Supplemental data

Comparative investigation into formycin A and pyrazofurin A biosynthesis reveals branch pathways for the construction of *C*-nucleoside scaffolds

Meng Zhang,[1,5] Peichao Zhang,[1,2,5] Gudan Xu,[1,5] Wenting Zhou,[1] Yaojie Gao,[1] Rong Gong,[1] You-Sheng Cai,[1] Hengjiang Cong,[3] Zixin Deng,[1] Neil P. J. Price,[4] Xiangzhao Mao[2]*, Wenqing Chen[1]*

[1]Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education, and School of Pharmaceutical Sciences, Wuhan University, Wuhan 430071, China

[2]College of Food Science and Engineering, Ocean University of China, Qingdao 266003, China

[3]College of Chemistry and Molecular Sciences, Institute for Advanced Studies (IAS), Wuhan University, Wuhan 430072, China

[4]Agricultural Research Service, US Department of Agriculture, National Center for Agricultural Utilization Research, Peoria, IL, USA

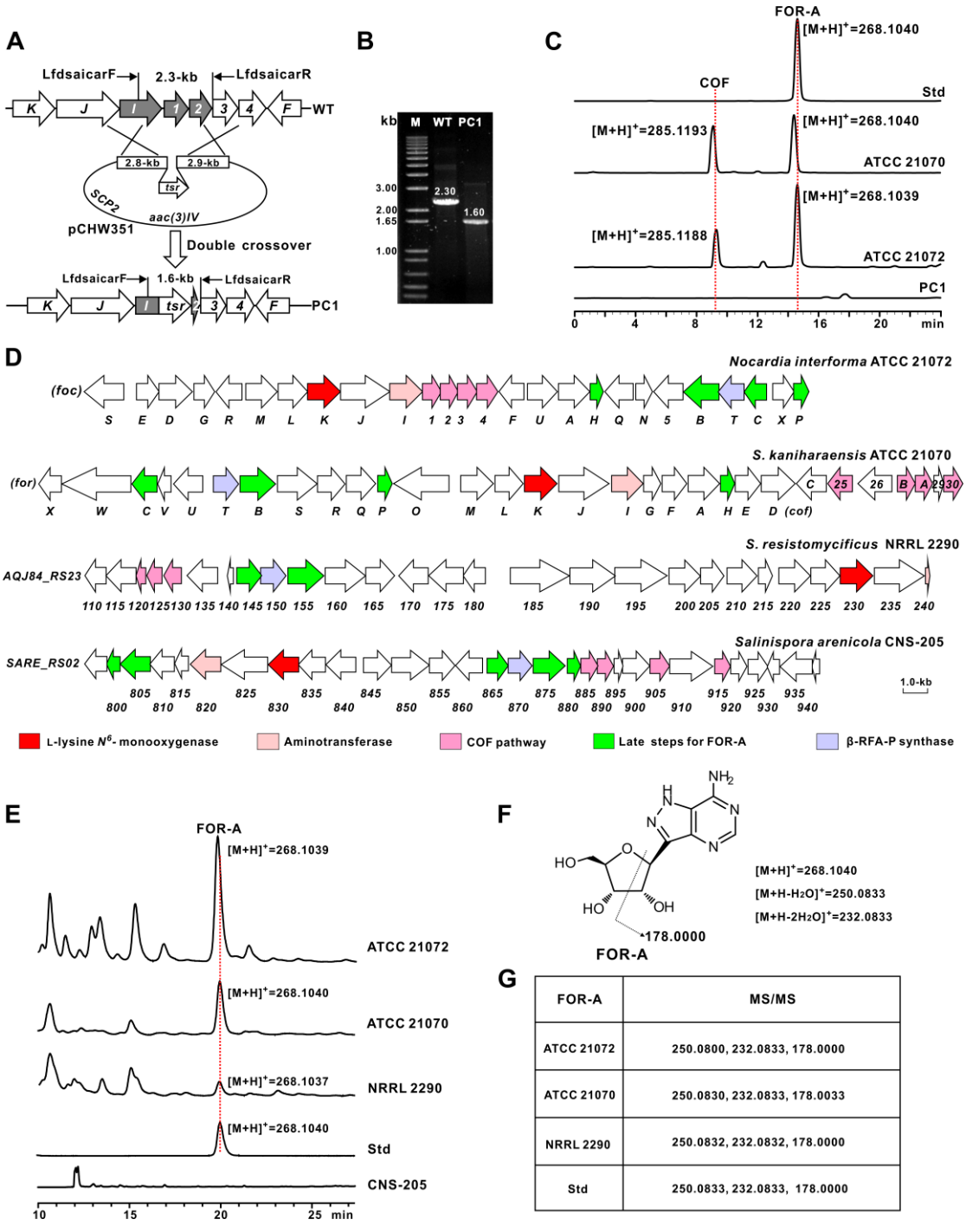
[5] These authors contributed equally to this paper.

*For Correspondence: **Wenqing Chen**, School of Pharmaceutical Sciences, Wuhan University, Wuhan 430071, China. **E-mail**: wqchen@whu.edu.cn

*For Co-correspondence: **Xiangzhao Mao,** College of Food Science and Engineering, Ocean University of China, Qingdao 266003, China. **E-mail**: xzhmao@ouc.edu.cn

Table of Contents

2. Supplementary Figures.



FOR-A	MS/MS
ATCC 21072	250.0800, 232.0833, 178.0000
ATCC 21070	250.0830, 232.0833, 178.0033
NRRL 2290	250.0832, 232.0832, 178.0000
Std	250.0833, 232.0833, 178.0000

Figure S1. Verification and genetic organizations of the FOR-A gene clusters

(A) Schematic illustration for the construction of *Nocardia interforma* PC1 mutant. (B) PCR identification of the mutant PC1. M, 1 kb plus DNA ladder; WT, PCR product using genomic DNA of *Nocardia interforma* ATCC 21072 as template; PC1, PCR product using genomic DNA of *Nocardia interforma* PC1 as template. (C) LC-HRMS

analysis of the target metabolites produced by related strains. Std, the authentic standard of FOR-A; ATCC 21070, the metabolites (COF and FOR-A) produced by *S. kaniharaensis* ATCC 21070 as positive control; ATCC 21072, the metabolites (COF and FOR-A) produced by the strains of *Nocardia interforma* ATCC 21072; PC1, the metabolites produced by the strains of *Nocardia interforma* PC1 mutant. The main fragment ions of the metabolite COF produced by *S. kaniharaensis* ATCC 21070 are as follows: *m/z* 267.1104, 152.9898, and 134.9333. (D) Genetic organizations of the FOR-A and COF (-related) gene clusters. The *foc* gene cluster is from *Nocardia interforma* ATCC 21072, the *for-cof* gene cluster is from *S. kaniharaensis* ATCC 21070, the *AQJ84_RS23* gene cluster is from *S. resistomycificus* NRRL 2290, and the *SARE_RS02* gene cluster (potentially responsible for FOR-A and COF biosynthesis) is from *Salinispora arenicola* CNS-205. (E) HPLC analysis of the metabolites individually produced by the strains of ATCC 21072, ATCC 21070, NRRL 2290, and CNS-205. ATCC 21072, the metabolites produced by *Nocardia interforma* ATCC 21072； ATCC 21070, the metabolites produced by *S. kaniharaensis* ATCC 21070; NRRL 2290, the metabolites of *S. resistomycificus* NRRL 2290; Std, the authentic standard of FOR-A; CNS-205, the metabolites produced by *Salinispora arenicola* CNS-205. The fermentation conditions of *S. resistomycificus* NRRL 2290 and *Salinispora arenicola* CNS-205 are identical to that of *S. kaniharaensis* ATCC 21070. (F) The fragmentation pattern of the FOR-A authentic standard. (G) The target main MS/MS fragments of the metabolite FOR-A produced by related strains. ATCC 21072, *Nocardia interforma* ATCC 21072; ATCC 21070, *S. kaniharaensis* ATCC 21070; NRRL 2290, *S. resistomycificus* NRRL 2290; Std, the authentic standard of FOR-A. As we expected to emphasize the core conserved genes (whose functions could be easily assigned) related with the present manuscript, as a result, we just marked them with color, and left other genes unmarked (blank).

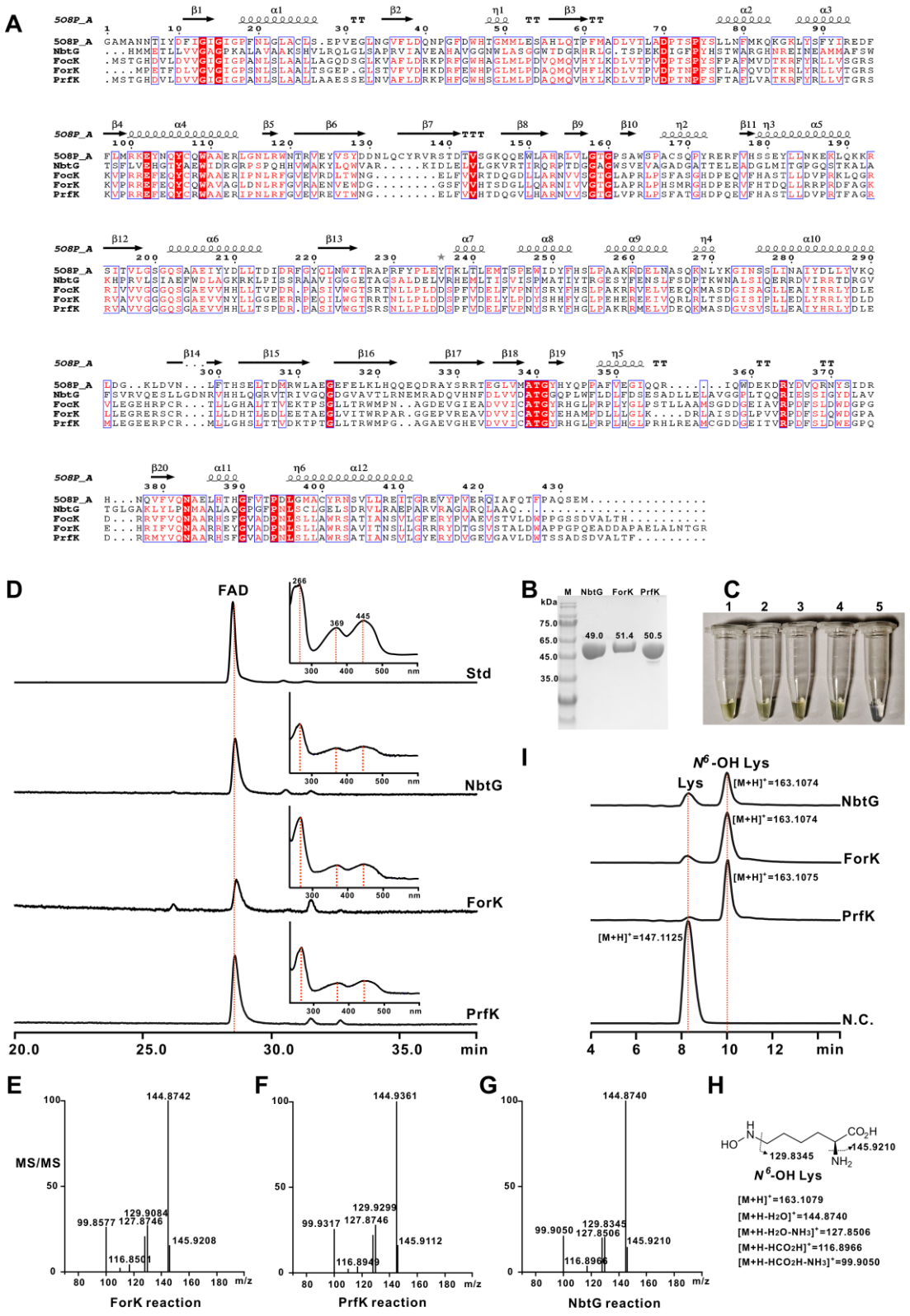
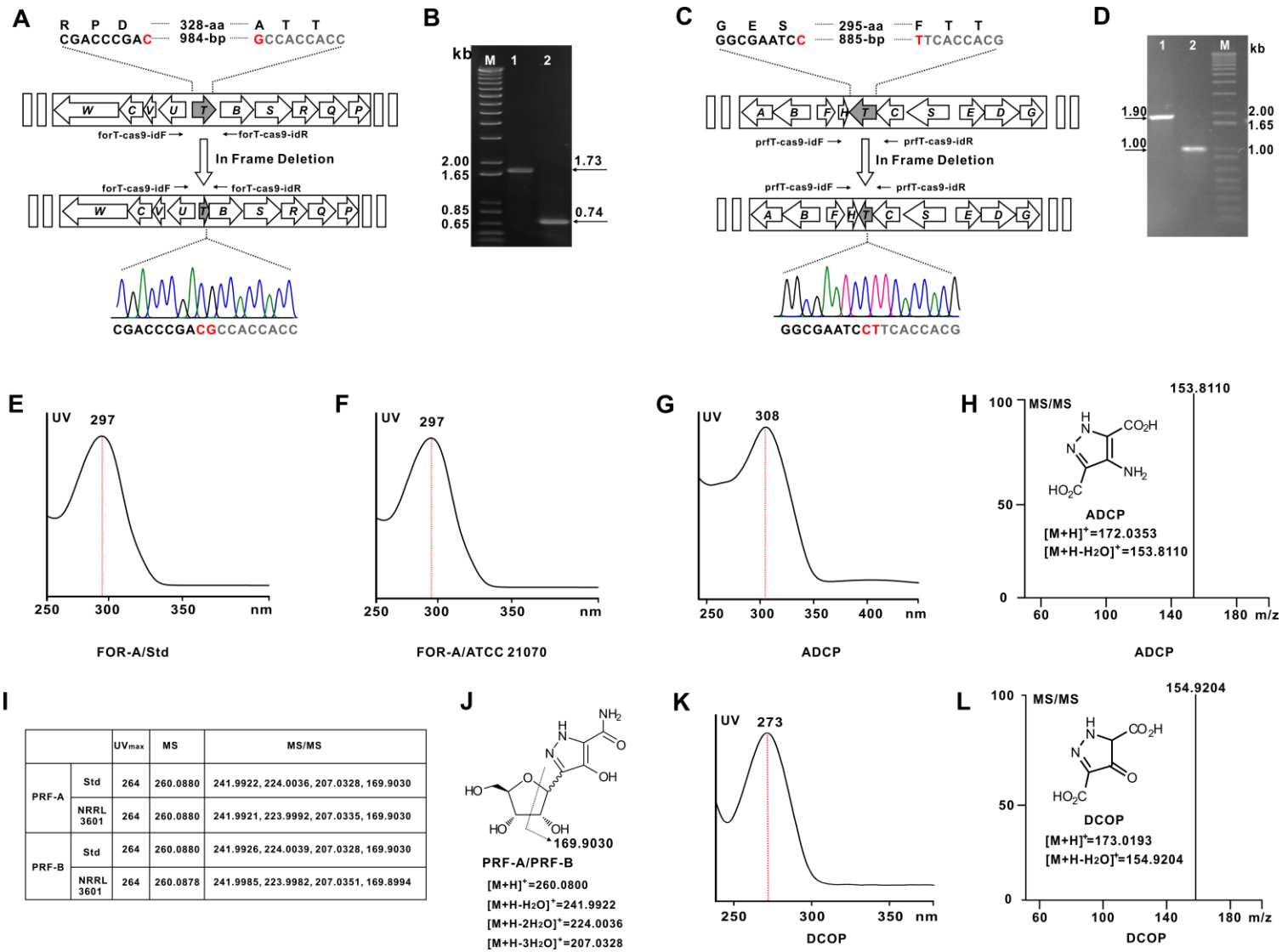


Figure S2. Enzymatic characterization of ForK/PrfK as a lysine N^6-monooxygenase

(A) Sequence alignment of FocK/ForK/PrfK with its homologs was performed using ESPript 3.0 (http://espript.ibcp.fr/ESPript/ESPript/). Secondary structure of 5O8P-A

(PDB: 5O8P-A), from *Erwinia amylovora* CFBP1430, is shown on the top. NbtG (GenBanK: BAD55606) is from *Nocardia farcinica* IFM 10152. Helices represent alpha-helices and arrows signify beta-strands. Amino acids (white) in red background with blue boxes means conserved regions, while the amino acids (red) within blue boxes means less conserved regions. (B) SDS-PAGE analysis of NbtG, ForK, and PrfK. The predicted molecular weights of NbtG (49.0 kDa), ForK (51.4 kDa) and PrfK (50.5 kDa) are consistent with the individual migrations on SDS-PAGE. (C) The color of the target purified proteins. 1, the authentic FAD standard as positive control; 2, the purified NbtG; 3, the purified ForK; 4, the purified PrfK; 5, the protein stock buffer used as negative control. (D) HPLC analysis of FAD isolated from proteins. The corresponding UV spectrum was included in the related trace. Std, the authentic FAD standard; NbtG, the FAD isolated from NbtG; ForK, the FAD isolated from ForK; PrfK, the FAD in purified protein PrfK. (E-G) LC-HRMS/MS analysis of the product N^6-OH Lys produced by the ForK/PrfK/NbtG reaction. (H) The fragmentation pattern of N^6-OH Lys produced by the NbtG reaction. N^6-OH Lys, N^6-OH Lysine. (I) LC-MS analysis of the ForK/PrfK/NbtG reactions. NbtG, the reaction catalyzed by NbtG; ForK, the reaction catalyzed by ForK; PrfK, the reaction catalyzed by PrfK; N.C, the reaction without enzyme added as negative control.



		UV_{max}	MS	MS/MS
PRF-A	Std	264	260.0880	241.9922, 224.0036, 207.0328, 169.9030
	NRRL 3601	264	260.0880	241.9921, 223.9992, 207.0335, 169.9030
PRF-B	Std	264	260.0880	241.9926, 224.0039, 207.0328, 169.9030
	NRRL 3601	264	260.0878	241.9985, 223.9982, 207.0351, 169.8994

Figure S3. *In vivo* functional characterization of *forT/prfT* in FOR-A/PRF-A biosynthesis

(A) Schematic illustration for the construction of *forT via* CRISPR-Cas9 technology. (B) PCR identification of the in-frame deletion mutant. M, 1-kb Plus DNA ladder; 1, PCR product using genomic DNA of *S. kaniharaensis* ATCC 21070 as template; 2, PCR product using genomic DNA of *S. kaniharaensis ΔforT* mutant as template. (C) Schematic illustration for the construction of *prfT via* CRISPR-Cas9 technology. (D) PCR identification of the in-frame deletion mutant. M, 1-kb Plus DNA ladder; 1, PCR product using genomic DNA of *S. candidus* NRRL 3601 as template; 2, PCR product using genomic DNA of *S. candidus ΔprfT* mutant as template. (E) UV spectrum of the FOR-A standard. (F) UV spectrum of FOR-A produced by *S. kaniharaensis* ATCC 21070. (G) UV spectrum of 4-amino-3,5-dicarboxypyrazole (ADCP) produced by *S. kaniharaensis ΔforT* mutant. (H) MS/MS analysis of ADCP. (I) The UV spectrums and MS/MS fragments of PRF-A/PRF-B. Std, the authentic standard of PRF-A/PRF-B; NRRL 3601, the target metabolite PRF-A/PRF-B produced by *S. candidus* NRRL 3601. (J) The fragmentation pattern of PRF-A/PRF-B. (K) UV spectrum of 3,5-dicarboxy-4-oxo-4,5-dihydropyrazole (DCOP). (L) MS/MS analysis of DCOP. The fragmentation pattern of DCOP is also indicated in this panel.

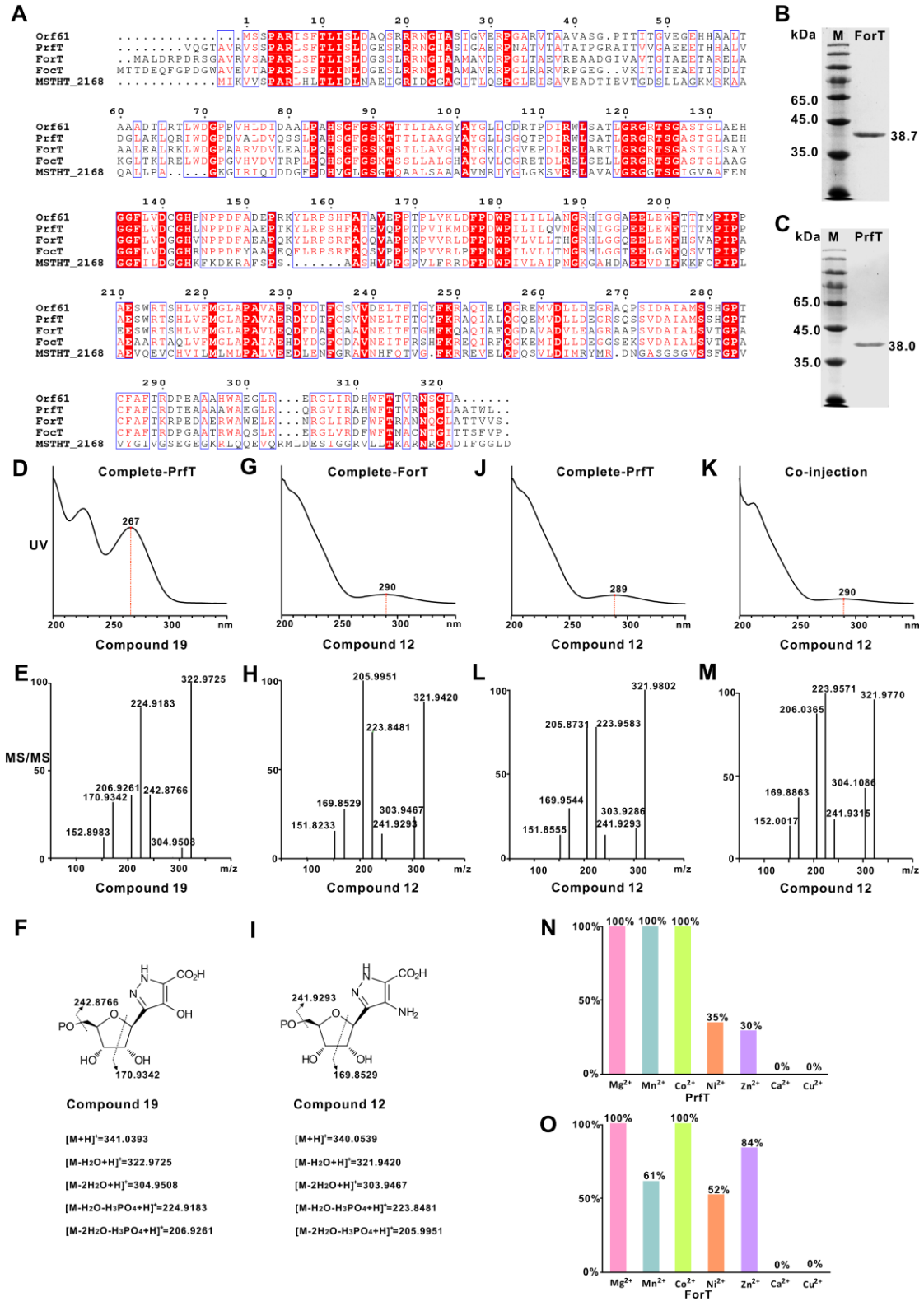


Figure S4. Biochemical characterization of PrfT/ForT as a β-RFA-P synthase-like enzyme

(A) Alignment of PrfT/ForT/FocT with relevant homologs. Orf61 (GenBanK:

AJO72754.1) is from *Nocardia terpenica* IFM 0406. MSTHT_2168 (GenBanK: AKB13926.1) is from *Methanosarcina thermophila* TM-1. White amino acids in red background with blue boxes means conserved regions, while the red amino acids with blue boxes means less conserved regions. (B) SDS-PAGE analysis of ForT. The predicted molecular weight of ForT (38.7 kDa) is consistent with the migration on SDS-PAGE. (C) SDS-PAGE analysis of PrfT. The predicted molecular weight of PrfT (38.0 kDa) is matched to the migration on SDS-PAGE. (D) UV spectrum of the PrfT-catalyzed product, compound **19**. (E) LC-HRMS/MS analysis of **19** produced by PrfT. (F) The fragmentation pattern of **19**. (G) UV spectrum of the ForT-catalyzed product, compound **12**. (H) LC-HRMS/MS analysis of **12** produced by ForT. (I) The fragmentation pattern of **12**. (J) UV spectrum of **12** produced by PrfT complete reaction. (K) UV spectrum of **12** (co-injection of the ForT and PrfT individual complete reactions). (L) LC-HRMS/MS analysis of **12** produced by PrfT. (M) LC-HRMS/MS analysis of **12** (co-injection of the ForT and PrfT individual complete reactions). (N) Relative efficiency of PrfT-catalyzed reaction with different divalent metal ions. (O) Relative efficiency of ForT-catalyzed reaction with different divalent metal ions.

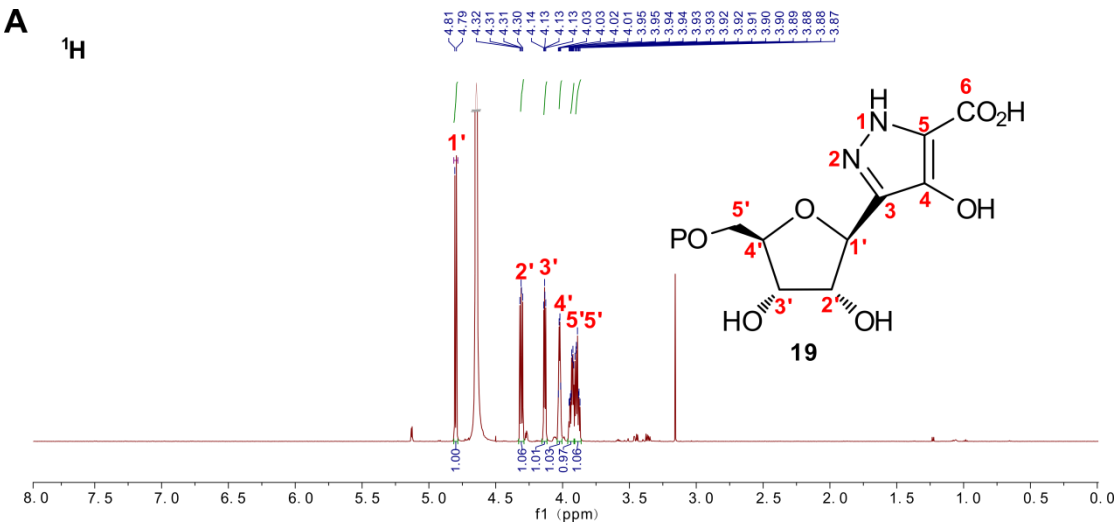


Figure S5A. ^{1}H NMR analysis of compound 19

^{1}H NMR data of **19** (600 MHz, D_2O).

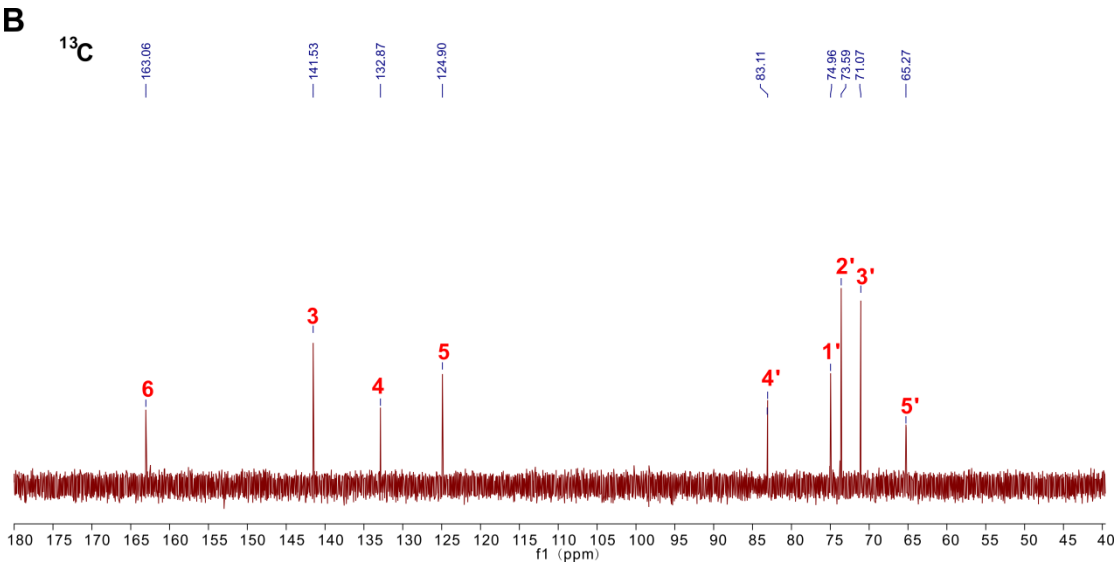


Figure S5B. ^{13}C NMR analysis of 19

^{13}C NMR data of **19** (600 MHz, D_2O).

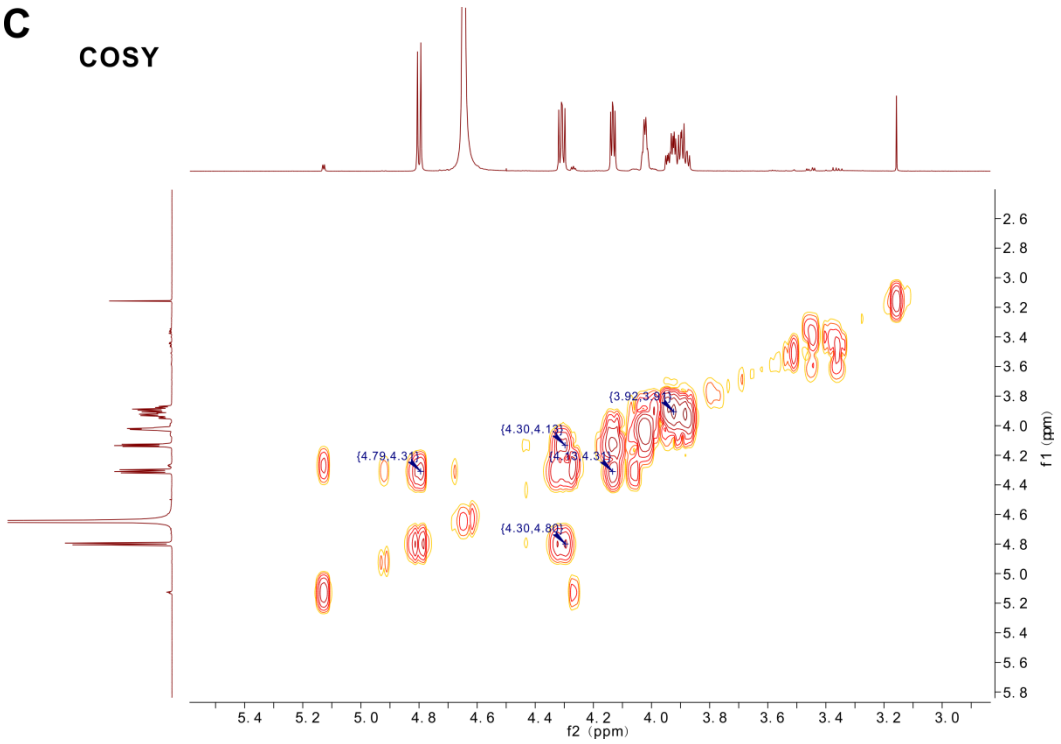


Figure S5C. COSY analysis of 19

COSY spectrum of **19** (600 MHz, D_2O).

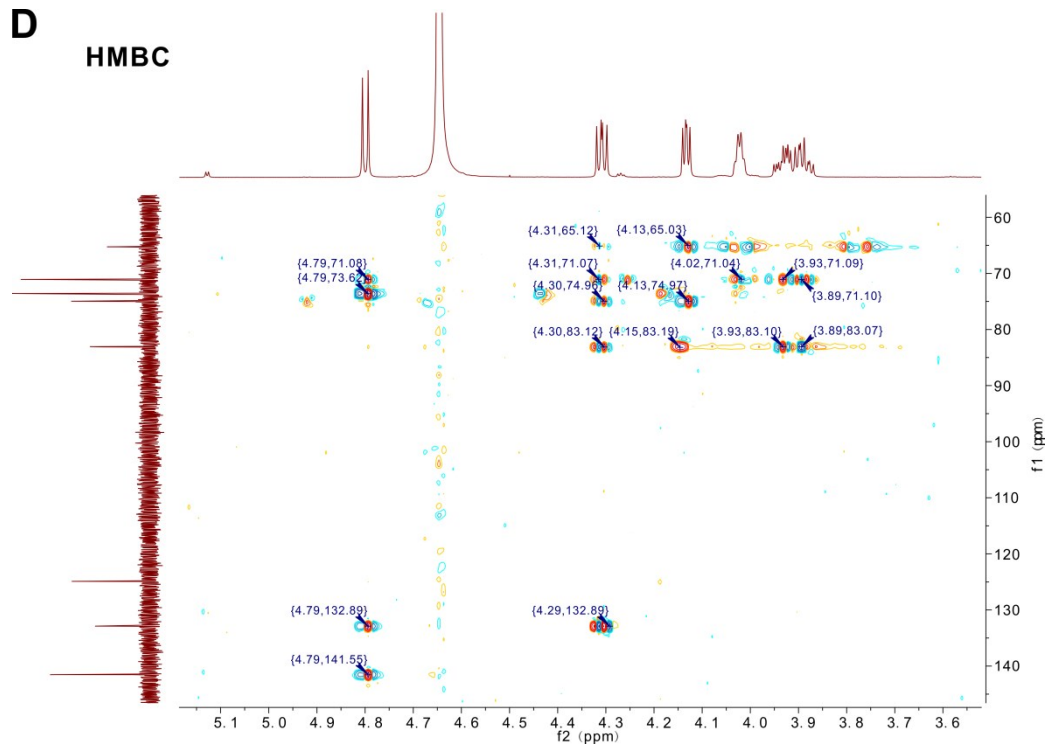


Figure S5D. HMBC analysis of 19

HMBC spectrum of **19** (600 MHz, D_2O).

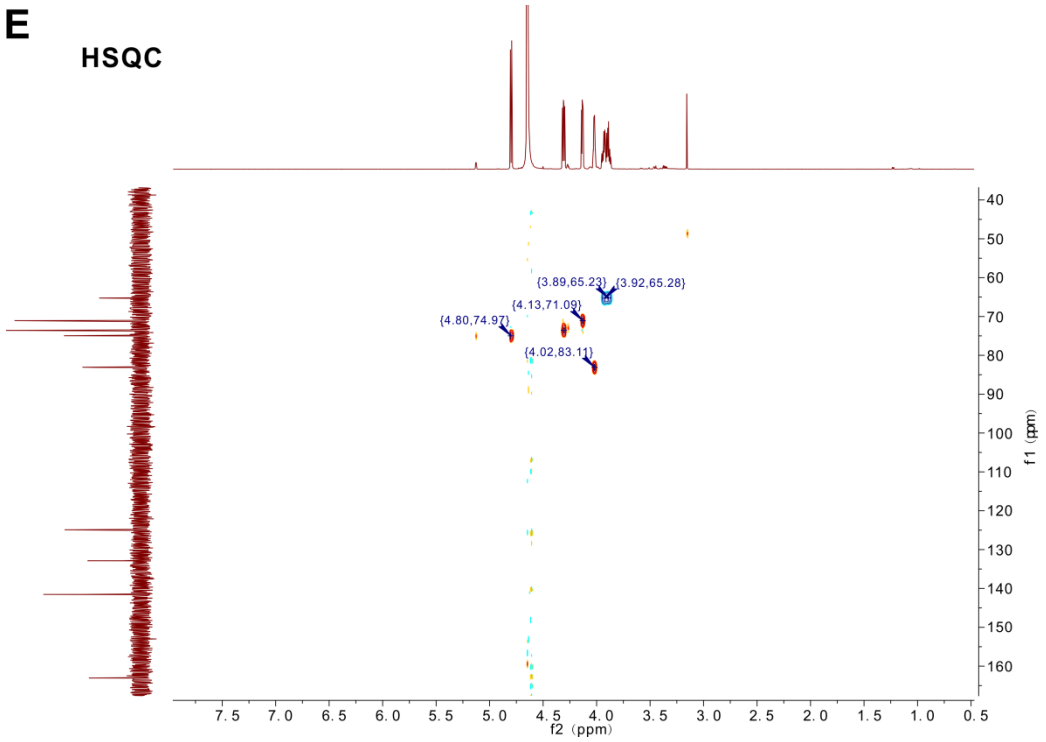


Figure S5E. HSQC analysis of 19

HSQC spectrum of **19** (600 MHz, D_2O).

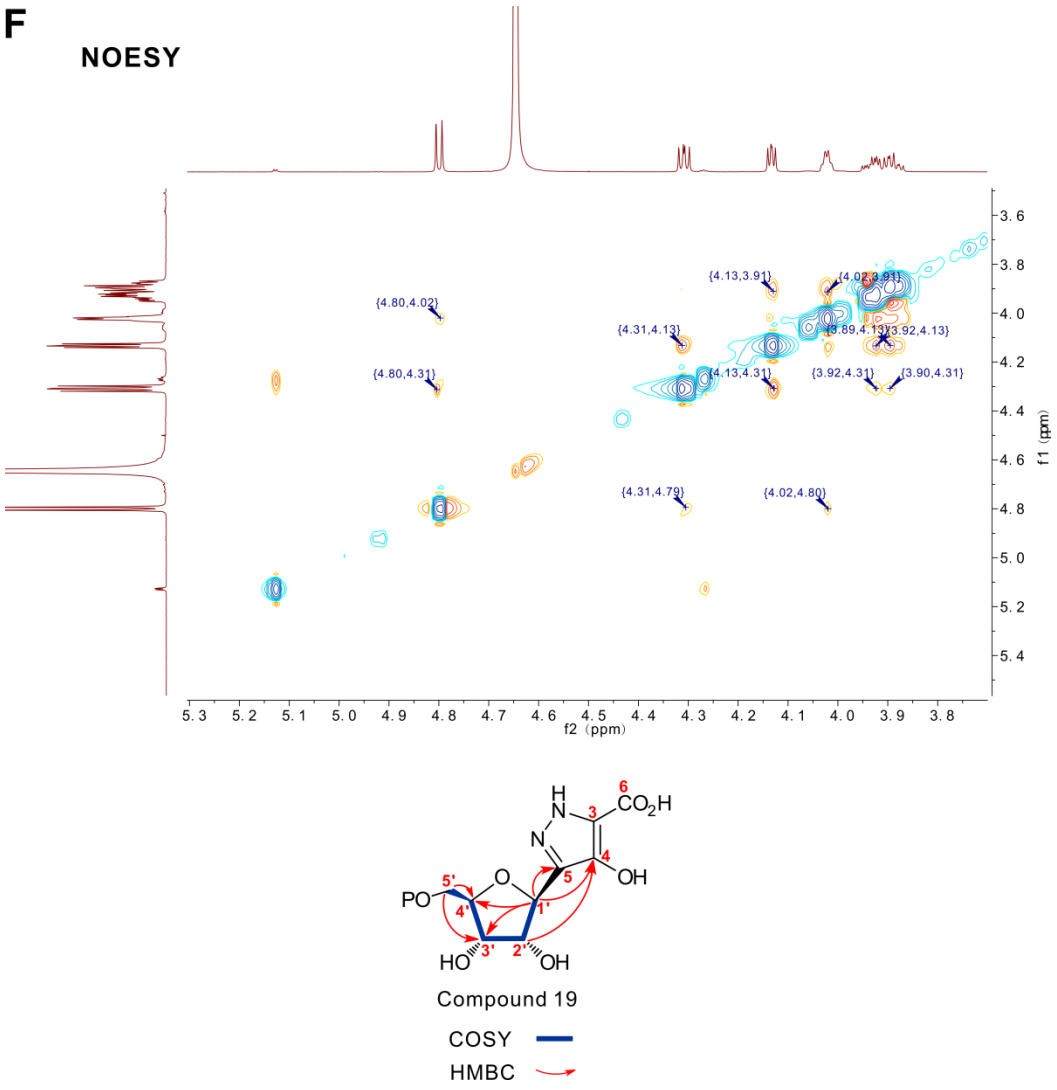


Figure S5F. NOESY analysis of 19

NOESY spectrum of **19** (600 MHz, D_2O).

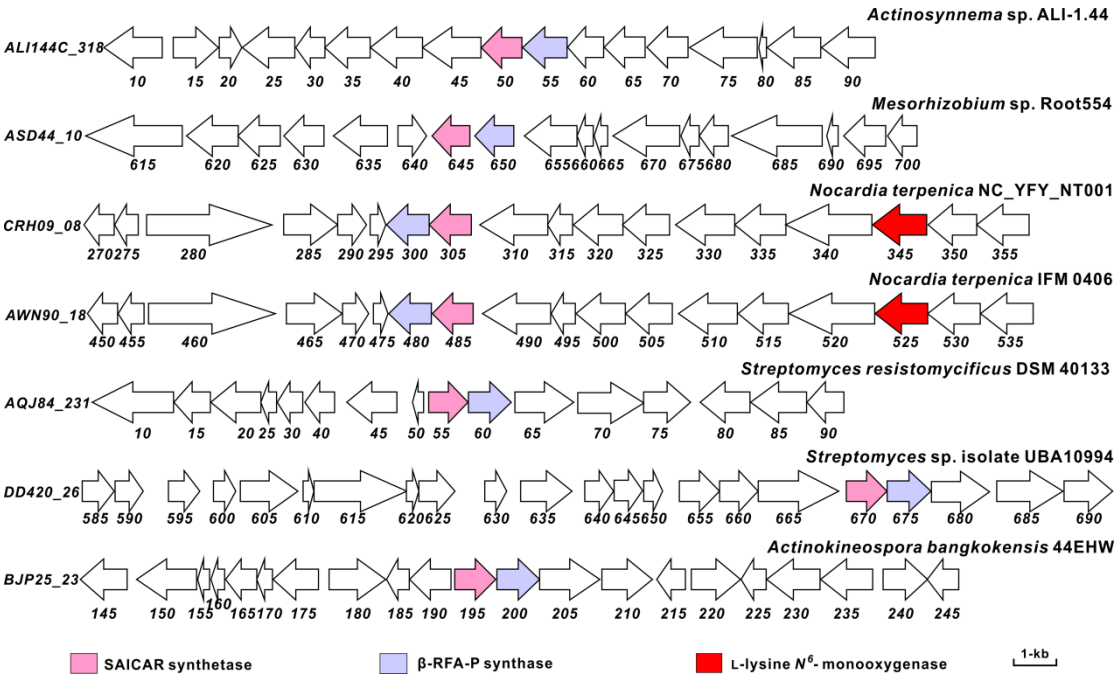


Figure S6. Target-oriented genome mining of the biosynthetic gene clusters for *C*-nucleoside antibiotics

The gene clusters for the potential *C*-nucleoside antibiotics were obtained using PrfT as enzyme probe. The proposed gene functions with highlighted colors were listed at the bottom. SAICAR, phosphoribosylaminoimidazole-succinocarboxamide.

2. Supplementary Tables.

Table S1. Deduced functions of the open reading frames in the *foc* and *for-cof* gene clusters*

Protein (aa) Identities, positives (%)	Protein Function	Homolog, Origin	Identities, positives (%)	Accession no.
ForX (295) FocX (288) 55, 66	phosphofructokinase	OV450_4865, *Actinobacteria bacterium* OV450	58, 71	KPI01287
ForW (931)	FAD-binding oxidoreductase	MPHLEI_25011, *Mycobacterium phlei* RIVM601174	69, 77	EID09270
ForC (333) FocC (294) 58, 68	SAICAR synthetase	OV450_4874, *Actinobacteria bacterium* OV450	67, 75	KPI01296
ForV (185)	FMN reductase (NADPH)	BIV23_36585*, Streptomyces* sp. MUSC 1	73, 83	OIJ93994
ForU (390) FocU (407) 73, 84	phthalate 4,5-dioxygenase	ADK37_11500, *S. resistomycificus*	87, 92	KOG37051
ForT (341) FocT (341) 69, 77	*C*-glycosyltransferase	Orf61, *Nocardia terpenica* IFM 0406	64,75	AJO72754
ForB (474) FocB (480) 66, 80	adenylosuccinate lyase	ADL05_15155, *Nocardiopsis* sp. NRRL B-16309	78, 86	KOX15518
ForS (529) FocS (528) 62, 69	fumarate reductase	CRH09_08310, *Nocardia terpenica*	64, 72	ATL71617
ForR (374) FocR (351) 49, 60	glycine/D-amino acid oxidase	Orf73, *Nocardia terpenica* IFM 0406	54, 65	AJO72766
ForQ (396) FocQ (387) 63, 75	amidohydrolase	A5717_10410, *Mycolicibacterium porcinum* ACS3670	62, 76	OCB14542
ForP (187) FocP (205) 65, 73	NUDIX hydrolase	BOQ63_27930, *S. viridifaciens* DSM 40239	82, 87	OJH68722
ForO (736)	S9 family peptidase	IX27_25180, *Streptomyces* sp. JS01	61, 70	KFK86867
ForM (436) FocM (413) 66, 79	phosphoribosylglycinamide synthetase	Orf72, *Nocardia terpenica* IFM 0406	69, 80	AJO72765
ForL (384) FocL (393) 57, 68	saccharopine dehydrogenase	Orf71, *Nocardia terpenica* IFM 0406	58, 67	AJO72764
ForK (440) FocK (436) 62, 72	L-lysine N6-monooxygenase	Spb38, *Streptomyces* sp. SoC090715LN-17	41, 55	BAW27702
ForJ (677) FocJ (668) 59, 68	methionine-tRNA ligase	Spb40, *Streptomyces* sp. SoC090715LN-17	36, 50	BAW27704
ForI (423) FocI (433)	aminotransferase	BJP25_23260, *Actinokineospora bangkokensis* 44EHW	64, 78	OLR92323

64, 78				
ForG (230)	RNA polymerase sigma-70 factor	GA0070616_0590, *Micromonospora nigra* DSM 43818	52, 67	SCL14959
ForF (333) FocF (334) 68, 81	phosphoglycerate dehydrogenase	BAL199_19391, *Alpha proteobacterium* BAL199	47, 63	EDP63049
ForA (423) FocA (426) 68, 79	adenylosuccinate synthetase	BJP25_23230, *Actinokineospora bangkokensis*	68, 77	OLR92236
ForH (196) FocH (187) 66, 73	AICAR transformylase	BJP25_23225, *Actinokineospora bangkokensis*	63, 72	OLR92235
ForE (342) FocE (312) 49, 60	putative monooxygenase	ASC89_11735, *Devosia* sp. Root413D1	41, 52	KQW78972
ForD (439) FocD (444) 78, 86	putative monooxygenase	Orf75, *Nocardia terpenica* IFM 0406	79, 86	AJO72768
CofC (404) Foc5 (394) 61, 74	MFS transporter	NbrT6, *Nocardia terpenica* IFM 0406	60, 74	AJO72759
Cof25 (335)	metal dependent phosphohydrolase	PenF, *S. antibioticus*	37, 49	AKA87335
Cof26 (452)	GntR family transcriptional regulator	SAZ_09915, *S. albulus* ZPM	71, 77	AKA02703
CofB (239) Foc1 (242) 61, 73	SAICAR synthetase	PenC, *S. antibioticus*	62, 75	AKA87338
CofA (233) Foc2 (231) 52, 66	short chain dehydrognase	PenB, *S. antibioticus*	49, 63	AKA87339
Cof29 (140)	hypothetic protein	GA0070616_0574, *Micromonospora nigra* DSM 43818	41, 52	SCL14901
Cof30 (264) Foc3 (257) 33, 42	HAD family phosphatase	Acel_2103, *Acidothermus cellulolyticus* 11B	35, 49	ABK53875
FocG (275)	transcriptional regulator	BJP25_23295, *Actinokineospora bangkokensis*	60,70	OLR92324
Foc4 (290)	ATP phosphoribosyltransferase	BJP25_23245, *Actinokineospora bangkokensis*	63,73	OLR92239
FocN (233)	putative HAD superfamily hydrolase	BJP25_23215, *Actinokineospora bangkokensis*	71, 81	OLR92233

*The *for* genes are from *S. kaniharaensis* ATCC 21070, while the *foc* genes are from *Nocardia interforma* ATCC 21072. Taking ForX as example, it shares 55% identities (66% positives) to FocX, and it shows 58% identities (71% positives) to OV450_4865 of *Actinobacteria bacterium* OV450.

Table S2A. Crystal data and structure refinement for ADCP

Identification code	x_a
Empirical formula	C5 H7 N3 O5
Formula weight	189.14
Temperature	200(2) K
Wavelength	1.54178 Å
Crystal system	Monoclinic
Space group	$P2_1/n$
Unit cell dimensions	a = 12.8777(6) Å $\alpha = 90°$
	b = 4.2342(2) Å $\beta = 103.553(2)°$..
	c = 14.2564(6) Å $\gamma = 90°$..
Volume	755.71(6) $Å^3$
Z	4
Density (calculated)	1.662 Mg/m^3
Absorption coefficient	1.320 mm^{-1}
F(000)	392
Crystal size	0.140 x 0.120 x 0.060 mm^3
Theta range for data collection	4.167 to 63.017°.
Index ranges	-14<=h<=12, -4<=k<=3, -16<=l<=16
Reflections collected	4581
Independent reflections	1213 [R(int) = 0.0217]
Completeness to theta = 63.017°	99.7 %
Absorption correction	Semi-empirical from equivalents
Max. and min. transmission	0.7530 and 0.5406
Refinement method	Full-matrix least-squares on F^2
Data / restraints / parameters	1213 / 0 / 146
Goodness-of-fit on F^2	1.108
Final R indices [I>2sigma(I)]	R1 = 0.0288, wR2 = 0.0729
R indices (all data)	R1 = 0.0300, wR2 = 0.0737
Extinction coefficient	n/a
Largest diff. peak and hole	0.179 and -0.240 $e.Å^{-3}$

Table S2B. Crystal data and structure refinement for DCOP

Identification code	x_a
Empirical formula	C5 H3 N2 O7
Formula weight	203.09
Temperature	296(2) K
Wavelength	1.54178 Å
Crystal system	Triclinic
Space group	P-1
Unit cell dimensions	a = 7.9342(7) Å α= 98.253(5)°
	b = 8.2041(8) Å β= 95.051(4)°
	c = 12.7997(12) Å γ = 97.782(4)°.
Volume	812.03(13) $Å^3$
Z	4
Density (calculated)	1.661 Mg/m^3
Absorption coefficient	1.441 mm^{-1}
F(000)	412
Crystal size	0.120 x 0.080 x 0.060 mm^3
Theta range for data collection	3.510 to 61.155°.
Index ranges	-9<=h<=9, -9<=k<=9, -10<=l<=14
Reflections collected	5416
Independent reflections	2365 [R(int) = 0.0464]
Completeness to theta = 61.155°	95.0 %
Absorption correction	Semi-empirical from equivalents
Max. and min. transmission	0.7522 and 0.5691
Refinement method	Full-matrix least-squares on F^2
Data / restraints / parameters	2365 / 1 / 260
Goodness-of-fit on F^2	1.135
Final R indices [I>2sigma(I)]	R1 = 0.0991, wR2 = 0.2719
R indices (all data)	R1 = 0.1026, wR2 = 0.2753
Extinction coefficient	n/a
Largest diff. peak and hole	0.884 and -0.573 $e.Å^{-3}$

Table S3. NMR data for 19

	δC	δH, *J*(Hz)
C-3	141.53	
C-4	132.87	
C-5	124.90	
C-6	163.06	
C-1’	74.96	4.80, d, *J*=7.24
C-2’	73.59	4.31, dd, *J*=5.40, 7.24
C-3’	71.07	4.13, dd, *J*=3.93, 5.22
C-4’	83.11	4.02, m
C-5’	65.27	3.93, dd, *J*=5.84, 11.37
		3.89, dd, *J*=6.45, 11.32