Supporting Information

Bioinformatic, Structural, and Biochemical Analysis leads to the Discovery of Novel Isonitrilases and Decodes their Substrate Selectivity

Tyler Hostetler, Tzu-Yu Chen and Wei-chen Chang*

Department of Chemistry, North Carolina State University, Raleigh, NC 27695, United States

Contents:

Cloning, protein expression and purification of isonitrilases and Srug variants

The DNA sequence for each Fe/2OG isonitrilase gene was codon-optimized for overexpression in *E. coli*, synthesized, and inserted into pET-28a between the *NdeI* and *XhoI* sites, providing each gene with an *N*-terminal His_6-tag. The codon optimized gene sequences are shown below:

Srug

ATGACATTAAATGTAAAAGGAGAAGGGCTAGGCGCGCAAATTACCGGGATCGATCC
GGAGGACCTGGACAGCATCCGTGCGGACGAGATCCGCGACATCGTGTACGAGAAC
AAGTTGGTCGTGCTGAAAGACGTTCGTCCGAGCGCAGAGCAATTTCTGCAGCTTG
GTCGCATCCTGGGCGAAATTGTGCCGTATTACGAGCCGATTTATCACCACAAGGAC
CACCCGGAAATTTTTGTGAGCAGCACCGAACAGGGTCAGGGAGTTCCGCGTACCG
GTGCGTTCTGGCATATCGACTATCAGTTCATGCCGAAACCGTTTGCATTTAGCATGG
TACTGCCGCTGGCTGTCCCGGGCCGTGATCGCGGTACCTACTTCATCGACTTGAAT
AAAGTTTGGGAATCCCTGCCGGCTGATTTGCAAGCGAAGGCCCGTGGTACGCGTT
CTATCCATTCTCCGCGCCGTTACATTAAAATTAGACCCTCGGATGTCTATCGTCCGAT
TGGCGAGATCTTGGCGGAGATCGAGGAAACCACCCCGCCTCAGACCTGGCCGAC
CGTTATTAAGCACCCGAAAACCGGCCAAGAAATCTTATATATTTGCGAAGCGGGTAC
GATGACCATCGAGGACGGCAATGGCAACGCGTTAGACCCGAGCCTGCTTCAAACC
CTGCTGGAAGCTACTGGCCAGCTGGACCCGGATTACGCCTCCCCGTTCATTCATGC
ACAGCACTACGAGGTTGGTGATATCGTGCTGTGGGATAACCGTGCGCTGGTTCATC
GTGCAAAGCACGGCACTGCCCCAGGTACACTCACGACCTACCGCCTGACCATGCT
GGATGGTCTGGCGACTCCAGGTTTGGCTGCGTAA

Rfas

ATGCTAGTATCACCCCAACCAAATGTTACAAGAGGCGTTCAGGTTCACGGGTTTCAT
CCGAGCACCGCAACCGAATCGGAAGTTTCTACGATCCTGGAGCACATTTACCGTGA
CAAAATCGTACTCCTGAAAGAACAGGAGTTGGATCACGACGAATTCGTGGCCCTGG
GCGCACTTTTGGGCACTCCGGCGGTCTACTACGAGTCCATGTATCATCTGCCGGGT
CACCCCAAGATTTTCGTGAGCTCCAACCACCCATCTGGTGGTAGCCGTATAGGCGT
GCCGAAGACCGGCAAATTCTGGCACAGCGACTATCAATTTATGCCGGACCCGTTTG
CAATCACCCTGATTTACCCGCAAATTCTGCCGGAACGTAATCGCGGCACGTACTTCA
TCGATATGGGTGAGGCGTTTCAGGATCTGGACCCGCGTATCAAAGATGTGGTTGTT
TTGTCTAACAGCTGGCATAGCTCTCGTCGTTATGTTAAGATCCGTCCGTCCGACGTG
TACCGCCCACTGGGTGAGATCGTGGCTGAGGTGGAAGCTAACACCCCTCCGCAAA
TGCGTCCGACCGTGATTGAGCACCCGGTTACCGGTGAGGCCGTCCTTTATATCAGC
GAGGCGTTCACGTACGCGATCGAAGACCGTGACGGCGAGGCTCTGCCGGCGGAA
CTGCTGACCGAATTGTTTGGTACTAGCGGTCAGCTGGACAGCACCTTTACGCATCC
AAATATTTTCCTGCAGACCTATGAGCCGGGTGACATCCTATTATGGGATAACCGCAG
CCTGGTTCACCGCGCCCTGCACACCAGCAGTAACGAACCGGTCGAGTCACATCGC
ATTACCGTTCACGATGGTCATCCGTTGAATCGTGTTCGTCATGCGGGTGTGGATCG
CGGCGAAGCGGCGGTGAAGTAA

Nfla

ATGAGGATAGACACTAGAACAGGGGATGGACGCGGGATCGAGGTGCGCGGTTTTT
CCGCGGCCGCCACCACCGACGACATGGCGCTGCTGCGTCAACACGTTTATTACGA
CAAAATTGTTGTTCTGAAAGAGCAGTCCCTGGGCATTGAAGAGTTCGTGACACTGG
GCAGTGCTTTAGGCAAGCCGGTGGCATACTACGAGCCGATGTATCATCACCCGGAT
AGCGAGTACGTTTTCGTTAGCAGCAATGTTAATCGCGACACGGGTCGCGTGGGTGT
ACCGAAGACCGGCGCTTTCTGGCATTCCGACTATCAGTTTATGCCGGAGCCGTTTG
CGATCACCTGTTTTTATCCGCAACGCCTGCCGGCTGCGGGTCGTGGCACGTACTTC
ATCGATATGGCGCAGGCATATAGACGTCTGTCGCCGCGTTTGCGTGACGCCATCGA
GAACACCTTCTGCCTGCACTCTGCCCGTCGTTATGTGAAGATCCGTCCGGAAGACG
TTTTTCGTCCGCTCGGCGAAGTCCTGGCGGAAGTCGAGGACCGTACCCCGCCTCA
ACGTAGACCGACCGTCCTTCACCACCCGGTGACCAAAGAGCCACTCCTGTACGTG
AGCGAAGCATTCACTTACGCGATTCAGGATGCGGCTGGCACGGCGTTACCGGGTA
CTCTGCTATCTGACTTGTTGGCGGAGAGCGGTCAACTGGATGATACGTACAGCCAT
CCCAACATCTTCCTGCAGCCGTATGAACCAGGTGATCTGGTGCTGTGGGACAACCG
CACCCTGATTCATCGTGCACTGCACAACCCGGGTAATGATTTGACCGAAAGCTATC
GTGTTACCGTTGTAGATGAATACCCGTTGACCTTGGAAGAAGCGGCATAA

Cpro

ATGGCAGTAGAAATAACACCCGCTCAAAATGGCAATATGGGTTCCGCGGTTACGGG
TTTCACCGTCGCCGGGGCCACCGCAGAAGATTATGCAGCGCTGCGTCAGGCGGTG
TATCGTGACCGTATTATTGTACTTAAGGACCAGCAAGATATCACCCCGGAGGAATTC
GTGGAGCTGGGCCGTCACTTTGGTACCATTGTGCCGTATTACGAGGAGACATACCA
CCACCCGGATCACCCGGAGATCTTTGTTTCTTCCAACCTGAGCGCGGACGGCCGG
CCGTTGGGCGTTCCGCGTACCGGTCGCTTCTGGCATGCCGACTACATGTTTGCGC
AAGAGCCGCTTAGTGTTACCGTCTTTGCCCAAAAAATCCTGCCACCGGGTCGCAGG
GGTACGTACTTCATTGATATGGTTAAAGCATTTCGTGACCTGCCGGAGAGCTTGAAG
CAGGAGGCCCGTGTAACTCGCGCCTCGCACAGCGTTCGTCGTTTCTTCAAGATCC
GCCCAGGCGACGTTTTCCGCCCTCTGGGTGACCTGATCAGAGAAGTTGAGAAAAT
CAGCCCGCCTGCAGTGCACAACACCGTTGTTACCCATCCGGTGACGGGTGAAGAG
ATCTTATATGTTTCTGAAGGTTTCACTGACCACCTGATCGGCGCGGAACGTGACAGC
CTGCTGCGCGACTTGTTGGAAGCTAGCGGTCAGCTGGATGAAACCTTTACCGACC
CGCATATTGTGTTGCACACCTACGAACCAGGTGAAATCGTGATTTGGGATAACCGTG
CTCTGGTGCATTGCGCGCTCCACGCGACCGATCCGTCCGCTCCGGTGATGAGCCA
TCGTGTCACGACCGTCGATGGTCATCCGTTTGATGCGAATCCGACCCCGGCGGGC
CGCGGCGCTGGCACCGGCTCTGCGAAATAA

Hneap

ATGAAAGGTCTGGAAATGAAAAACAACGTTGCGGTTATCTCTCGTCAGGTTACCGAA
ATGACCGACGACGAAATCCACAACCTGAAGAAAATCGTTTTCGACTCTGGTATCGTT
GTTCTGAAAGCGCAGAACGCGACCGCATCTGACTTCGTTGATTTCGGTCGTCGTAT
CGGTGAACTGTCTCCGTACTACGAAGAAATGTACCACCACCCGAACCACAAAGAAC

TGTTCGTTTCTTCTAACGTTCAGACCGATGGTAAAGTTATCGGTGTTCCGCGTACCG
GTAAATTCTGGCACGCGGATTACGCATTCATGGCTAAACCGTTCGCGTTCACCATCA
CCTACCCGCAGGTTGTTAGCTCTCAGGAACGTGGTACTTACTTCATCGATATGGCTT
CTGCTTACGAACGTCTGTCTCCGGAAATGAAACGTAAAATCGAAGGCGGTGTTGGC
ACCCACTCTGTTCGTCGTTACTTCAAAATCCGTCCGACCGATGTTTACCGTCCGATC
TCTGAAATCCTGCACGAAATCGACGCTAAAACCCCGACCGTTACCCACCCGCTGGT
TGTTAACCACCCGGTTACCGGCGCGAAAATCCTGTACGTTTCCCGTGGTTTCACCG
AAACCATCTCTCTGAAAGATGACGACCTGGACGCTGACGAAGTTCTGAAAGACCTG
CTGGTTGAATCTGGTCAGTCTGACGACACCTTCACCCACCCGGACATCCGTCAGAT
CAACATCAACGAAGGTGACATCTTCCTGTGGGATAACCGTCGTTACGTTCACCACG
CGAAACACAACGACAAAGTTGAACCGACCAAAACCTACCGTCTGACCGCTTACGAC
GGTCTGCCGTTCTCTGCGGAAATCGACTTCGCGCTGGACTCTGTTAAAGAAGTTGG
TCTGGTTTAA

Mlep

ATGACATTACACGTAAAAGGAGAAGGGCTAGGCGCGCAGGTCACCGGTGTGGACC
CGAAGAACCTGGACAACATTAGCACTGCCGAGATCCGCAAGATCGTCTATGTTAATA
AACTCGTGGTGTTGAAGAACGTGCACCCGACCCCGGAAGAGTTCATCAAGCTGGG
TCGTATTATTGGTGAAATCGTACCTTATTACGAGCCGATTTACCGCCATAAAGACTAC
CCGGAGATTTTTGTTTCTTCCACCGAAGAGGGCCAAGGTGTTCCGAATACCGGCGC
GTTCTGGCATGTTGACTATATCTTCATGCCCAAGCCGTTTGCATTCAGCATGACCCT
TCCGCTGGCTATGCCGGGTAACGATCGTGGCACTCACTTTATTGATCTGAGCCAGG
CATGGGAATCTCTGCCGTCAACGACCAAAAACAGCGTTCGTGGTACGTACAGCACC
CACTCGCCGAGACGTTATATCAAGATTCGTCCGAGCGATGTTTACCGTCCGATCGG
CGAAGTGTTGGCGGAAATTGAAGAGGTTACCCCGCCACAGAAATGGCCGACGGTG
ATTAAGCACCCGAAAACCGGCCAAGAGATCTTATACATCTGCGAAGCGGCGACCGT
CTCCGTAGAAGGTAAAAACGGGAACTTGTTGGACCCGATGGTTCTGCAAGAGCTG
CTGACCGCGAGCGGTCAGCTGGACCCGGATTGTAAAAGCCTGTTGATCCATACCCA
GCATTATGAGGTGGGCGATGTGATCCTGTGGGATAATCGTGCGCTGGTGCACCGC
GCTAAACACTCCACCGTTAGTGGCACGCTGATTACTTACCGCCTGACCCTTCTGGA
CGGTCTGAAGACCCCAGGTTATGCCGCATAA

The expression vectors were transformed into *E. coli* BL21 (DE3) cells (New England Biolabs, MA). The transformed cells were grown at 37 °C in Terrific Broth (TB) medium supplemented with 50 μg/mL kanamycin added. After the optical density at 600 nm (OD_{600}) reached ca. 0.6, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to the culture to a final concentration of 0.5 mM. The cells were incubated for another 16 h at 18 °C and then were harvested by centrifugation (9,000 x g, 25 min, 4 °C). The cell pellets were collected and stored at –20 °C. Cell pellets were resuspended in lysis buffer (100 mM Tris-HCl, 10 mM imidazole, pH 7.6). After sonication and centrifugation (34,000 x g, 30 min, 6 °C), the supernatant was loaded onto a nickel-nitrilotriacetic acid (Ni-NTA) agarose column that was equilibrated with lysis buffer. The column was washed using lysis buffer (ca. 9 column volumes). The target proteins were eluted with elution buffer

(100 mM Tris-HCl, 250 mM imidazole, pH 7.6, 6 column volumes). The pooled protein fractions were concentrated using a protein concentrator (30-K Pall®) and dialyzed at 4 °C first against 100 mM Tris-HCl and 5 mM ethylenediaminetetraacetic acid (EDTA, pH 7.6) followed by 100 mM Tris-HCl (pH 7.6) for 12 h twice. After dialysis, the proteins were aliquoted and stored at –80 °C. Protein concentration was determined by UV absorption at 280 nm using their respective calculated molar absorptivity (http://ca.expasy.org). SDS-PAGE with Coomassie staining of the purified proteins is shown in Figure S17.

To improve the protein solubility, a vector encoding the chaperone protein (GroEL, Takara®) was co-transformed with Srug variants and Mlep. Similar procedure was applied, while additional 35 μg/mL chloramphenicol and 0.1% arabinose were supplemented in Terrific Broth (TB) medium.

Bioinformatic workflow using BioSynthNexus

A sequence similarity network (SSN) is constructed from a query gene sequence using EFI-EST[1,2] (https://efi.igb.illinois.edu/efi-est/) and subsequently processed by EFI-GNT[1,2] (https://efi.igb.illinois.edu/efi-gnt/) to generate a genome neighborhood network (GNN). The GNN can be visualized using the genome neighborhood diagram (EFI-GND) available on EFI website. The GNN consists of a list of genome neighborhoods, each containing a parent gene (a homolog of the initial query) and its neighboring genes. A representative example is shown in Figure S1, which genes from the same protein family (Pfam) are represented in the same color.

BioSynthNexus is a custom Python program designed to refine potential homologs within a given GNN by efficiently filtering Pfam IDs in the neighborhoods. It also allows remote access to the UniProt database to retrieve gene information for neighborhoods of interest. Installation instructions for BioSynthNexus can be found on its GitHub repository (https://github.com/Tyler-Hostetler/BioSynthNexus). The representative example shown in Figure S1 will be used as an example to illustrate the key features of BioSynthNexus. These key features are listed below, with examples and step-by-step instructions provided in Figure S2–S6.

1. Retrieval of gene information from the UniProt database according to the selected output type, such as a FASTA format of protein sequences, protein/genome accession IDs in GenBank, and ORF names in the corresponding genome (Figure S2).
2. Retrieval of genome neighborhoods containing neighboring gene(s) within the specific Pfam ID(s) (Figure S3).
3. Retrieval of accession IDs for neighboring genes within a designated Pfam ID from the given genome neighborhoods (Figure S4).
4. Retrieval of all Pfam IDs for neighboring genes from the given genome neighborhoods (Figure S5).

5. Retrieval of genome neighborhood information from the given PFam ID(s) (Figure S6).

Bioinformatic analysis of potential Fe/2OG isonitrilases

The sequence similarity network (SSN) and subsequent genome neighborhood network (GNN) of ScoE was generated via EFI-EST[1,2] (https://efi.igb.illinois.edu/efi-est/) and EFI-GNT[1,2] (https://efi.igb.illinois.edu/efi-gnt/), respectively. Initially, over 9,000 biosynthetic gene clusters (BGCs) were returned. To avoid tedious *in silico* analysis, a custom python program, BioSynthNexus, was developed. In this study, BioSynthNexus was used to filter the BGCs by the protein family domain (Pfam) of the dual functional thioesterase (PF10862). Subsequently, the number of Fe/2OG isonitrilases is down to 365. The sequences of potential Fe/2OG isonitrilases were exported in a FASTA format, and the SSN was constructed and shown in Figure 2. The SSN is visualized with Cytoscape[3] where edges indicate at least 60% identity. A multiple sequence alignment (MSA) is constructed with T-Coffee[4] (Figure S7).

In vitro assays of Fe/2OG isonitrilases and Srug variants

Liquid chromatography (LC) with detection by mass spectrometry (MS) was utilized an Agilent Technologies (Santa Clara, CA) 1290 Infinity II system coupled to an Agilent Technologies 6530 Quadrupole Time of Flight (Q-TOF) mass spectrometer. The enzymatic reactions were chromatographed on an Agilent Zorbax Extend-C18 column (4.6 x 50 mm, 1.8 µm). Solvent A contained 0.1% formic acid in H_2O, and solvent B was acetonitrile. The Agilent MassHunter software package was used for data collection and analysis. Samples with tetrazine derivatization were analyzed with the following gradient program: 0–3 min 80% A isocratic, 3–4.5 min 80% to 72% A, 4.5–5 min 72% A isocratic, 5–5.5 min 72% to 80% A, 5.5–6 min 80% A isocratic. The flow rate was 0.4 mL/min. Mass spectra were monitored with electrospray ionization in positive mode (ESI^+). Samples without tetrazine derivatization were analyzed the following gradient program: 0–1 min 75% A isocratic, 1–3 min 75% to 20% A, 3–5 min 20% A isocratic, 5–6 min 20% to 75% A, 6–8 min 75% A isocratic. The flow rate was 0.4 mL/min. Mass spectra were monitored with electrospray ionization in negative mode (ESI^-).

Enzymatic reactions containing 100 µL of 50mM Tris-HCl (pH 7.6) with final concentrations of 0.2 mM enzyme, 0.2 mM Fe(II), 2 mM 2-oxoglutarate, and 1 mM substrate (**6** with odd alkyl chain n=1–11) were exposed to air at 4 °C for 12 h. The reactions were quenched by adding 100 µL of 2.5 mM 3,6-di-2-pyrid-yl-1,2,4,5-tetrazine in methanol. After incubating the reaction mixtures at 42 °C for 30 min, the samples were centrifuged at 19,000 x g for 30 min to precipitate the protein before LC-MS analysis. The common pyrazole product **8** was quantified (Figure 3G, 4, 5B, S13, S14, and S15). For

samples without derivatization, the reaction mixtures were quenched with an equal amount of methanol and analyzed using LC-MS (Figure S12).

Site-directed mutagenesis

PCR samples containing a final concentration of 0.5 µM forward primer, 0.5 µM reverse primer, 20 pg DNA template, and 12.5 µL New England Biolabs (NEB) Q5® High-Fidelity 2X Master Mix were prepared in a total volume of 25 µL. The PCR program was: 98 °C for 30s, 28 cycles at 98 °C for 10s, T_m for 30s, 72 °C for 2.15 min, then 72 °C for 4 min. The primers and the corresponding T_m used in this paper are listed in Table S2.

The linear PCR products were then re-circularized with NEB KLD Mix, comprising of 1 µL T4 DNA Kinase, 1 µL T4 DNA Ligase, 1 µL 10X T4 DNA Ligase Buffer, 1 µL DpnI, and 1 µL PCR product prepared in a total volume of 10 µL. The mixtures were incubated at room temperature for 1 hour followed by 37 °C for 20 min. Entire KLD reaction contents were transformed into *E. coli* DH10β cells. Mini-prep cultures for each variant were cultivated overnight. The DNA was then extracted and purified with a plasmid miniprep kit (Zymo Research), and the mutations confirmed with Sanger sequencing.

Predicted structure and substrate docking of Srug

The heptyl substrate **6** (n=7) was first prepared as a MOL2 molecular structure file in Avogadro.[5] The protonation state of **6** was adjusted based on a pH 7.4 and its conformational energy was minimized with the MMFF94s forcefield. The structure of Srug was generated using AlphaFold3.[6] The substrate and protein structure were then prepared as PDBQT molecular structure files using AutoDock Tools.[7] With the predicted Srug structure, polar hydrogens were added and Kollman charges were computed. A gridbox of 10x10x10 Å was used to cover the entire binding site of Srug. The substrate and protein PDBQT molecular structures were then docked using AutoDock Vina (version 1.1.2)[8,9], and the result was visualized with ChimeraX.[10]

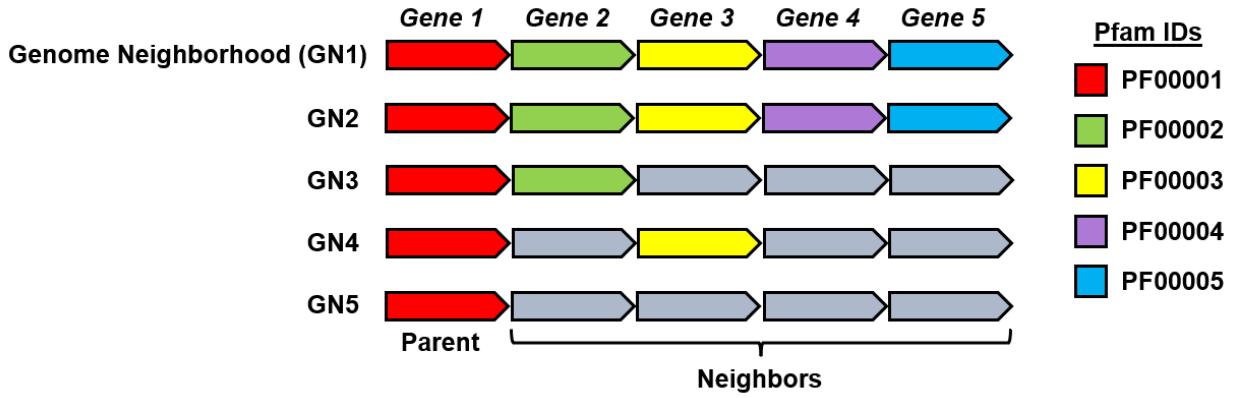


Figure S1. Representative biosynthetic gene clusters (BGCs) used for the following BioSynthNexus demonstrations. Examples and step-by-step instructions are detailed in the Figure S2–S6. In all examples, Gene 1 is used as a query to generate a sequence similarity network (SSN) and a subsequent genome neighborhood network (GNN) via the EFI-EST and EFI-GNT websites[1,2], respectively. A list of genome neighborhoods (GNs) can be visualized on EFI-GNT website. Genes from the same protein family (Pfam) are represented in the same color, while genes not belonging to PF00001–PF00005 are shown in gray for clarity.

1. Select UniProt for Request Type.
2. Select desired Output Type to request from UniProt (e.g. FASTA).

Enter Accession ID(s) in Input Field.

1. Click Search.
2. Displayed in the output field will be data retrieved from UniProt. For this example, the FASTA-formatted sequences are shown.

Figure S2. Retrieval of gene information from the UniProt database according to the selected the output type. In this example, the FASTA format of sequences is displayed as results, which can be used for further multiple sequence alignment.

Note: Other output types, such as protein accession ID in GenBank, genome accession ID in GenBank, and ORF names in the corresponding genome, can be selected for different purposes. This feature does not require uploading a GNN sqlite file.

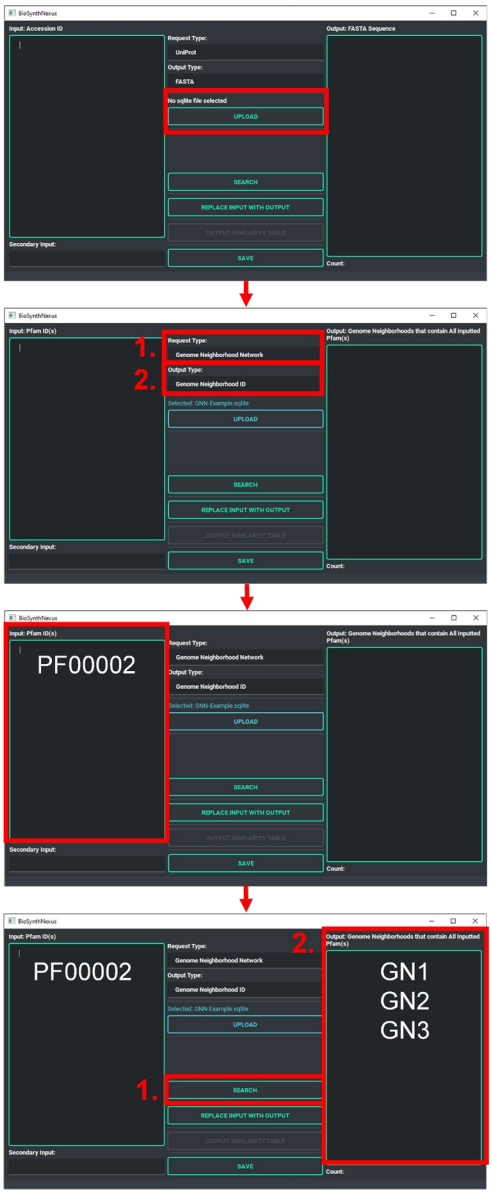
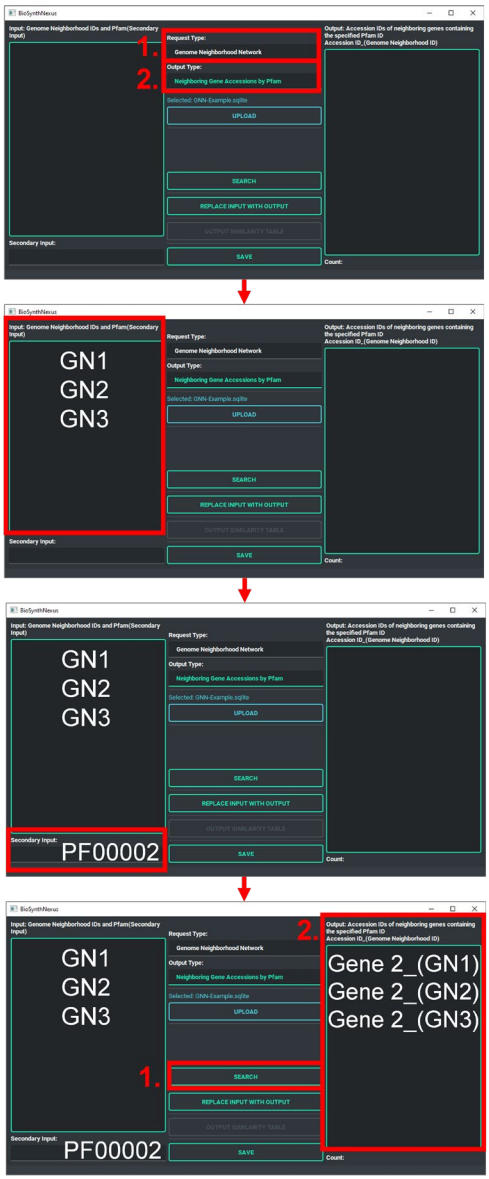


Figure S3. Retrieval of genome neighborhoods containing neighboring gene(s) within the specific Pfam ID(s). In this example, GN1–GN3 are displayed as output results because these genome neighborhoods include a neighboring gene from PF00002 (input) (Figure S1). This strategy was employed in this study.

Note: The accession IDs of Gene 1 from GN1–GN3 will be as displayed as results if "Parent Accession ID" is selected as the output type.



1. Select Genome Neighborhood Network for Request Type.
2. Select Neighboring Gene Accessions by Pfam.

Enter Genome Neighborhood ID(s) into Input Field.

Enter a single neighboring gene Pfam ID into Secondary Input.

1. Click Search.
2. Displayed in the Output Field are the gene neighbors (Gene 2) that contain the designated Pfam ID (PF00002) for each Genome Neighborhood searched. Formatted as: 'Gene Accession_(Genome Neighborhood ID)'

Figure S4. Retrieval of accession IDs for neighboring genes within a designated Pfam ID from the given genome neighborhoods. In this example, the accession IDs of Gene 2 (secondary input) from GN1–GN3 (input) are displayed as results, with the origin of each Gene 2 specified in parentheses indicating the corresponding genome neighborhood ID.

Note: After filtering the Pfam ID of interest in Figure S3, clicking "REPLACE INPUT WITH OUTPUT" button and following the instructions above enables efficient retrieval of the accession IDs of the targeted neighboring genes.

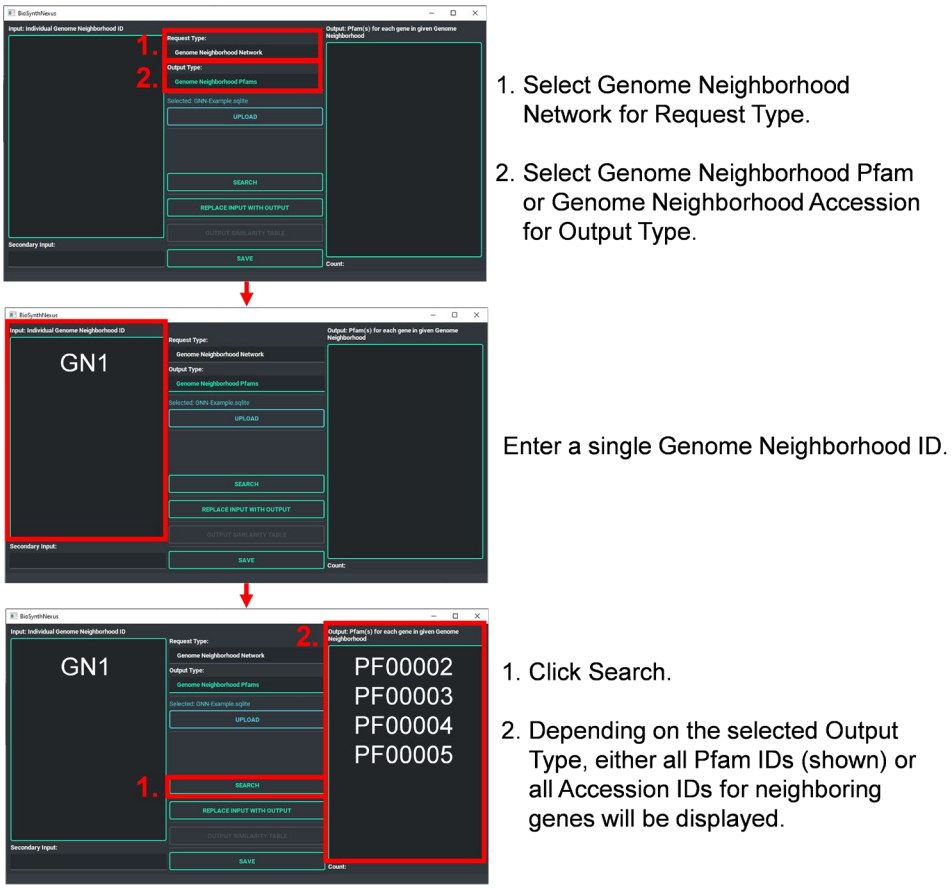
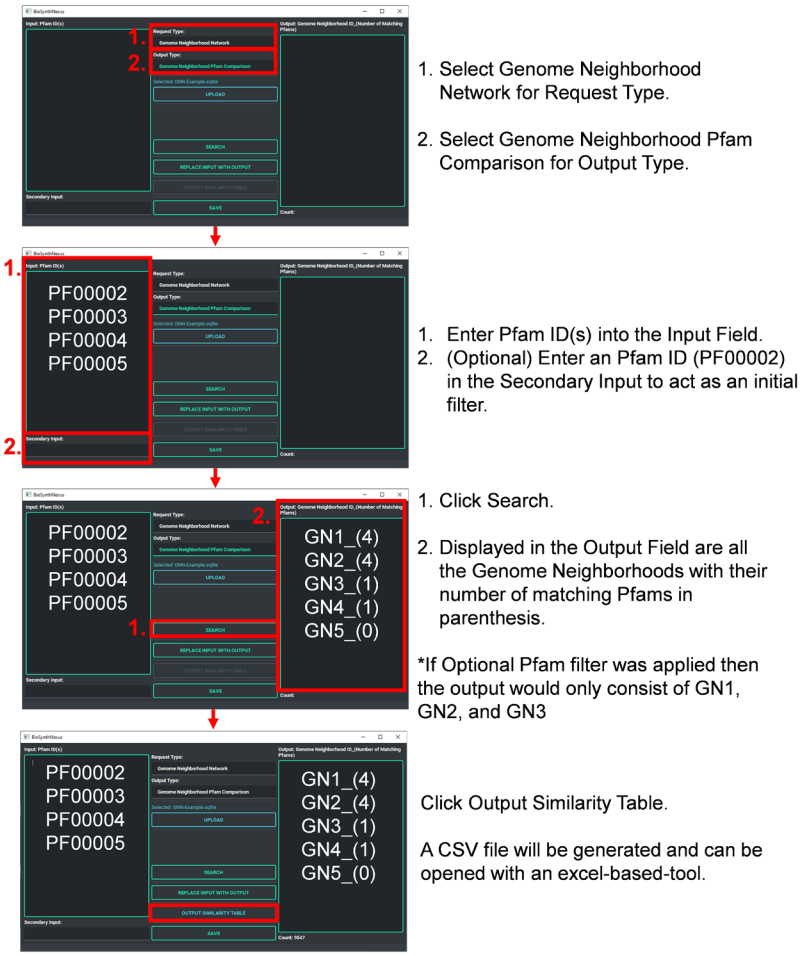


Figure S5. Retrieval of all Pfam IDs for neighboring genes from the given genome neighborhoods. In this example, PF00002–PF00005 are displayed as results because these neighboring genes are within GN1 (input).

Note: The accession IDs of GN1 neighboring genes will be displayed as results if "Genome Neighborhood Accession" is selected as the output type. The number of retrieved neighboring genes depends on the neighborhood size specified when generating the GNN sqlite file from the EFI-GNT website.



Genome Neighborhood ID	PF00002	PF00003	PF00004	PF00005	TRUE COUNT
GN1	TRUE	TRUE	TRUE	TRUE	4
GN2	TRUE	TRUE	TRUE	TRUE	4
GN3	TRUE	FALSE	FALSE	FALSE	1
GN4	FALSE	TRUE	FALSE	FALSE	1
GN5	FALSE	FALSE	FALSE	FALSE	0

Figure S6. Retrieval of genome neighborhood information from the given Pfam ID(s). In this example, all genome neighborhoods are displayed as results, with the number of matching Pfam IDs in parentheses. Moreover, a CSV file can be retrieved to show which Pfam IDs match in each genome neighborhood.

Note: The secondary input is an optional choice to pre-filter the Pfam IDs. If “PF00002” is applied as the secondary input, only GN1–GN3 will be displayed as results, as they are the only genome neighborhoods containing a neighboring gene from PF00002 (Figure S1).

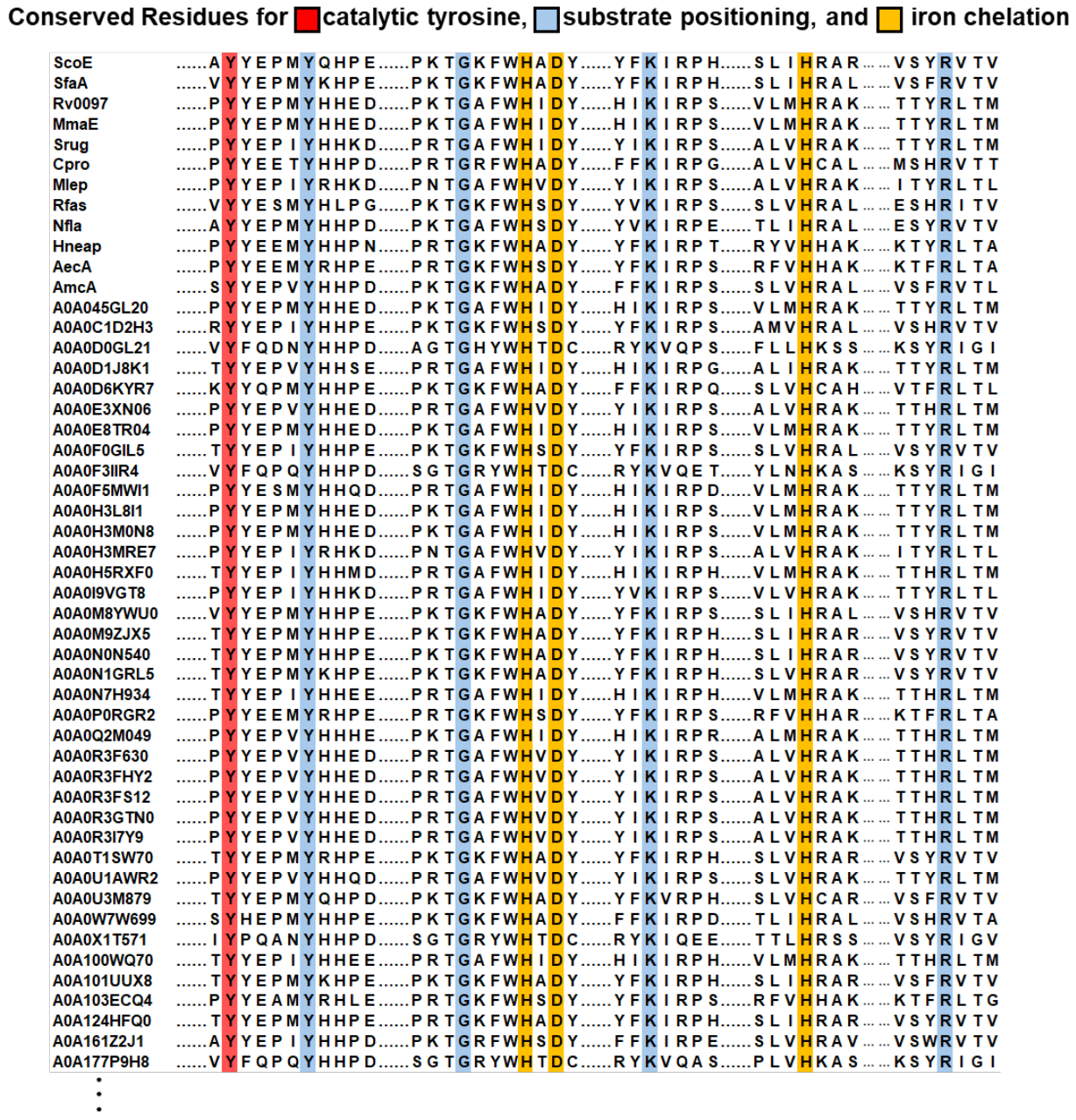


Figure S7. Multiple sequence alignment of 365 potential isonitrilases confirms that all key residues critical for isonitrilase activity are conserved. Due to space limitations, only a few genes are shown as examples.

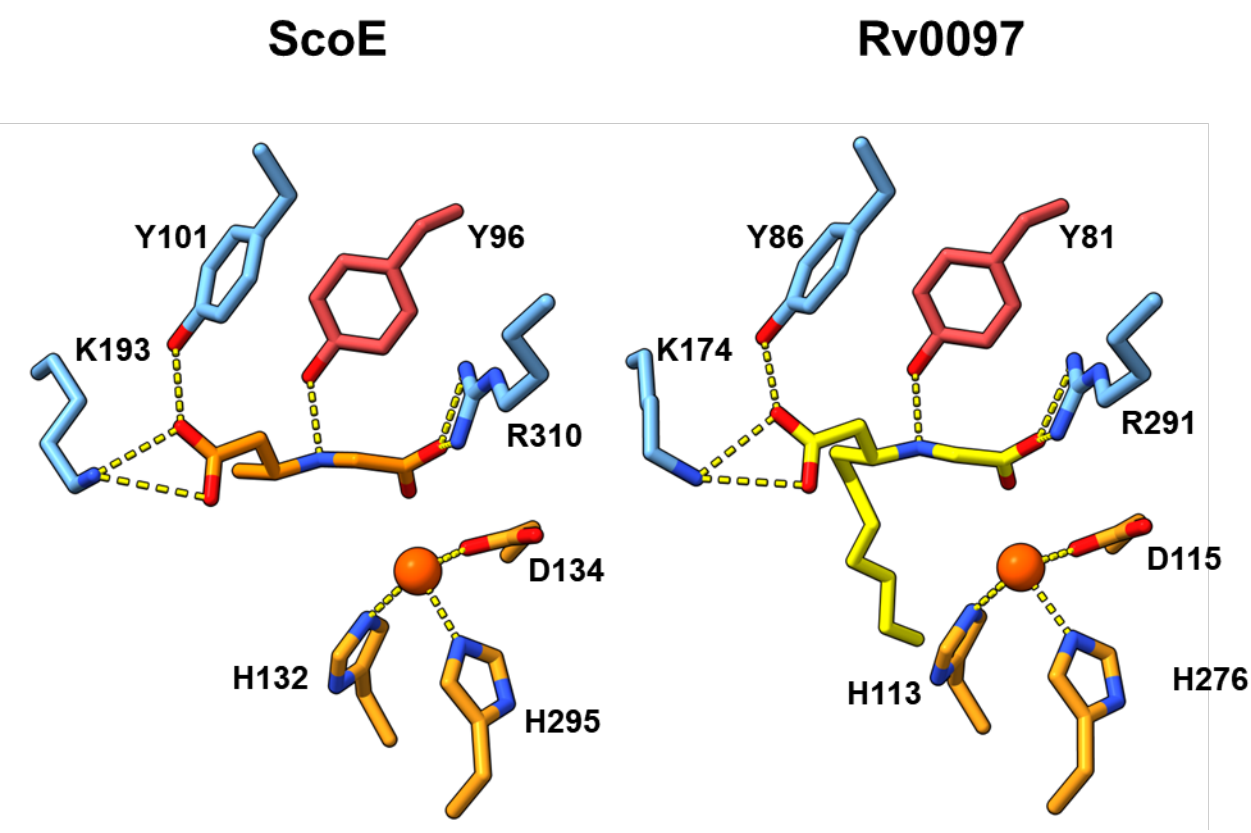


Figure S8. Conserved residues found in characterized isonitrilases. The 2-His/1-Asp triad coordinates iron, a catalytic tyrosine (Y96 in ScoE and Y81 in Rv0097) involves the formation of an aldimine intermediate, and other residues are responsible for substrate positioning.

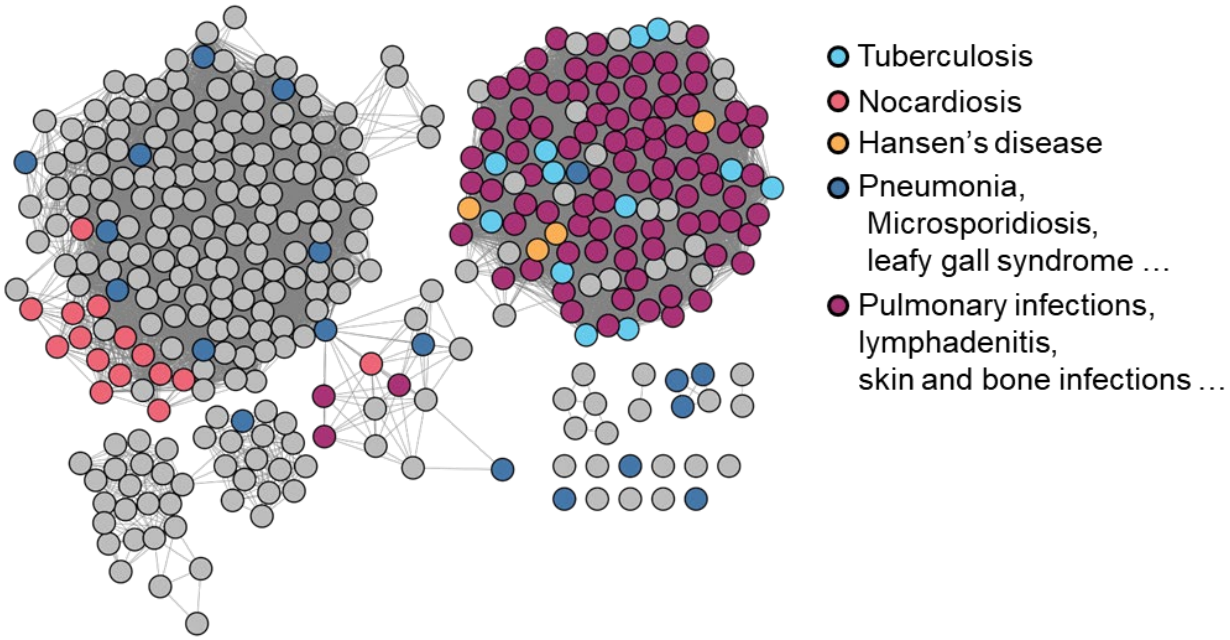


Figure S9. SSN of 365 putative Fe/2OG isonitrilases. Pathogenic associated species are colored respective to their symptoms/ associated diseases.

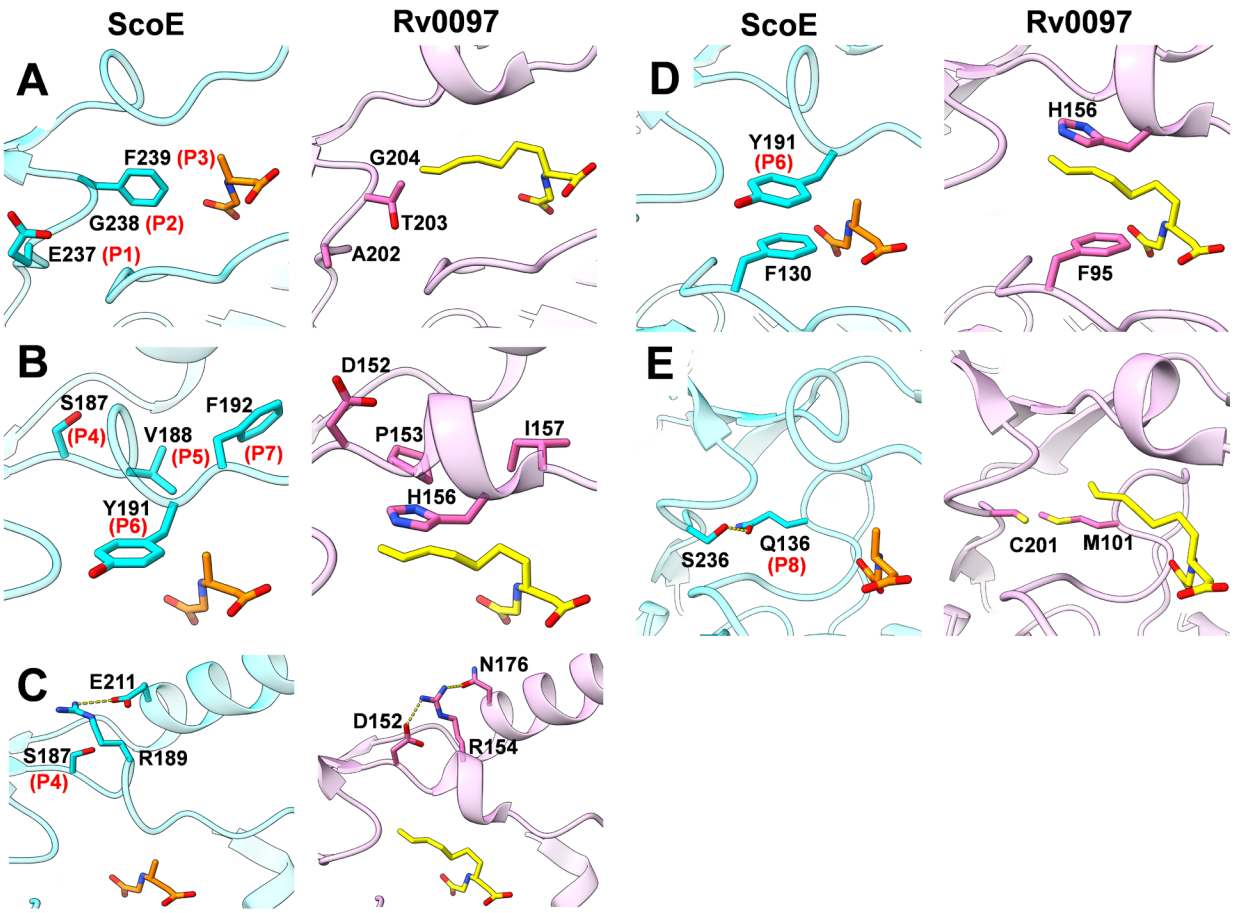


Figure S10. Eight selected positions in reported structures of ScoE bound with methyl-**6** (left, PDB ID: 6L6X) and Rv0097 bound with heptyl-**6** (right, PDB ID: 8KHT). (A) Significant steric differences are found between the two enzymes at P1–3, where a phenylalanine (P3) in ScoE block the entrance to the pocket. (B) An α-helix observed in Rv0097 shifts the entire loop away from the active site. As a result, (C) no π-π interaction is observed between P6 and a conserved phenylalanine, and (D) a hydrogen interaction is identified between R154 and P4 in Rv0097. (E) Different interactions with surrounding residues and P8 are detected in ScoE and Rv0097.

	P1	P2	P3	P4	P5	P6	P7	P8
ScoE	E	G	F	S	V	Y	F	Q
SfaA	E	G	F	S	V	Y	F	Q
Rv0097	A	T	G	D	P	H	I	M
MmaE	A	T	G	D	P	H	I	M
Srug	E	A	G	S	P	Y	I	Q
Cpro	E	G	F	S	V	F	F	M
Mlep	E	A	A	S	P	Y	I	I
Rfas	E	A	F	S	S	Y	V	Q
Nfla	E	A	F	S	A	Y	V	Q
Hneap	R	G	F	S	V	Y	F	A
AecA	R	G	F	S	V	Y	F	A
AmcA	E	G	F	S	P	F	F	Q
A0A045GL20	A	T	G	D	P	H	I	M
A0A0C1D2H3	E	G	F	S	V	Y	F	G
A0A0D0GL21	S	G	F	G	G	R	Y	Q
A0A0D1J8K1	S	T	G	D	P	H	I	M
A0A0D6KYR7	E	G	F	S	V	F	F	A
A0A0E3XN06	E	A	G	T	P	Y	I	M
A0A0E8TR04	A	T	G	D	P	H	I	M
A0A0F0GIL5	E	G	F	S	V	Y	F	Q
A0A0F3IIR4	S	G	F	E	G	R	Y	Q
A0A0F5MWI1	A	S	G	D	P	H	I	M
A0A0H3L8I1	A	T	G	D	P	H	I	M
A0A0H3M0N8	A	T	G	D	P	H	I	M
A0A0H3MRE7	E	A	A	S	P	Y	I	I
A0A0H5RXF0	A	S	G	D	P	H	I	M
A0A0I9VGT8	E	A	A	S	P	Y	V	M
A0A0M8YWU0	E	G	F	S	P	Y	F	Q
A0A0M9ZJX5	E	G	F	S	V	Y	F	Q
A0A0N0N540	E	G	F	S	V	Y	F	Q
A0A0N1GRL5	E	G	F	S	V	Y	F	Q
A0A0N7H934	A	S	G	D	P	H	I	M
A0A0P0RGR2	R	G	F	S	V	Y	F	A
A0A0Q2M049	A	T	G	D	P	H	I	M
A0A0R3F630	E	A	G	T	P	Y	I	M
A0A0R3FHY2	E	A	G	T	P	Y	I	M
A0A0R3FS12	E	A	G	T	P	Y	I	M
A0A0R3GTN0	E	A	G	T	P	Y	I	M
A0A0R3I7Y9	E	A	G	T	P	Y	I	M
A0A0T1SW70	E	G	F	S	V	Y	F	Q
A0A0U1AWR2	E	A	G	T	P	Y	I	M
A0A0U3M879	E	G	F	S	V	Y	F	Q
A0A0W7W699	E	G	F	S	V	F	F	Q
A0A0X1T571	G	G	F	E	G	R	Y	Q
A0A100WQ70	A	S	G	D	P	H	I	M
A0A101UUX8	E	G	F	S	V	Y	F	Q
A0A103ECQ4	R	G	F	S	V	Y	F	A
A0A124HFQ0	E	G	F	S	V	Y	F	Q
A0A161Z2J1	E	G	F	S	V	F	F	Q
A0A177P9H8	S	G	F	E	G	R	Y	A

⋮

Figure S11. Multiple Sequence Alignment of eight selected positions for the 365 potential isonitrilases. Positions are displayed in a rearranged order that does not match their sequential arrangement in the genes. Due to space limitations, only a subset of representative genes are shown.

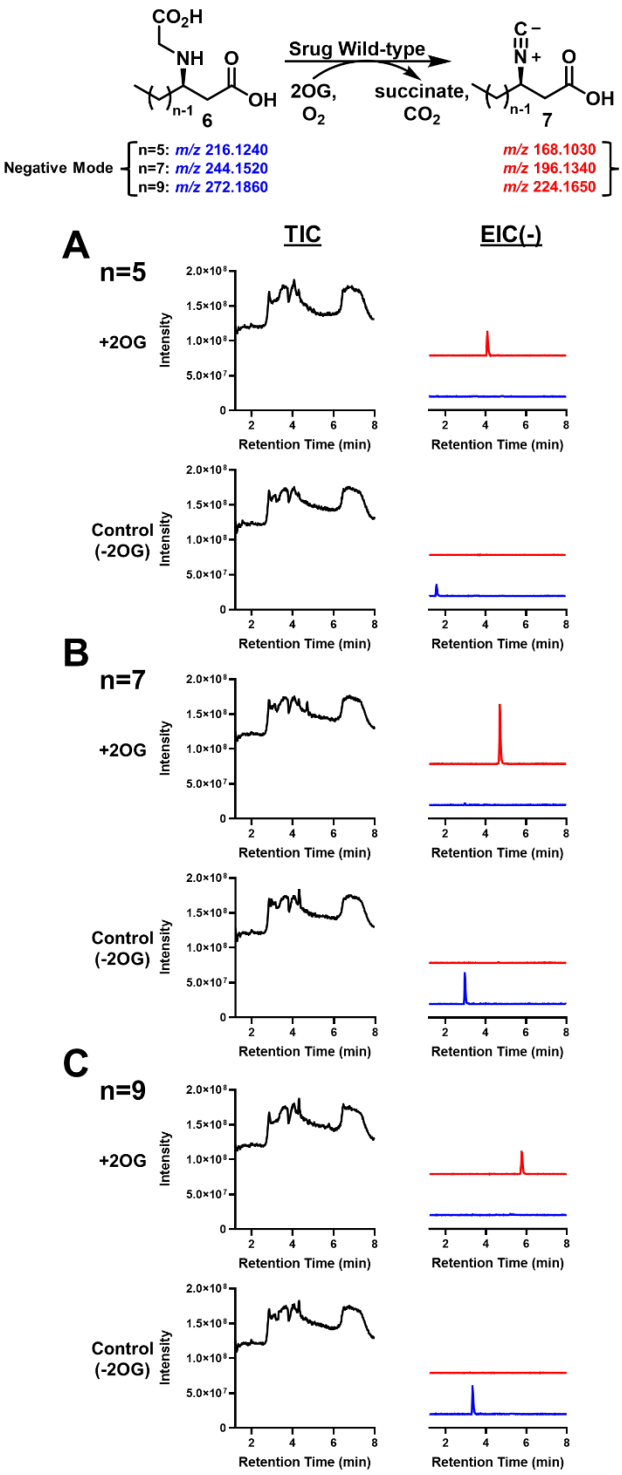


Figure S12. LC-MS analysis of wild-type Srug reactions without derivatization. Both the total ion chromatogram (TIC, left panel) and the extracted ion chromatogram (EIC, right panel) were used for analysis. The corresponding isonitrile product was detected upon incorporation of 2OG, while no other products were observed. The EIC was conducted in negative mode, with the corresponding *m*/*z* values indicated in the scheme.

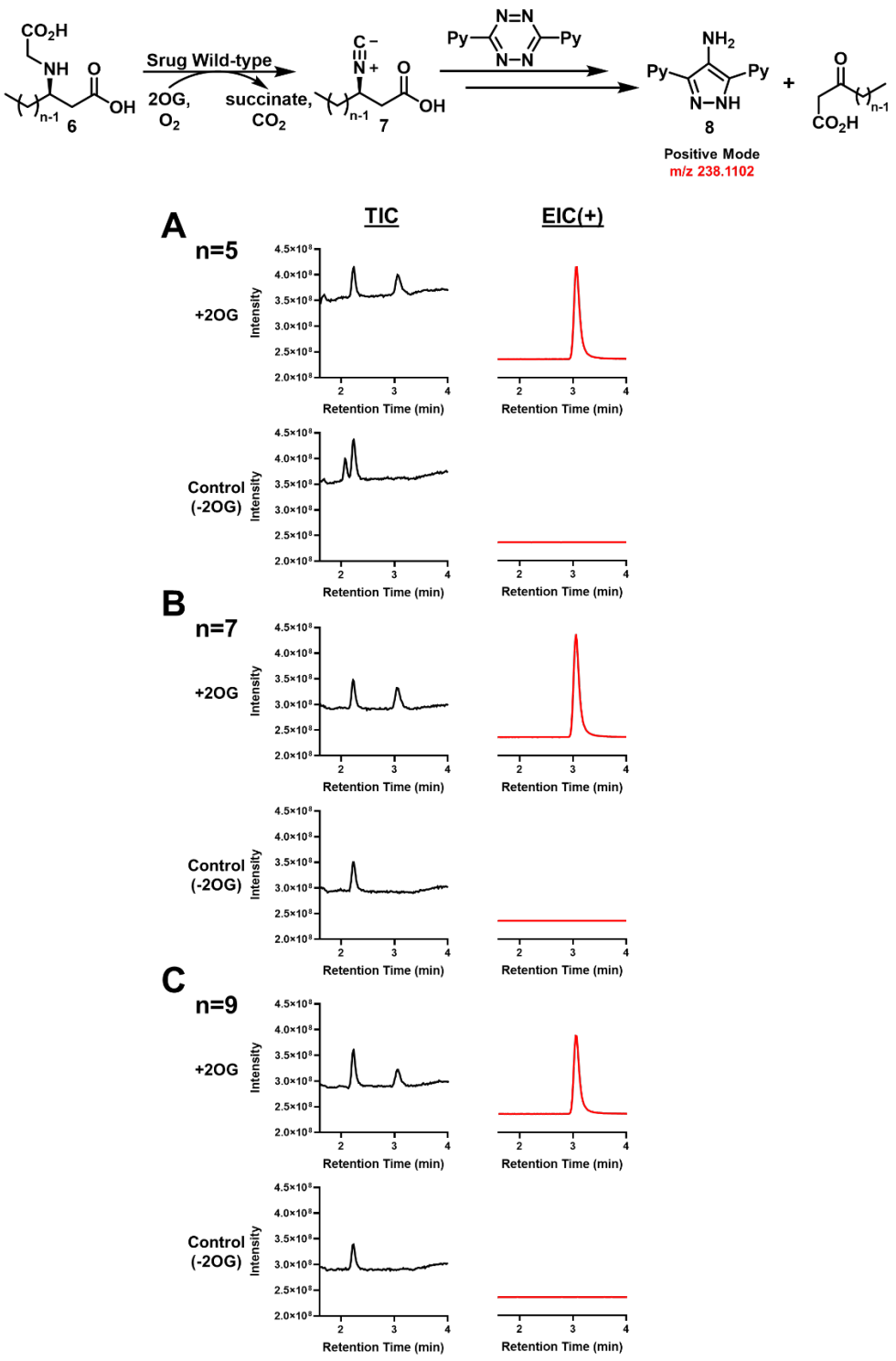


Figure S13. LC-MS analysis of wild-type Srug reactions with tetrazine derivatization. Both the total ion chromatogram (TIC, left panel) and the extracted ion chromatogram (EIC, right panel) were used for analysis. With tetrazine derivatization, the same pyrazole product **8** (*m*/*z* 238.1102) is generated, regardless of the substrate used. The EIC was conducted in positive mode.

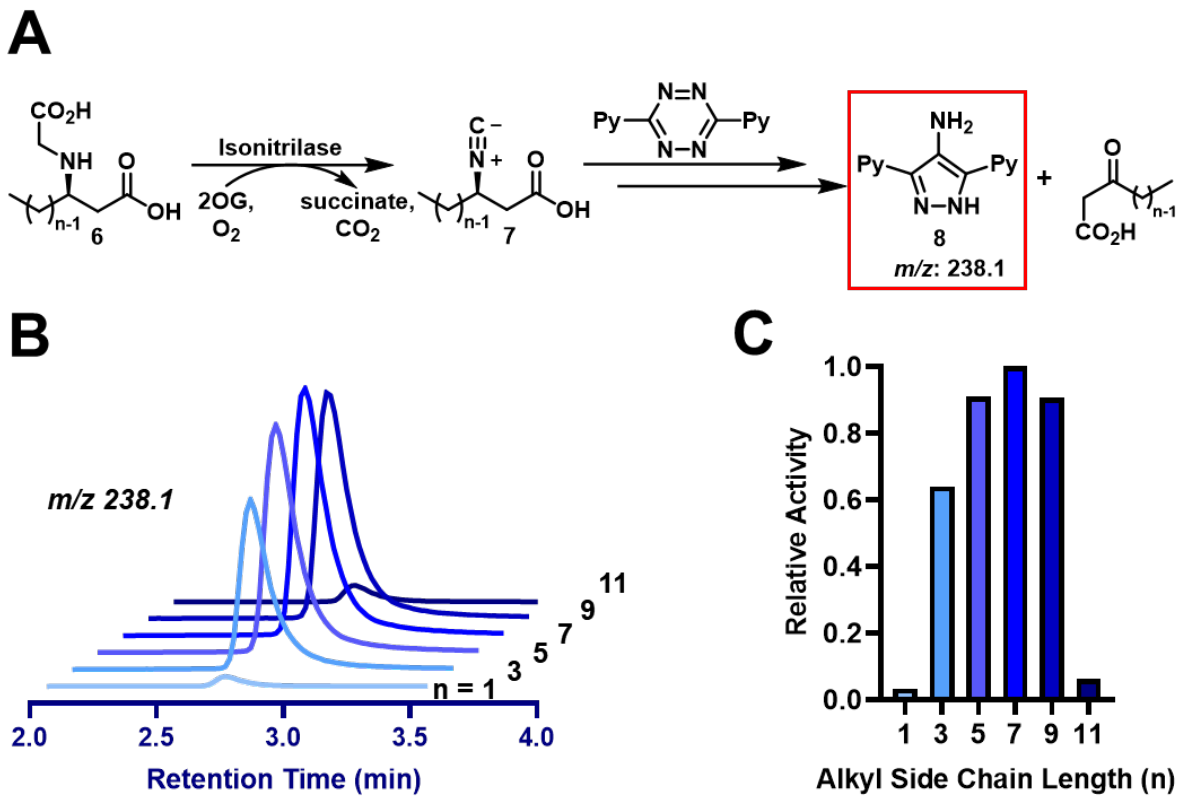


Figure S14. Analysis of isonitrilase substrate selectivity. (A) After **7** formation, tetrazine is used to derivative the isonitrile group, yielding the common pyrazole product **8** with *m*/*z* of 238.1. (B) LC-MS analysis of isonitrilase substrate selectivity regrading various alkyl chain length (**6**, n=1–11). Srug wild-type is shown as an example. (C) Integration of each chromatogram is further plotted into a bar graph.

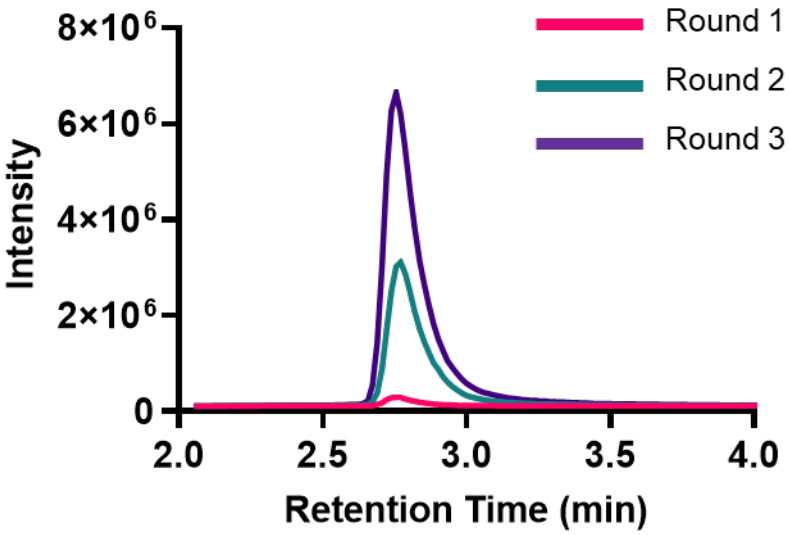


Figure S15. LC-MS analysis of Srug variants with undecyl substrate (n=11). The relative activity toward the undecyl alkyl chain substrate increased 36-fold after multiple-round mutagenesis. Round 1: S152D/Y156H; Round 2: S152D/Y156H/E202A/A203T; Round 3: S152D/Y156H/E202A/A203T/Q102M

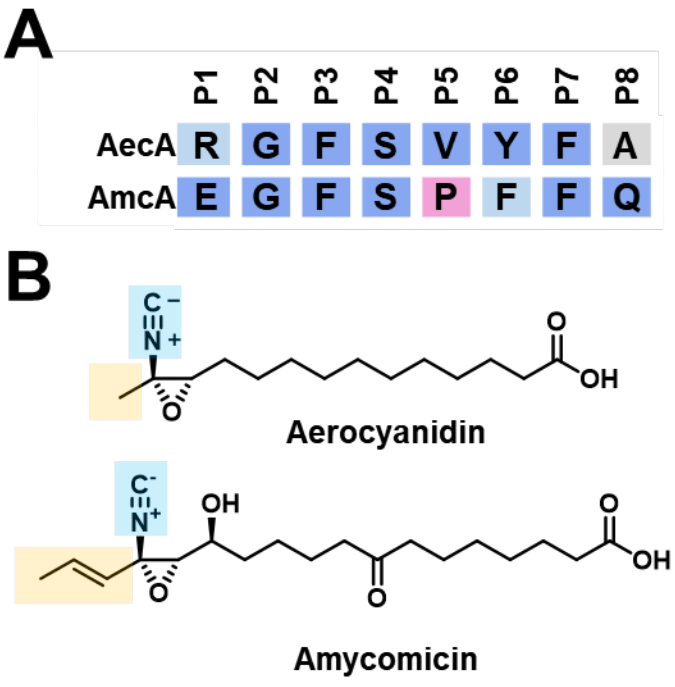


Figure S16. (A) Eight selected positions of AecA and AmcA. (B) Structures of aerocyanidin and amycomicin. The alkyl chain lengths of isonitrile moieties in these two natural products align with our predictions based on eight selected positions.

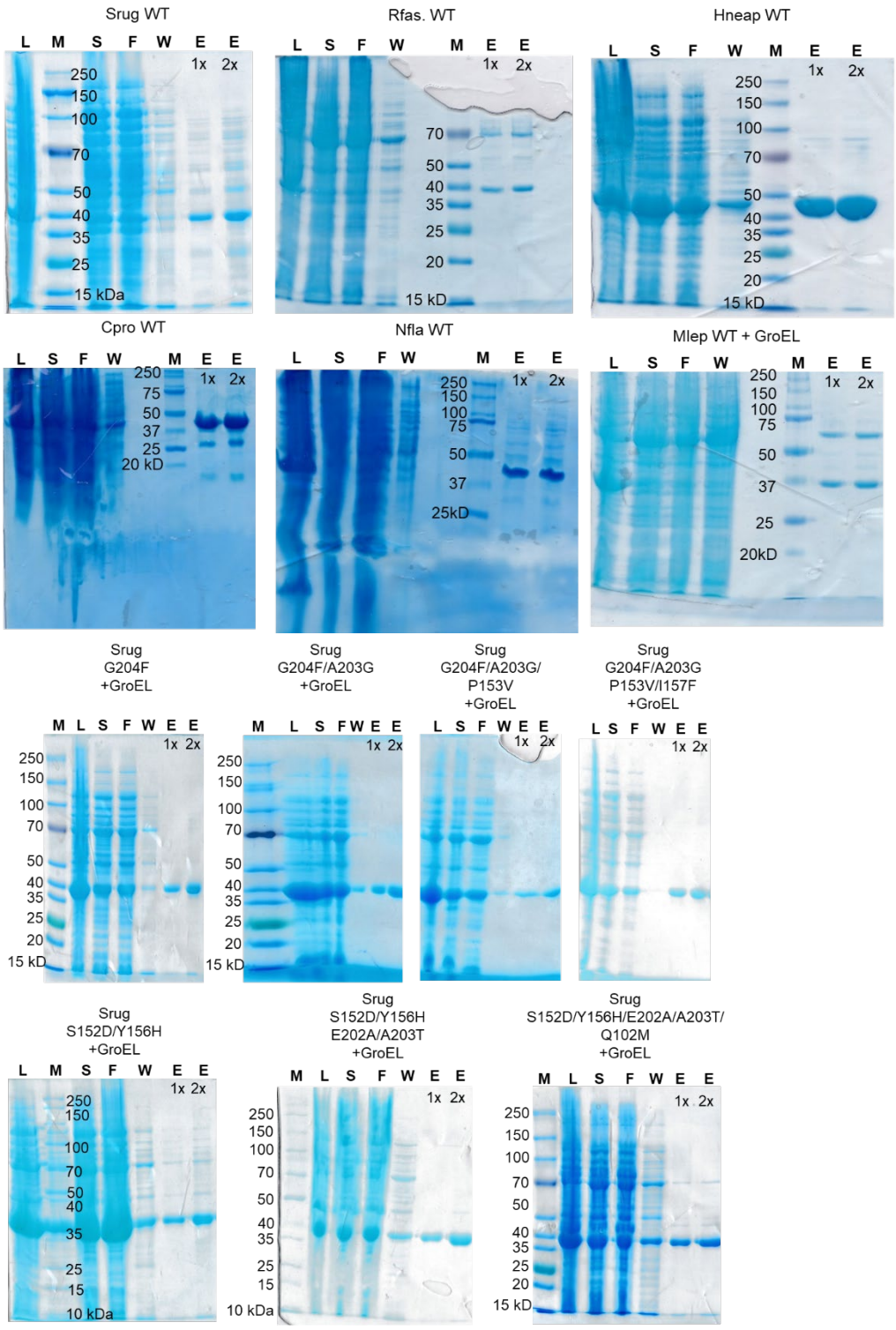


Figure S17. Coomassie-stained SDS-PAGE (L: lysate, S: supernatant after centrifuge, F: flow-through, W: wash, E: elution and M: marker).

Table S1. Eight selected residues with respective protein position numbers

Position	P1	P2	P3	P4	P5	P6	P7	P8
ScoE	E237	G238	F239	S187	V188	Y191	F192	Q136
Rv0097	A202	T203	G204	D152	P153	H156	I157	M101
SfaA	E207	G208	F209	S157	V158	Y161	F162	Q106
MmaE	A202	T203	G204	D152	P153	H156	I157	M101
Srug	E202	A203	G204	S152	P153	Y156	I157	Q101
Mlep	E202	A203	A204	S152	P153	Y156	I157	I101
Nfla	E206	A207	F208	S156	A157	Y160	V161	Q105
Rfas	E207	A208	F209	S157	S158	Y161	V162	Q106
Hneap	R204	G205	F206	S154	V155	Y158	F159	A103
Cpro	E210	G211	F212	S160	V161	F164	F165	M109
AecA	R204	G205	F206	S154	V155	Y158	F159	A103
AmcA	E207	G208	F209	S157	P158	F161	F162	Q106

Table S2. Primers for Srug variants.

Variant	Primer	Sequence (5' – 3')	Tm (°C)	Template
S152D/Y156H	F	CCGTCACATTAAAATTAGACCCTCGG	68	WT
	R	CGCGGATCATGGATAGAACGC		
S152D/Y156H/ E202A/A203T	F	GATGACCATCGAGGACGG	68	S152/Y156H
	R	GTACCCGTTGCGCAAATATATAAGATTTC		
S152D/Y156H/ E202A/A203T/Q101M	F	GACTATATGTTCATGCCGAAACCGTTTG	68	S152D/Y156H/ E202A/A203T
	R	GATATGCCAGAACGCACCG		
G204F	F	CGTTTACGATGACCATCGAGG	65	WT
	R	CTTCGCAAATATATAAGATTTCTTGGCC		
G204F/A203G	F	GGTTTACGATGACCATCGAGG	65	G204F
	R	CTTCGCAAATATATAAGATTTCTTGGCC		
G204F/A203G/ P153V	F	CGTTACATTAAAATTAGACCCTCGG	64	G204F/A203G
	R	GCGCACAGAATGGATAGAAC		
G204F/A203G/ P153V/I157F	F	GCGCCGTTACTTTAAAATTAGACC	64	G204F/A203G/ P153V
	R	GGAGAATGGATAGAACGCG		

References

1. N. Oberg, R. Zallot and J. A. Gerlt, *J. Mol. Biol.*, 2023, **435**, 168018.
2. R. Zallot, N. Oberg and J. A. Gerlt, *Biochemistry*, 2019, **58**, 4169–4182.
3. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
4. C. Notredame, D. G. Higgins and J. Heringa, *J. Mol. Biol.*, 2000, **302**, 205–217.
5. M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, *J. Cheminformatics*, 2012, **4**, 17.
6. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
7. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785–2791.
8. J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, *J. Chem. Inf. Model.*, 2021, **61**, 3891–3898.
9. O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
10. E. C. Meng, T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J. Pearson, J. H. Morris and T. E. Ferrin, *Protein Sci.*, 2023, **32**, e4792.