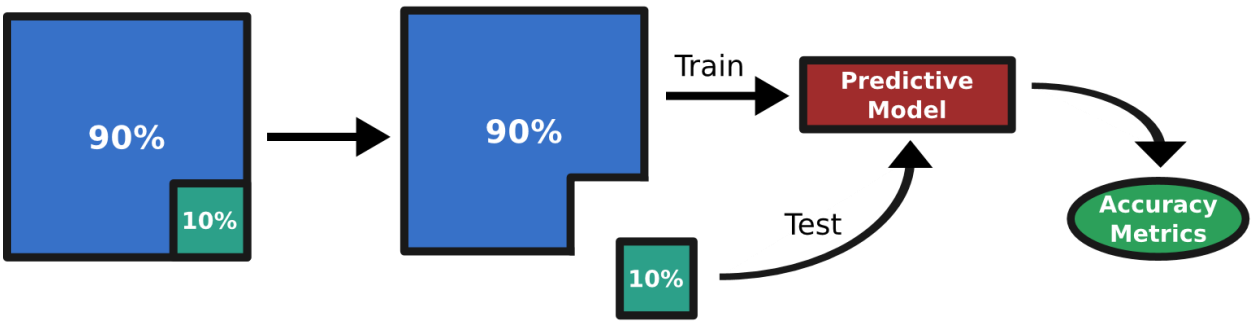

Fig. S3. Decision tree accuracies across max depths and minimum leaf supports.

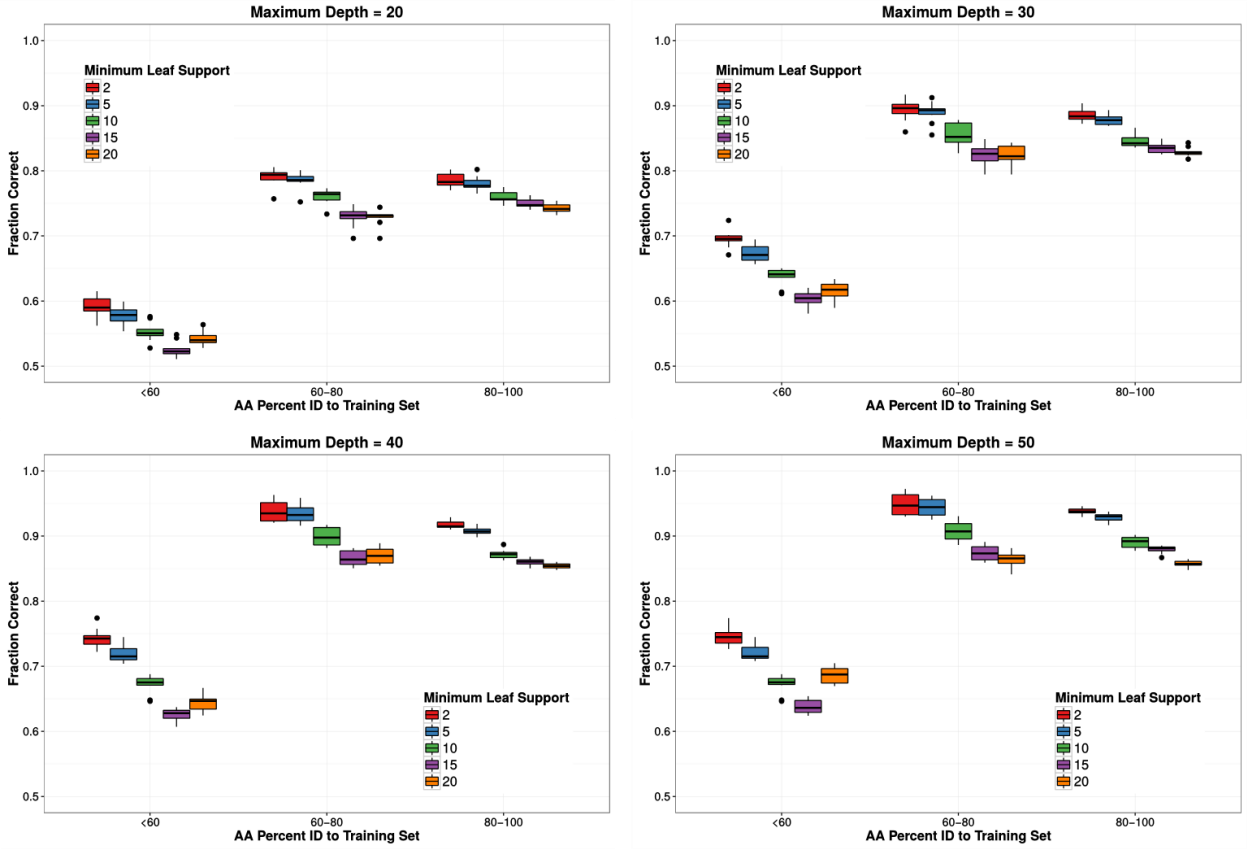

Fig. S4. SANDPUMA SML decision training with methods removed for full dataset (a) and for amino acids with 2 or less examples in the dataset (b).

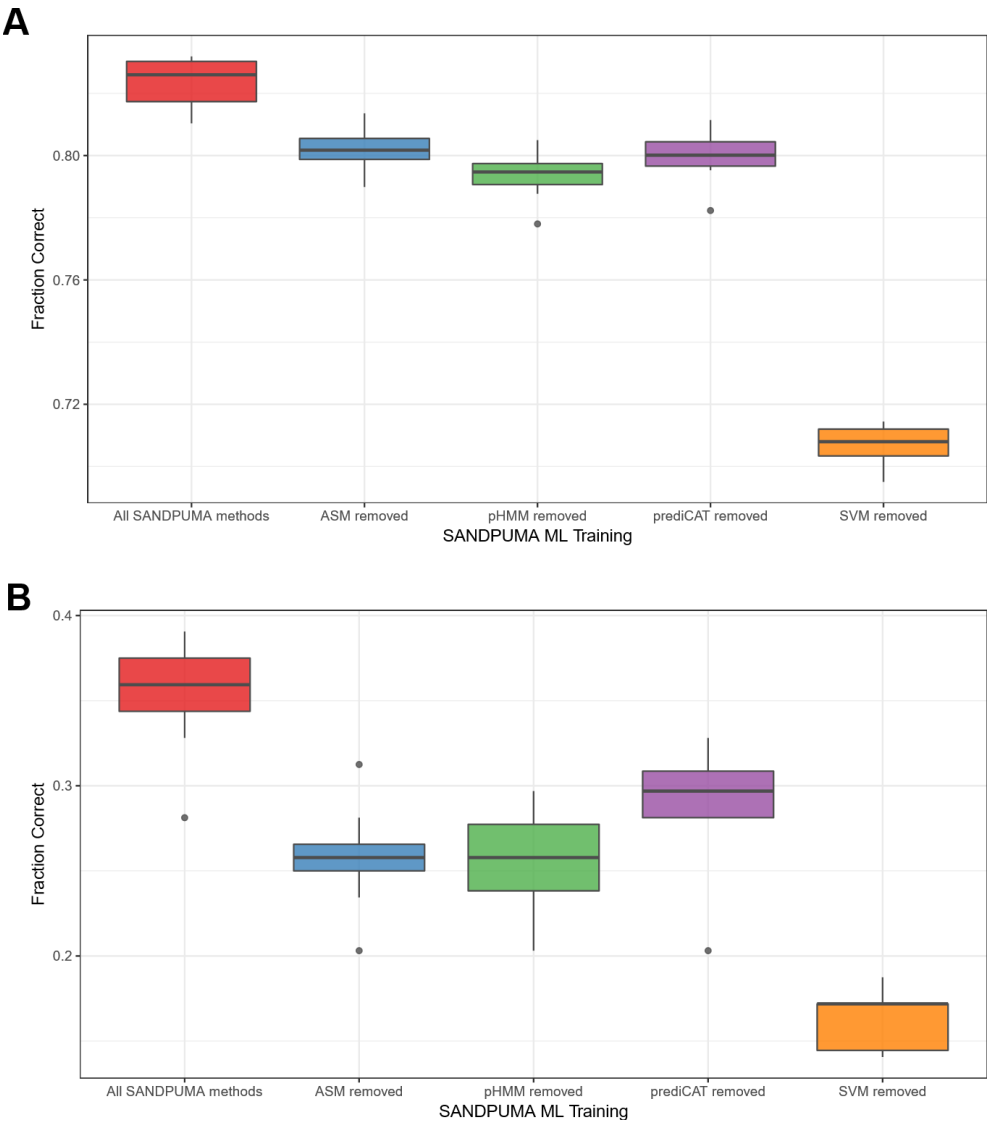

Fig. S5. Accuracy of SML decision paths.

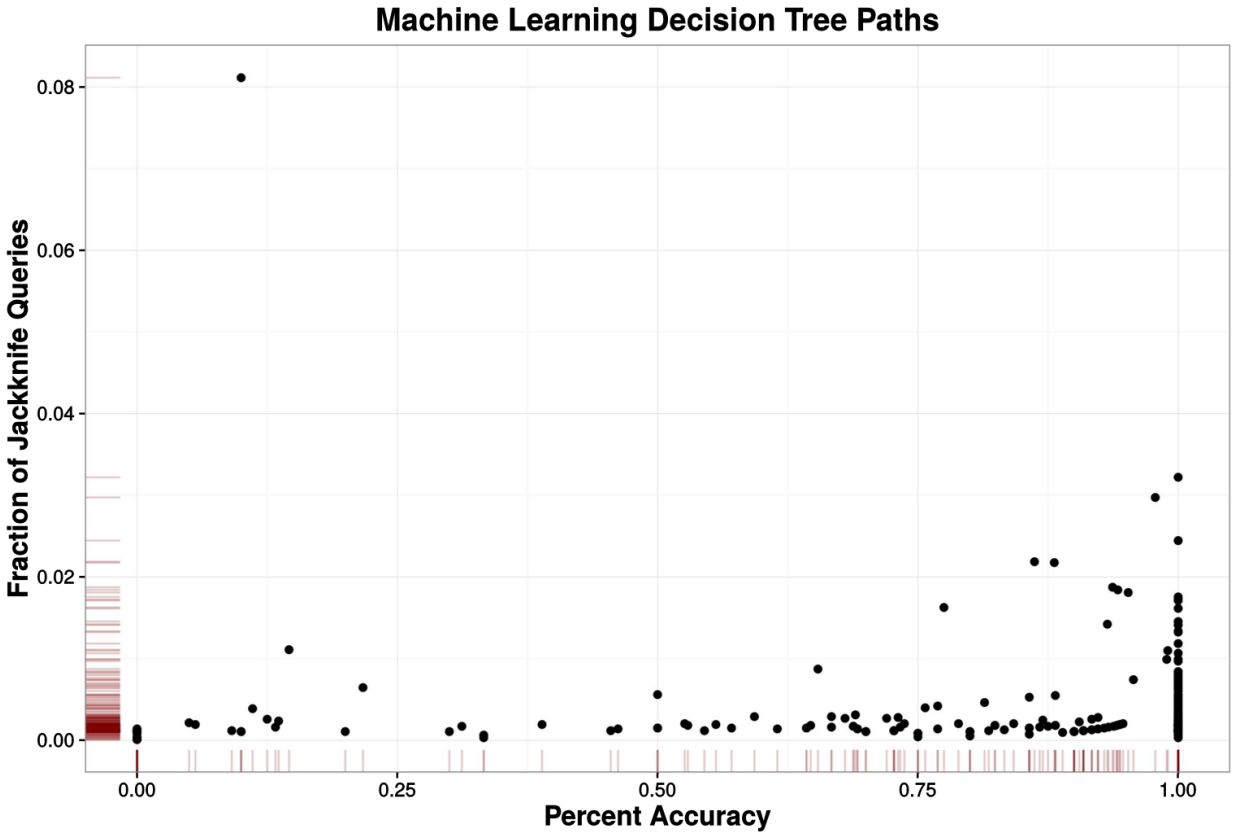

Fig. S6. Cross-validation shared coverage for individual methods (a), fraction correct (b), and coverage (c).

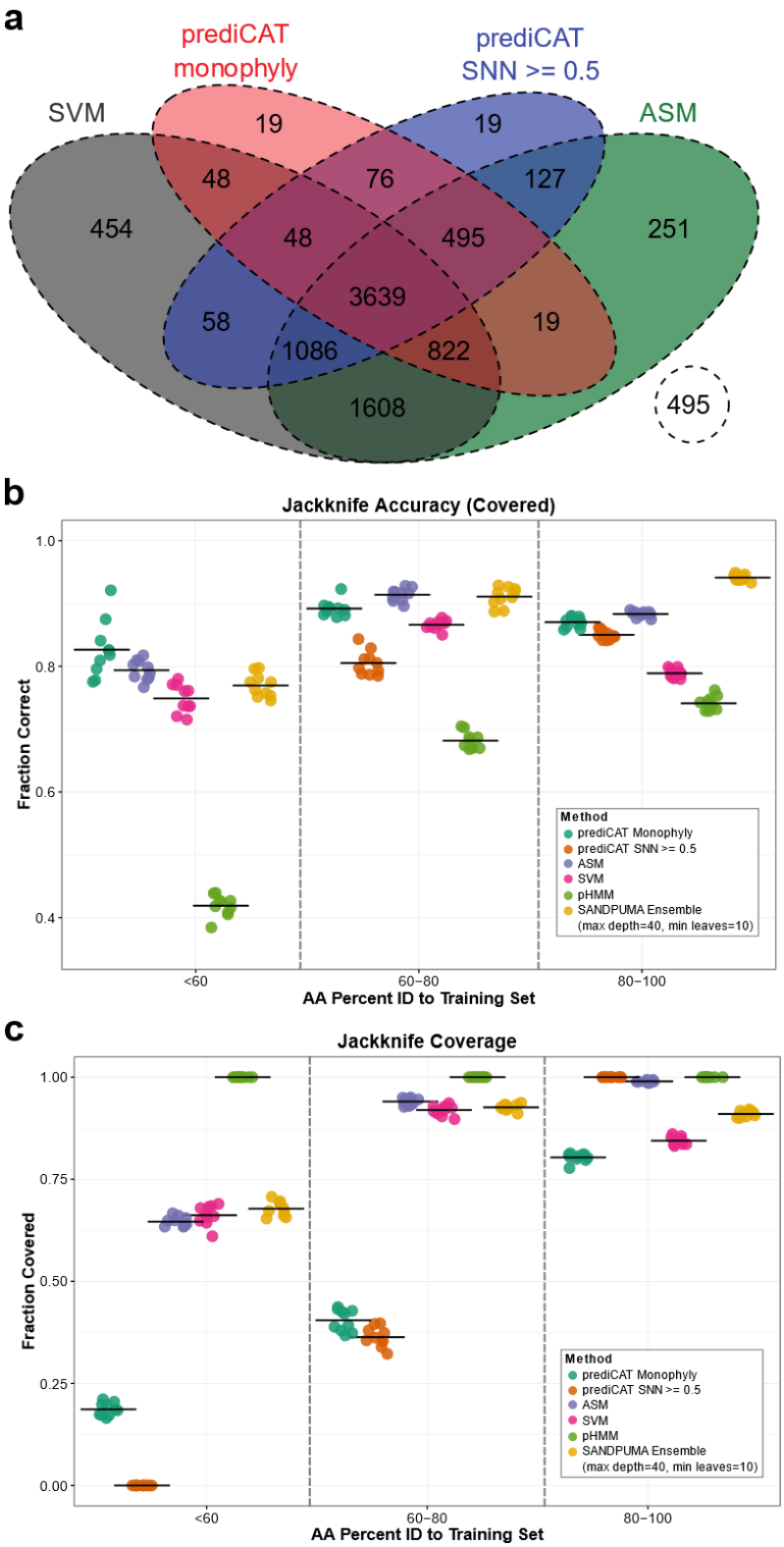

Fig. S7. Accuracy (a) and coverage (b) comparisons of cross-validations.CV2 represents a second, independent cross-validation.

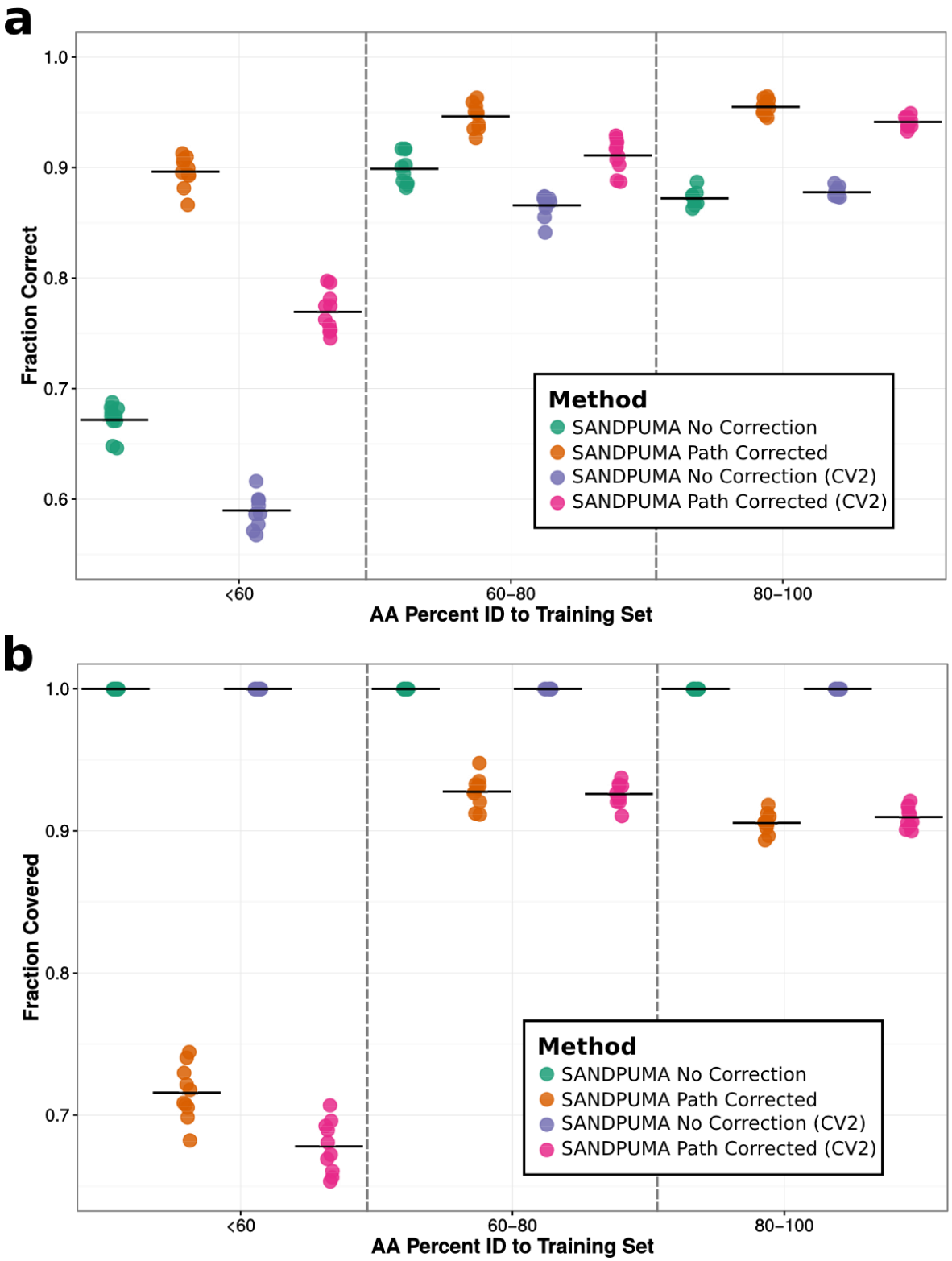

Fig. S8. Multilocus phylogeny of Actinobacteria denoting number of genomes available and NRP BGCs per genome. Branches with bootstrap values <100 are marked.

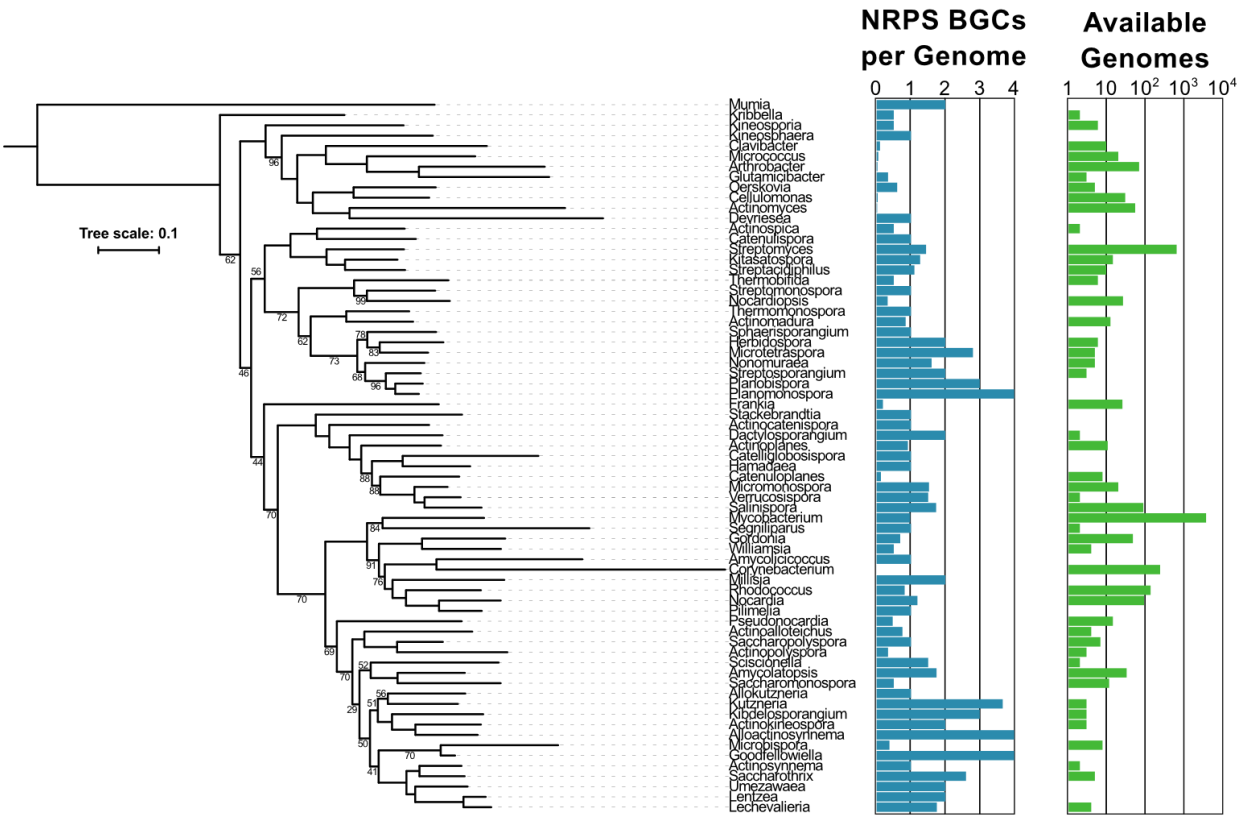

Fig. S9. Distribution of the number of network subgraphs at each size for all NRP BGCs (blue) and NRP BGCs >10kb from the end of its contig (red).