Spectroscopic, Kinetic, and Bioinformatic Characterization of Bacterial Cysteine Dioxygenases

By

Rebecca L. Schultz

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 5 June 2023

The dissertation is approved by the following members of the Final Oral Committee:

Thomas C. Brunold, Professor, Inorganic Chemistry
Andrew R. Buller, Professor, Chemical Biology
Judith N. Burstyn, Professor, Chemical Biology
Brian G. Fox, Professor, Biochemistry

Spectroscopic, Kinetic, and Bioinformatic Investigation of Bacterial Cysteine Dioxygenases

Rebecca L. Schultz

Under the supervision of Professor Thomas C. Brunold at the University of Wisconsin-Madison

Abstract

Cysteine dioxygenase (CDO) is a non-heme iron-containing enzyme that catalyzes the oxidation of cysteine (Cys) to cysteine sulfinic acid (CSA). Crystal structures of eukaryotic CDOs revealed the presence of an unusual crosslink between the sulfur of a cysteine residue (C93 in *Mus musculus* CDO, *Mm*CDO) and a carbon atom adjacent to the phenyl group of a tyrosine residue (Y157). Formation of this crosslink occurs over time as a byproduct of catalysis and increases the catalytic efficiency of CDO by at least 10-fold. Interestingly, in bacterial CDOs, the residue corresponding to C93 is replaced by a highly conserved glycine (G82 in *Bacillus subtilis* CDO, *Bs*CDO), which precludes the formation of a C–Y crosslink in these enzymes; yet bacterial CDOs achieve turnover rates paralleling those of fully crosslinked eukaryotic CDOs. We prepared the G82C variant of *Bs*CDO to determine if a single DNA point mutation could lead to C–Y crosslink formation in this enzyme. We used gel electrophoresis, peptide mass spectrometry, electron paramagnetic resonance spectroscopy, and kinetic assays to characterize this variant alongside the natively crosslinked wild type (WT) *Mm*CDO and the natively non-crosslinked WT *Bs*CDO. Collectively, our results provide compelling evidence that the G82C *Bs*CDO variant is indeed capable of C–Y crosslink formation. Our kinetic studies indicate that G82C *Bs*CDO has a reduced catalytic efficiency compared to WT *Bs*CDO and that activity increases as the ratio of crosslinked to non-crosslinked enzyme increases. Finally, by carrying out a bioinformatic analysis of the CDO family, we were able to identify a large number of putatively crosslinked bacterial CDOs, the majority of which are from gram-negative pathogenic bacteria.

To better understand the extreme substrate specificity of CDO, we characterized various substrate (analogue)-bound complexes of WT *Mm*CDO, WT *Bs*CDO, and G82C *Bs*CDO. MCD, Abs, and EPR spectroscopies were used to evaluate the relative binding affinities of substrate (analogues), active site structures and conformational flexibility, and the role of the substrate Cys amine, thiol, and carboxylate moieties with regard to both binding and turnover. UPLC-MS and ^{1}H-NMR activity assays were performed to evaluate the extent of substrate (analogue) turnover at different time points and concentrations. We found that the *Bs*CDO active site is much more flexible and thus more accommodating of substrate analogues than the *Mm*CDO active site. The increased conformational freedom experienced by substrate (analogues) bound to the *Bs*CDO active site are not only due to the lack of a C-Y crosslink that is important for the function of *Mm*CDO, but may also be due to the absence of a small two-strand beta sheet above the flexible active site wall nearest the substrate amino group.

Acknowledgements

Like most things in life, a Ph.D. can't be completed alone. The work presented in this dissertation would not have been possible without the support of my department, advisor, coworkers, friends, and family.

To Thomas, your mentorship has allowed me to grow into a confident scientist and researcher. You gave me the support to master techniques beyond my comfort zone as well as the space to work independently and creatively. I hope I've been able to help shape this group as much as it has helped shape me.

To my group members, past and present. Rebeca, Elizabeth, Joshua, Mickie, Laura, Ryan, Zach, and Cristina—thank you for your mentorship and your friendship. You let me be my realest, weirdest self around you.

To Brian, Judith, and Andrew and all the wonderful students in your labs, thank you for sharing your resources, time, and knowledge. My work would not have been nearly as productive without your support.

To the incredible scientific, teaching, and administrative staff in and associated with the Chemistry Department, thank you for your guidance and enthusiasm. You all made accessing resources within the department incredibly easy and made the chemistry building a welcoming, supportive place to work.

To my friends, thank you for helping me grow not just as a scientist but as a person. Lydia, Oliver, and Vinnie, you showed me how to be spontaneous and weird and not afraid of mud. My soccer teams—Tots and Bananas and PPI—and all the friends I've met playing over the years, you kept me sane and healthy and reminded me to have fun. My hockey friends, thank you for welcoming and encouraging me as I grow. You really can learn anything you put your mind to.

To Meeko, thank you for putting up with my long nights in lab and boring work-from-home writing days. We can play fetch at the park as often as you want this summer.

To Mom, Dad, and Sam, thank you for your support through graduate school and through life. I love you and I wouldn't be anywhere without you.

To Ryan, I'm incredibly lucky to have experienced graduate school with you, and I can't wait to experience the rest of life with you as well.

Dedication

For Mom and Dad.

Table of Contents

List of Figures

List of Schemes

List of Tables

Abbreviations

CDO	cysteine dioxygenase
Cys	cysteine
CSA	cysteine sulfinic acid
ADO	cysteamine dioxygenase
MDO	3-mercaptopropionate dioxygenase
MSDO	mercaptosuccinate dioxygenase
CT	charge transfer
Mm	*Mus musculus*
Sec	selenocysteine
Bs	*Bacillus subtilis*
PCO	plant cysteine dioxygenase
3-His	three-histidine
MCD	magnetic circular dichroism
QM/MM	quantum mechanics/molecular mechanics
Pa	*Pseudomonas aeruginosa*
EPR	electron paramagnetic resonance
FAS	ferrous ammonium sulfate
IPTG	isopropyl-β-D-thiogalactopyranoside
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
MBP	maltose binding protein
TFA	trifluoroacetic acid
CID	collision-induced dissociation
FDR	false discovery rate
TCEP	tris(2-carboxyethyl)phosphine

UPLC-MS	ultra-high performance liquid chromatography mass spectrometer
MES	2-(N-morpholino)ethanesulfonic acid
BCS	bathocuproine disulfonate
SSN	sequence similarity network
EFI-EST	Enzyme Function Initiative-Enzyme Similarity Tool
MSA	multiple sequence alignment
MUSCLE	Multiple Sequence Comparison by Log-Expectation
FCB	Fibrobacteres-Chlorobi-Bacteroidetes
TSP	trimethylsilylpropanoic acid
2-AET	cysteamine
3-MPA	3-mercaptopropionic acid
HCys	homocysteine
HT	hypotaurine
3-SPA	3-sulfinopropionic acid
HCSA	homocysteine sulfinic acid
LT	low temperature
RT	room temperature
^{1}H-NMR	^{1}H nuclear magnetic resonance

CHAPTER 1

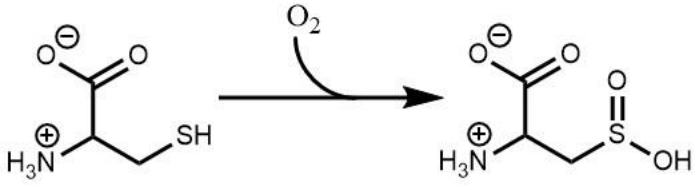
Introduction:

The Cysteine Dioxygenase Family and How Homologues Can Help Us Understand the Functional Role of Unique Enzyme Structural Features

Chapter 1 : Introduction: The Cysteine Dioxygenase Family and How Homologues Can Help Us Understand the Functional Role of Unique Enzyme Structural Features

Biological Role of Cysteine Dioxygenase

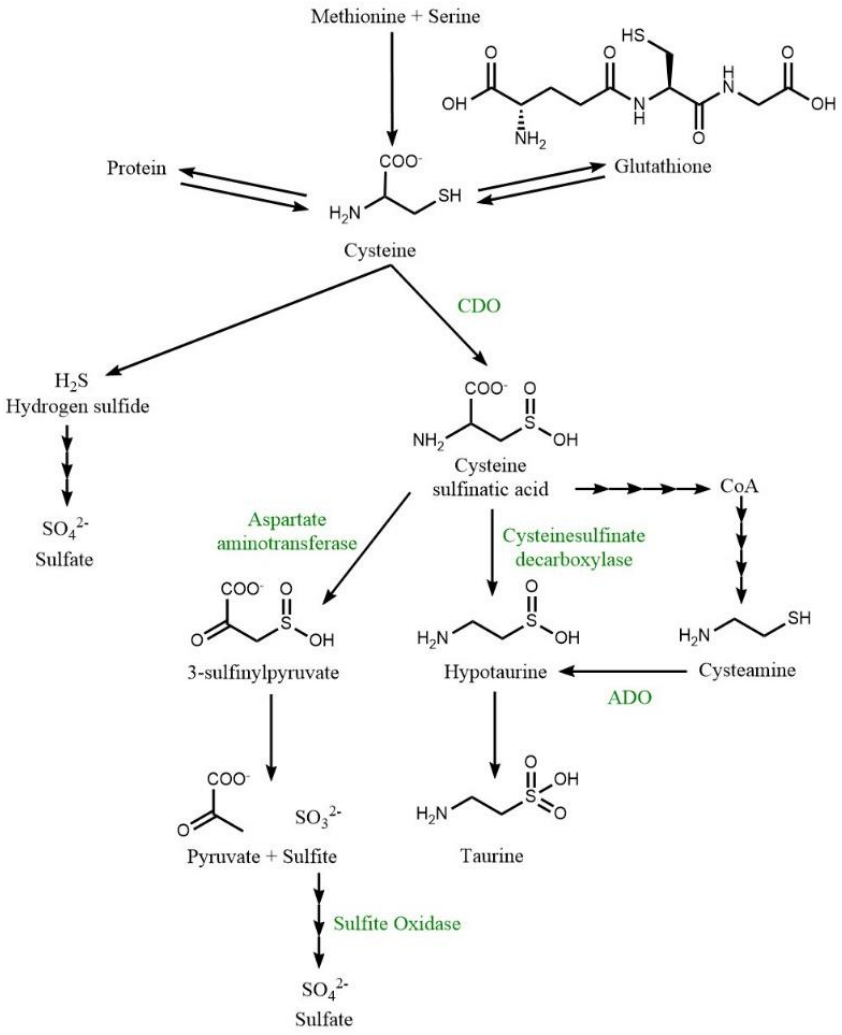
Cysteine dioxygenase (CDO) is a non-heme Fe(II)-dependent enzyme that catalyzes the irreversible oxidation of cysteine (Cys) to cysteine sulfinic acid (CSA) via addition of molecular oxygen (**Scheme 1**).[1] It is a member of the larger group of thiol dioxygenases, ferrous iron-containing enzymes that oxidize thiols to their corresponding sulfinates. This class includes CDO, cysteamine dioxygenase (ADO), 3-mercaptopropionate dioxygenase (MDO), and mercaptosuccinate dioxygenase (MSDO).



Scheme 1.1 CDO-catalyzed conversion of L-cysteine to L-cysteine sulfinic acid.

Cysteine is an essential building block of proteins as well as a precursor for the biosynthesis of the metabolites coenzyme A, glutathione, taurine and pyruvate[2]; however, excess cysteine is both cytotoxic[3] and neurotoxic[4], and is easily oxidized to the poorly soluble cystine, which can accumulate as stones in the human renal system[5]. Abnormal or deficient CDO activity, primarily observed as elevated cysteine-to-sulfate ratios, is reported in patients suffering from several autoimmune and neurodegenerative diseases, including rheumatoid arthritis, Parkinson's disease, Alzheimer's disease, and motor neuron diseases[6–9]. The ability of organisms to regulate intracellular cysteine levels within a narrow range, typically between 0.03 and 0.2 μmol per gram in most tissues[10], is therefore of vital importance.

There are several pathways of cysteine catabolism: a desulfuration pathway that gives hydrogen sulfide, a cysteamine-dependent pathway that uses ADO to oxidize cysteamine to hypotaurine, and the cysteine sulfinate-dependent pathway that starts with the CDO catalyzed oxidation of cysteine to cysteinesulfinate and branches off to give either pyruvate and sulfate or hypotaurine and taurine (**Figure 1**). The relative flux through these different pathways varies depending on the amount of sulfur-containing amino acids in the diet, and much of this regulation is controlled at the level of CDO[2,11].



Scheme 1.2 Metabolic pathway of sulfur metabolism in eukaryotic organisms. Enzymes are labeled in green.

CDO activity changes quickly in response to diet, reaching a new steady-state within hours in animals switched from a low-protein to high-protein diet[12]. In hepatic and adipose cells, cysteine levels have been shown to regulate CDO concentration not at the level of mRNA, but instead at the level of CDO degradation. At low cysteine levels, CDO is readily ubiquitinated and degraded by the 26S proteasome system; at high cysteine levels, ubiquitination of CDO is inhibited and CDO concentration is increased[13]. Activity is regulated up to 45-fold by controlling the enzyme concentration[12].

Enzyme Structure

Crystal structures show that CDO is a member of the functionally diverse cupin superfamily, whose members generally conserves a β-barrel fold and two sequence motifs: $G(X)_5HXH(X)_{3,4}E(X)_6G$ and $G(X)_5PXG(X)_2H(X)_3N$[14]. The active site of CDO contains several unique conserved residues that help to coordinate the iron molecule and facilitate the binding and reactivity of cysteine and molecular oxygen. The nitrogen atoms of His-86, -88, and -140 (*Mus musculus* CDO, *Mm*CDO numbering) coordinate the iron instead of the 2-His, 1-carboxylate facial triad typically found in many metalloproteins[14]. Other highly conserved active site residues of *Mm*CDO include His-155, Ser-153, and Trp-77, which form a hydrogen bonding network with Tyr157. Residues Ser-153, His-155, and Tyr-157 are reminiscent of a catalytic triad motif seen in many enzymes. Both Tyr-58 and Arg-60 hydrogen bond with the carboxylate group of the substrate cysteine and are thought to be important for correctly orienting cysteine in the active site[14].

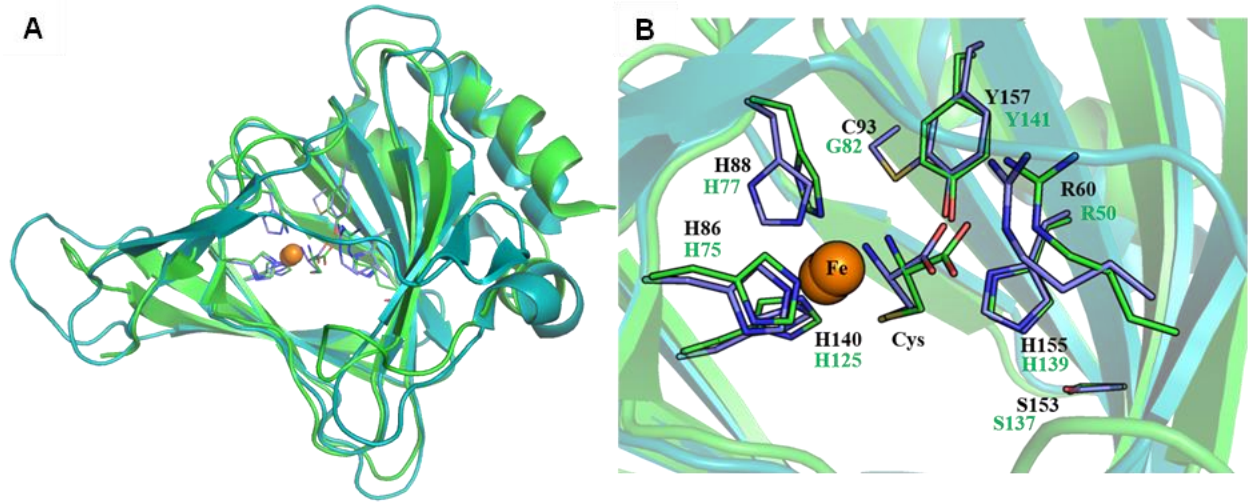


Figure 1.1 Alignment of Cys-bound *Rattus norvegicus* CDO (PDB= code 4JTO, blue) and *Bs*CDO (PDB= code 4QM9, green). *(A)* Overall structure of the CDO enzymes. (*B)* Active site of CDO enzymes with key residues highlighted. *Mm*CDO numbering in black text. *Bs*CDO numbering in green text.

The resting state has a 6-coordinate pseudo-octahedral geometry, with the iron coordinated by the three histidines and three water molecules[14]. Structures of CDO crystallized in the presence of cysteine show that cysteine displaces all three active site water molecules to coordinate the iron in a bidentate fashion via its sulfur and amine groups. The sixth coordination site is then available for molecular oxygen to bind end-on to the iron, facilitated by a hydrophobic pocket surrounding the site[15].

The active site of eukaryotic CDO also contains a unique thioether crosslink between the sulfur of C93 and an ortho carbon of T157[14] (*Mm*CDO numbering). A similar crosslink has been observed by X-ray crystallography in only a small number of other enzymes, including galactose oxidase and the bacterial sulfite reductase NirA. The CDO crosslink forms after several hundred catalytic turnovers with the substrate cysteine and increases the catalytic efficiency of the enzyme by 10-fold[16]. The crosslinked and non-crosslinked forms of the protein travel as two bands of slightly different apparent molecular weight on an SDS-PAGE gel[16]. The mechanism of formation of the crosslink is unknown, but its putative role is to facilitate the proper positioning of

Tyr-157 relative to the two substrates and the iron center of the enzyme, and to prevent the binding of a water molecule to cysteine-bound Fe(II) CDO[17].

Spectroscopic Methods for Studying CDO

Because it is difficult to observe short-lived reaction intermediates, our lab has turned to cysteine analogue- and oxygen mimic-bound forms of CDO as close proxies to identify important mechanistic constraints and on-pathway intermediates of the enzyme. In particular, we have used Selenocysteine in place of Cys, and nitric oxide, cyanide, and azide in place of oxygen. Abs, MCD, rRaman, and EPR spectroscopies have all proven to be useful tools for studying these CDO-complex mimics, particularly when used in conjunction with computational models of the enzyme.

As-isolated CDO exists as a mixture of Fe(II) and Fe(III)-bound active sites. Cys-bound Fe(II) CDO is 5-coordinate, with an open coordination site for O_2. In contrast, Cys-Fe(III) CDO is a 6-coordinate complex with direct sulfur ligation and a hydroxide in the position where O_2 binds in the reduced state[18]. MCD spectra of Cys-bound CDO exhibit signals for both the Cys→Fe(III) charge transfer (CT) transitions (lower energy features) and the Cys→Fe(II) CT transitions (higher energy features). Comparing spectra of CDO with native substrate cysteine versus substrate analogue selenocysteine (Sec), which the enzyme can bind but not turn over, allows for the unambiguous identification of S/Se→Fe CT bands. Only the MCD features associated with Cys-Fe(II) CDO disappear after the addition of O_2, indicating Cys turnover and thus that Fe(II) CDO is the catalytically relevant form of the enzyme.

In addition to binding cysteine and cysteine analogues to CDO for spectroscopic analysis, O_2 analogues have also been used. One of these is nitric oxide, NO. NO shows limited binding to CDO unless cysteine is already bound, indicating a required ordered binding of substrate before O_2.

Cyanide has also been used as an O_2 analogue, in particular to investigate the role of the Cys-Tyr crosslink. Like NO, cyanide cannot bind to the CDO active site unless cysteine is already bound. EPR spectra of Fe(II) CDO species are difficult to obtain, since large spin-orbit coupling leads to zero-field splitting for the non-Kramers doublet that exceeds the microwave radiation of the instrument. Instead, EPR spectra of (CN^-/Cys)-Fe(III) CDO were taken, which mimics a putative Fe(III)-superoxo intermediate. Cyanide has strong pi-accepting character, which forces the iron complex to adopt a low spin ground state. This gives rise to a rhombic S=1/2 EPR signal that is sensitive to the presence of the crosslink. Fully crosslinked *Mm*CDO exhibits a wide spread of g-values, a fully non-crosslinked variant shows a narrower spread of g-values, and as-isolated CDO, which is a roughly 50/50 mix of crosslinked and non-crosslinked enzyme, displays a mix of the two signals[19].

To further investigate if a superoxo-Fe(III) complex is involved in the CDO reaction mechanism, azide was used as a superoxide mimic due to its similar frontier orbitals and same charge. Unlike NO and CN^-, which require pre-bound cysteine in order to coordinate to the iron center of CDO, azide is capable of binding to the iron center in the absence of the substrate cysteine. Hydroxide was shown computationally to occupy another coordination site of iron upon azide binding to Fe(III) CDO in the absence of cysteine, so that there are two anionic ligands. However, it is unclear if this complex is 5- or 6-coordinate. When azide is added to Cys-Fe(III)CDO, however, it does not bind directly to the iron center in the majority of active sites. Azide binding slightly changes the electronic structure of the iron, though, indicating it binds somewhere in or near the active site pocket. A small fraction of azide is able to bind directly to the iron when azide is present in large molar excess to yield a low spin Fe(III) complex[20].

Bacterial CDOs

Until recently it was believed that prokaryotes lacked CDO, and this pathway of cysteine degradation was therefore omitted from schemes of prokaryotic cysteine metabolism. However,

in 2006, Stipanuk and coworkers used sequence alignment tools to identify a total of 38 putative bacterial CDOs. Four of these enzymes were heterologously expressed in *E. coli* and confirmed to have CDO activity[21]. One of these four, the CDO from *Bacillus subtilis*, was successfully crystallized and its structure solved[22] (PDB= code 4QM9, **Figure 1**). The active site of *Bs*CDO is nearly identical to that of mammalian CDO, except that the position corresponding to C93 is occupied by a glycine (G82), and the enzyme is therefore incapable of forming an active site C-Y crosslink. Despite the lack of this crosslink, *Bs*CDO is as active as crosslinked mammalian CDO[21]. Comparing these two naturally occurring forms of the enzyme is a promising opportunity to study the formation and functional role of the C-Y crosslink found in eukaryotic CDOs.

When bacterial CDOs were first discovered, it was thought that there were two types, “Arg-type” and “Glu-type” bacterial CDOs, which differ in having either an Arg or a Glu at the position equivalent to R60 in *Mm*CDO. However, it was later realized that Glu-type bacterial CDO is actually an MDO[22], which catalyzes the oxidation of 3-MPA. R60 is thought to play an important role in substrate positioning in eukaryotic CDO[14], as R60 is positioned corrected to form a salt bridge with the carboxylate group of the substrate Cys. In 2019, Jameson and coworkers directly compared *Mm*CDO and *Pa3*MDO along with several variants to investigate the role of specific active site residues in the binding and turnover of Cys and 3-MPA[23]. They found that permutations of three specific active site features controlled the enzyme’s preference for one substrate over another: (1) the R or Q at position 60, (2) a cis-peptide bond between residues S158 and P159 that rotates the 159-168 loop in closer to the active site, and (3) a C or R at position 164 that, when the cis-peptide bond is not present, is positioned for potential salt bridge formation with the carboxylate group of substrate 3-MPA (all *Mm*CDO numbering). *Mm* and *Bs*CDO both have (1) R60 and (2) the cis-peptide bond, but differ in (3) the residue at position 164 (C for *Mm*CDO, M for *Bs*CDO). Residue 164 sits very near the substrate channel into the active site, and thus a comparison of these two WT CDOs would afford a better

understanding of the specific role this region of the enzyme plays in determining substrate specificity and reactivity.

Research Overview

In this dissertation, I present my research into the unique structural features of CDO and their role in determining the substrate scope and reactivity of the enzyme. I employ a combined spectroscopic and kinetic methodology that allows for observation of the direct effects of metal-center and active site structure on enzyme function. This is then contextualized using modern bioinformatic tools in the framework of the larger family of CDO homologues.

- Chapter 2: The G82C variant of *Bs*CDO was generated via a single DNA point mutation. Unlike the WT enzyme, this variant displays the ability to form an active site C-Y crosslink analogous to that found in eukaryotic forms of CDO. Crosslink formation was confirmed via SDS-PAGE gel electrophoresis, mass spectrometry, and Cys/CN^--Fe(III)CDO EPR spectroscopy. All three experiments allowed for a quantitative measurement of the ratio of crosslinked to non-crosslinked protein, and this ratio could be altered by changing the amount of iron in the growth media during protein expression. A UPLC-MS assay and a colorimetric assay using Ellman's reagent were employed to determine the catalytic efficiency of the G82C variant in comparison to WT *Bs*CDO and WT *Mm*CDO. Crosslink formation in the G82C *Bs*CDO had a similar effect on activity as it does in WT *Mm*CDO. A sequence similarity network of the CDO family was generated that revealed a significant number of putatively crosslinked bacterial CDOs. Many of these were from organisms that are eukaryotic pathogens, suggesting potential gene transfer from a eukaryotic host.

- Chapter 3: Various substrate analogue-bound complexes of WT *Mm* and *Bs*CDO were compared spectroscopically and kinetically to investigate the origin of substrate specificity in *Mm*CDO. G82C *Bs*CDO was included to isolate effects of the C-Y active site crosslink. MCD spectroscopy showed that the Fe(II) state of *Bs*CDO can accommodate bulkier substrates than *Mm*CDO. UPLC-MS and NMR assays of substrate analogue turnover by *Mm* and *Bs*CDO indicate that *Bs*CDO is better able to turnover non-native substrates, if only to a small degree. EPR spectroscopy revealed that Fe(III)*Bs*CDO has a much more flexible active site than *Mm*CDO, but introduction of a crosslink via the G82C mutation only removed some of that additional flexibility. *Bs*CDO's comparatively weaker binding of substrate analogue 2-AET suggests that the extra flexibility not accounted for by the absence of crosslink is in the region of the active site around where the amine group of substrate binds. A comparison of the crystal structures of *Mm* and *Bs*CDO show that the active site wall in this area has no secondary structure, but *Mm*CDO has a small beta sheet on the C-terminus of the protein that sits on top of this flexible active site wall.

- Chapter 4: This chapter is provided as a resource for future students on the many bioinformatic tools currently available for their use. More detail is given as to how each was used in the context of studying CDO, as well as background information on the programs themselves, instructional details for their use, and references to helpful resources, files, and example MATLAB codes.

References

(1) Lombardini, J. B.; Singer, T. P.; Boyer, P. D. Cysteine Oxygenase II. Studies on the Mechanism of the Reaction with 18Oxygen. *J. Biol. Chem.* **1969**, *244* (5), 1172–1175. https://doi.org/https://doi.org/10.1016/S0021-9258(18)91825-9.

(2) Stipanuk, M. H.; Dominy, J. E.; Lee, J.-I.; Coloso, R. M. Mammalian Cysteine Metabolism: New Insights into Regulation of Cysteine Metabolism. *J. Nutr.* **2006**, *136* (6), 1652S-1659S. https://doi.org/https://doi.org/10.1093/jn/136.6.1652S.

(3) Lehmann, A.; Hagberg, H.; Orwar, O.; Sandberg, M. Cysteine Sulphinate and Cysteate: Mediators of Cysteine Toxicity in the Neonatal Rat Brain? *Eur. J. Neurosci.* **1993**, *5* (10), 1398–1412. https://doi.org/10.1111/j.1460-9568.1993.tb00926.x.

(4) Montine, T. J.; Picklo, M. J.; Amarnath, V.; Whetsell, W. O.; Graham, D. G. Neurotoxicity of Endogenous Cysteinylcatechols. *Exp. Neurol.* **1997**, *148* (1), 26–33. https://doi.org/10.1006/exnr.1997.6662.

(5) Fjellstedt, E.; Harnevik, L.; Jeppsson, J.-O.; Tiselius, H.-G.; Söderkvist, P.; Denneberg, T. Urinary Excretion of Total Cystine and the Dibasic Amino Acids Arginine, Lysine and Ornithine in Relation to Genetic Findings in Patients with Cystinuria Treated with Sulfhydryl Compounds. *Urol. Res.* **2003**, *31* (6), 417–425. https://doi.org/10.1007/s00240-003-0366-6.

(6) Thomas Heafield, M.; Fearn, S.; Steventon, G. B.; Waring, R. H.; Williams, A. C.; Sturman, S. G. Plasma Cysteine and Sulphate Levels in Patients with Motor Neurone, Parkinson's and Alzheimer's Disease. *Neurosci. Lett.* **1990**, *110* (1–2), 216–220. https://doi.org/https://doi.org/10.1016/0304-3940(90)90814-P.

(7) Emery, P.; Salmon, M.; Bradley, H.; Wordsworth, P.; Tunn, E.; Bacon, P. A.; Waring, R. Genetically Determined Factors as Predictors of Radiological Change in Patients with Early Symmetrical Arthritis. *Br. Med. J.* **1992**, *305* (6866), 1387–1389. https://doi.org/10.1136/bmj.305.6866.1387.

(8) Davies, M. H.; Ngong, J. M.; Pean, A.; Vickers, C. R.; Waring, R. H.; Elias, E. Sulphoxidation and Sulphation Capacity in Patients with Primary Biliary Cirrhosis. *J. Hepatol.* **1995**, *22* (5), 551–560. https://doi.org/10.1016/0168-8278(95)80450-1.

(9) Bradley, H.; Gough, A.; Sokhi, R. S.; Hassell, A.; Waring, R.; Emery, P. Sulfate Metabolism Is Abnormal in Patients with Rheumatoid Arthritis. Confirmation by in Vivo Biochemical Findings. *J. Rheumatol.* **1994**, *21* (7), 1192–1196. https://doi.org/10.1016/s0950-3579(05)80178-5.

(10) Finkelstein, J. D.; Kyle, W. E.; Harris, B. J.; Martin, J. J. Methionine Metabolism in Mammals: Concentration of Metabolites in Rat Tissues. *J. Nutr.* **1982**, *112* (5), 1011–1018. https://doi.org/10.1093/jn/112.5.1011.

(11) Stipanuk, M. H.; Ueki, I.; Stipanuk, M. H.; Ueki, I. Dealing with Methionine/Homocysteine Sulfur: Cysteine Metabolism to Taurine and Inorganic Sulfur. *J. Inherit. Metab. Dis.* **2011**, *34* (1), 17–32. https://doi.org/10.1007/s10545-009-9006-9.

(12) Lee, J. I.; Londono, M.; Hirschberger, L. L.; Stipanuk, M. H. Regulation of Cysteine Dioxygenase and γ-Glutamylcysteine Synthetase Is Associated with Hepatic Cysteine Level. *J. Nutr. Biochem.* **2004**, *15* (2), 112–122. https://doi.org/10.1016/j.jnutbio.2003.10.005.

(13) Dominy, J. E.; Hirschberger, L. L.; Coloso, R. M.; Stipanuk, M. H. In Vivo Regulation of Cysteine Dioxygenase via the Ubiquitin-26S Proteasome System. In *Advances in Experimental Medicine and Biology*; Springer New York, 2006; Vol. 583, pp 37–47. https://doi.org/10.1007/978-0-387-33504-9_4.

(14) Mccoy, J. G.; Bailey, L. J.; Bitto, E.; Bingman, C. A.; Aceti, D. J.; Fox, B. G.; Phillips, G. N. Structure and Mechanism of Mouse Cysteine Dioxygenase. *PNAS* **2006**, *103* (9), 3084–3089. https://doi.org/https://doi.org/10.1073/pnas.0509262103.

(15) Driggers, C. M.; Cooley, R. B.; Sankaran, B.; Hirschberger, L. L.; Stipanuk, M. H.; Karplus, P. A. Cysteine Dioxygenase Structures from PH 4 to 9: Consistent Cys-

Persulfenate Formation at Intermediate PH and a Cys-Bound Enzyme at Higher PH. *J. Mol. Biol.* **2013**, *425* (17), 3121–3136. https://doi.org/10.1016/j.jmb.2013.05.028.

(16) Dominy, J. E.; Hwang, J.; Guo, S.; Hirschberger, L. L.; Zhang, S.; Stipanuk, M. H. Synthesis of Amino Acid Cofactor in Cysteine Dioxygenase Is Regulated by Substrate and Represents a Novel Post-Translational Regulation of Activity. *J. Biol. Chem.* **2008**, *283* (18), 12188–12201. https://doi.org/10.1074/jbc.M800044200.

(17) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of the H155A Variant of Cysteine Dioxygenase: Geometric and Electronic Consequences of a Third-Sphere Amino Acid Substitution. *Biochemistry* **2015**, *54* (18), 47. https://doi.org/10.1021/acs.biochem.5b00171.

(18) Gardner, J. D.; Pierce, B. S.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Characterization of Substrate-Bound Mouse Cysteine Dioxygenase: Nature of the Ferrous and Ferric Cysteine Adducts and Mechanistic Implications. *Biochemistry* **2010**, *49* (29), 6033–6041. https://doi.org/10.1021/bi100189h.

(19) Li, W.; Blaesi, E. J.; Pecore, M. D.; Crowell, J. K.; Pierce, B. S. Second-Sphere Interactions between the C93−Y157 Cross-Link and the Substrate-Bound Fe Site Influence the O2 Coupling Efficiency in Mouse Cysteine Dioxygenase. *Biochemistry* **2013**, *52* (51), 9104–9119. https://doi.org/10.1021/bi4010232.

(20) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of Iron(III) Cysteine Dioxygenase: Implications for the Nature of the Putative Superoxo-Fe(III) Intermediate. *Biochemistry* **2014**, *53* (36), 5759–5770. https://doi.org/10.1021/bi500767x.

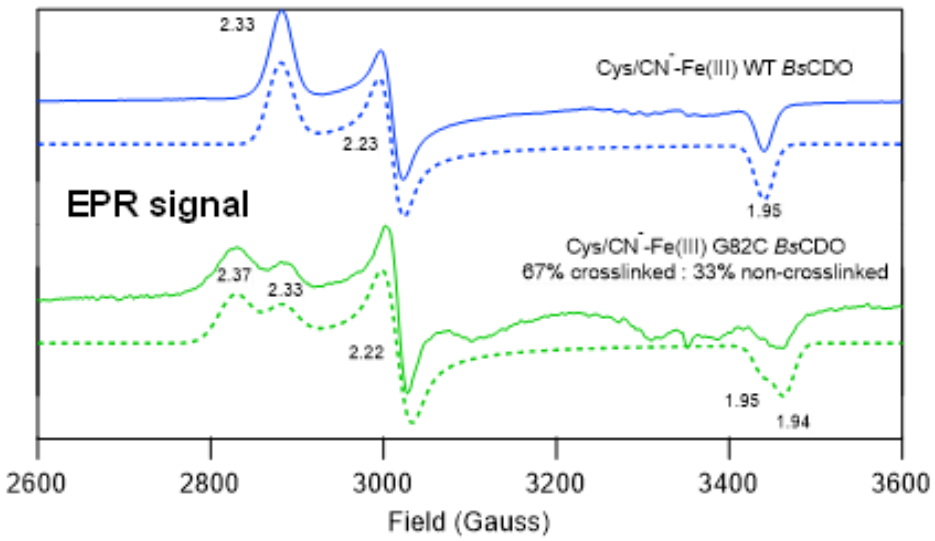
(21) Dominy, J. E.; Simmons, C. R.; Karplus, P. A.; Gehring, A. M.; Stipanuk, M. H. Identification and Characterization of Bacterial Cysteine Dioxygenases: A New Route of Cysteine Degradation for Eubacteria. *J. Bacteriol.* **2006**, *188* (15), 5561–5569. https://doi.org/10.1128/JB.00291-06.

(22) Driggers, C. M.; Hartman, S. J.; Karplus, P. A. Structures of Arg- and Gln-Type Bacterial Cysteine Dioxygenase Homologs. *Protein Sci.* **2015**, *24* (1), 154–161. https://doi.org/10.1002/pro.2587.

(23) Aloi, S.; Davies, C. G.; Karplus, P. A.; Wilbanks, S. M.; Jameson, G. N. L. Substrate Specificity in Thiol Dioxygenases. *Biochemistry* **2019**, *58* (19), 2398--2407. https://doi.org/10.1021/acs.biochem.9b00079.

CHAPTER 2

A Single DNA Point Mutation Leads to the Formation of a Cysteine-Tyrosine Crosslink in the Cysteine Dioxygenase from *Bacillus subtilis*



This chapter has been accepted for publication under the following: Schultz, R. L.; Sabat, G.; Fox, B. G.; Brunold, T. C. A single DNA point mutation leads to the formation of a cysteine-tyrosine crosslink in the cysteine dioxygenase from *Bacillus subtilis*. *Biochemistry* **2023**. https://doi.org/10.1021/acs.biochem.3c00083.

Chapter 2 : A Single DNA Point Mutation Leads to the Formation of a Cysteine-Tyrosine Crosslink in the Cysteine Dioxygenase from *Bacillus subtilis*

2.1 Introduction

Cysteine dioxygenase (CDO) is a member of the thiol dioxygenases, a family of non-heme iron-containing cupin enzymes that oxidize thiols to their corresponding sulfinates.[24] This family additionally includes cysteamine dioxygenase (ADO), 3-mercaptopropionic acid dioxygenase (MDO), mercaptosuccinate dioxygenase (MSDO),[24] and plant cysteine dioxygenase (PCO).[25,26] CDO catalyzes the oxidation of cysteine (Cys) to cysteine sulfinic acid (CSA) via incorporation of molecular oxygen.[1] Cys is an essential protein building block as well as a precursor for the biosynthesis of the metabolites coenzyme A, glutathione, taurine, and pyruvate.[2] However, excess free Cys is cytotoxic[3] and neurotoxic,[4] is easily oxidized to the insoluble cystine,[5] and has been linked to several autoimmune and neurodegenerative diseases.[6–9] Cys oxidation is a key enzymatic step in the sulfur metabolic pathway and therefore plays a vital role in maintaining homeostasis of sulfur-containing compounds.[24]

The active site of CDO contains several conserved residues that coordinate the iron atom and promote the binding and activation of Cys and O_2. X-ray crystal structures revealed a three-histidine (3-His) coordination of the iron ion instead of the 2-His, 1-carboxylate facial triad typical of non-heme iron proteins.[14] In the resting state of CDO, the Fe(II) ion resides in a 6-coordinate pseudo-octahedral coordination environment comprised of the three His residues and three water molecules.[27] Structures of CDO crystallized in the presence of Cys revealed that substrate displaces all three active site water molecules to coordinate to the iron in a bidentate fashion via its sulfur and amine groups. The sixth, open coordination site is then available for O_2 to bind end-on to the iron, facilitated by a local hydrophobic pocket.[15]

Eukaryotic CDO additionally contains a unique thioether crosslink between the sulfur of C93 and an ortho carbon of Y157 (*Mus musculus* CDO, *Mm*CDO numbering).[14] A similar

crosslink has been observed by X-ray crystallography in only a small number of other enzymes, including galactose oxidase and the bacterial sulfite reductase NirA.[28,29] Iron, molecular oxygen, and substrate Cys must all be present for the CDO crosslink to form, and ~50% of the protein population forms this crosslink after about 800 catalytic turnovers.[16] Crosslink formation increases the catalytic efficiency of the enzyme by at least 10-fold. The crosslinked "mature" and non-crosslinked "immature" forms of eukaryotic CDO travel as two distinct bands on an SDS-PAGE gel.[16] The exact role of the crosslink in CDO catalysis is not fully understood. X-ray crystal structures of crosslinked wild type (WT) and the non-crosslinked C93G *Mm*CDO are essentially identical, indicating that the crosslink does not serve to stabilize an otherwise unfavorable protein conformation.[30] Also, the C93G *Mm*CDO variant, which is incapable of C–Y crosslink formation, has comparable activity to the crosslinked WT enzyme.[30] This led to the hypothesis that the crosslink serves not to directly improve enzyme efficiency, but rather to prevent deleterious effects associated with the presence of an untethered Cys residue in the enzyme active site.[30]

Using a combination of spectroscopic and computational techniques, we identified H155 in *Mm*CDO as an important residue in maintaining proper positioning of the C93–Y157 crosslink. Residue H155 is a member of the S153-H155-Y157 catalytic triad motif[14] (**Figure 1**) that is highly conserved among CDOs and crucial for enzymatic activity, as evidenced by the fact that the H155A *Mm*CDO variant is ~100-fold less active than the WT enzyme.[19] Exposure of as-isolated H155A *Mm*CDO (which contained similar amounts of Fe(II) and Fe(III)) to Cys led to the appearance of magnetic circular dichroism (MCD) features similar to those observed for Cys-bound WT CDO; however, the features of the Cys- and selenocysteine (Sec)-bound H155A Fe(II)*Mm*CDO fractions were markedly blue-shifted from their WT counterparts.[17] These shifts, in conjunction with results obtained from quantum mechanics/molecular mechanics (QM/MM) computations, suggested that in H155A *Mm*CDO a six-coordinate (H_2O/Cys)-Fe(II) complex is stabilized by ~39 kcal/mol over its five-coordinate, Cys-only bound analogue, while in WT Cys-

Fe(II)*Mm*CDO the Fe(II) center favors a five-coordinate ligand environment.[17] The stabilization of a six-coordinate (H_2O/Cys)-Fe(II) complex in H155A *Mm*CDO was attributed to the increased conformational freedom of the C93–Y157 crosslink in the absence of H155, thus allowing the C93 sulfur atom to reposition itself so as to accept a hydrogen bond from a coordinated water molecule. Because in non-crosslinked eukaryotic WT CDOs the C93 residue is granted even more conformational freedom than in the H155A variant, we proposed that the low activity of non-crosslinked *Mm*CDO arises from the formation of a similar six-coordinate (H_2O/Cys)-Fe(II)CDO complex.[17]

Selective evolutionary pressure to conserve C93 in eukaryotic CDOs is curious considering this residue decreases catalytic efficiency unless it is crosslinked to Y157. However, Dominy et al. rationalized the conservation of C93 as an additional level of CDO regulation that provides fine-tuned control of intracellular Cys concentration.[31] When Cys levels increase, more "mature" enzyme is produced as crosslink forms during turnover and catalytic efficiency improves. Additionally, at high levels, Cys blocks the ubiquitination and proteasomal degradation of CDO, thus increasing the intracellular CDO concentration ~10-fold. Together, these two layers of regulation give eukaryotes the capacity to drastically increase CDO activity in response to a sudden rise in intracellular Cys levels.[31]

Until quite recently, it was believed that only eukaryotes can degrade Cys to CSA. However, in 2006 Stipanuk and coworkers identified numerous bacterial enzymes that could potentially possess CDO activity.[21] Four members of this family were heterologously expressed and found to oxidize Cys to CSA.[21] Although bacterial CDOs generally have overall sequence identities of less than 30% compared with *Mm*CDO, many of the active site residues are conserved, the most notable exception being C93 (*Mm*CDO numbering). In bacterial CDOs, this position is occupied by a highly conserved glycine (**Figure 1**), which implies that these enzymes are not capable of forming a C–Y crosslink. Indeed, an X-ray crystal structure of the *Bacillus subtilis* CDO (*Bs*CDO) revealed that the tyrosine residue corresponding to Y157 in mouse CDO

is unmodified in the bacterial enzyme.[22] With this important exception, the *Bs*CDO active site structure is almost identical to that of mammalian crosslinked CDO, even in the Cys-bound state (**Figure 1**). Despite the drastically decreased activity of non-crosslinked eukaryotic CDOs, bacterial CDOs achieve turnover rates paralleling those of fully crosslinked eukaryotic CDOs.[21]

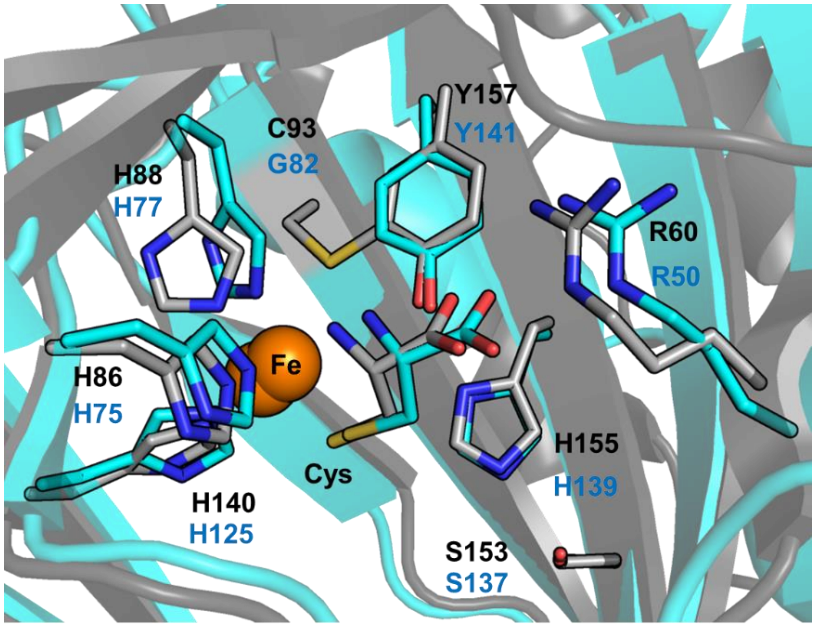


Figure 2.1 Overlay of the active site regions of Cys-bound MmCDO (gray, PDB ID 4JTO) and BsCDO (cyan, PDB ID 4QM9).

Alignment of known CDO sequences revealed a seemingly new category of bacterial CDOs in which the R60 residue (*Mm*CDO numbering) is substituted by a Q. These "Gln-type" CDOs were subsequently found to display increased catalytic efficiency for the conversion of 3-mercaptopropionic acid (3-MPA) to 3-sulfinopropionate than for Cys oxidation and were thus reclassified as MDOs.[22] Jameson and coworkers successfully produced a G95C variant of *Pseudomonas aeruginosa* MDO (*Pa*MDO) capable of forming a crosslink analogous to that found in mammalian CDOs.[32] X-ray crystal structures revealed that when non-crosslinked, the Cys thiol excludes Y159 from its native position. A kinetic analysis showed that the K_M remained relatively unchanged between WT and ~50% crosslinked G95C *Pa*MDO for both the native substrate 3-mercaptopropionic acid and the non-native substrate Cys.[32] In contrast, the k_{cat}

values for both substrates were significantly smaller for the G95C variant compared to WT *Pa*MDO, even though half of the active sites contained the crosslink. The loss of activity of the non-crosslinked isoform of G95C *Pa*MDO was attributed to the mispositioning of Y159 observed in the crystal structure, but the dramatic decrease in activity of the crosslinked fraction remained unexplained.[32]

In the present study, we prepared the G82C variant of *Bs*CDO to determine if a single DNA point mutation could introduce into this enzyme the ability to form a C–Y crosslink. We used gel electrophoresis, mass spectrometry, electron paramagnetic resonance (EPR) spectroscopy, and kinetic assays to characterize this variant along with the natively crosslinked WT *Mm*CDO and the natively non-crosslinked WT *Bs*CDO. Finally, bioinformatic tools were used to identify trends of C–Y crosslink conservation across the entire CDO family.

2.2 Materials and Methods

Recombinant Gene Expression and Protein Purification. Gene expression and protein purification of WT *Bs*CDO were conducted as described previously[21] with a few minor alterations. In brief, a codon-optimized *cdoA* gene was purchased from Integrated DNA Technologies and inserted into a pQE-30 expression vector using restriction digest with BamHI-HF and HindIII-HF and subsequent ligation. Insertion was confirmed via colony PCR and the gene sequence was verified by Sanger sequencing at the University of Wisconsin-Madison Biotechnology Center. The plasmid was then transformed into *Escherichia coli* Rosetta 2(DE3) cells, the cells were grown in a modified TB+G medium (12 g/L tryptone, 24 g/L yeast extract, 8 mL/L glycerol with 100 mL/L 0.17 M KH_2PO_4/0.72 M K_2HPO_4, 25 mL/L 15% aspartate, and 2 mL/L 1 M $MgSO_4$) at 37 °C and 250 rpm, and the *cdoA* gene was expressed via induction with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at an OD_{600} of 8.0. Ferrous ammonium sulfate (FAS) was added to a final concentration of 500 μM at the time of induction to increase Fe incorporation into the CDO active site. Cells grew for an additional four hours after induction. Filtered cell lysate in IMAC A buffer (20 mM Tris, 5 mM imidazole, 500 mM NaCl, pH 8.0) was

applied to an immobilized metal affinity chromatography column and eluted with an increasing gradient of IMAC B (20 mM Tris, 500 mM imidazole, 500 mM NaCl, pH 8.0) buffer. Fractions containing CDO as determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) were pooled and activity of the purified enzyme was confirmed qualitatively using thin-layer chromatography as previously described.[18]

The G82C mutation was introduced into the *cdoA* gene using site-directed mutagenesis with the primers 5'-GGCAGAGTATTTGTTGCGCCATGG-3' and 5'-CCATGGCGCAACAAATACTCTGCC-3' purchased from Integrated DNA Technologies. Correct mutagenesis was confirmed via Sanger sequencing at the University of Wisconsin-Madison Biotechnology Center and variant protein was produced as described above for WT protein. Four preparations of G82C *Bs*CDO were completed with different media and supplements to achieve variable iron loading and thus varying degrees of crosslink formation. Preps 1 and 2 were done in the modified TB+G medium described above. Preps 3 and 4 were carried out in LB medium, with Prep 3 induced at an OD_{600} of ~4 and Prep 4 induced at an OD_{600} of ~0.8. Preps 2, 3, and 4 were supplemented 15 min before induction with 1,10-phenanthroline to a final concentration of 100 mM in 100 mM HCl to prevent iron binding to the enzyme, as previously done by Ellis and coworkers.[33] The rest of the growth and purification was performed as described above for WT *Bs*CDO.

WT *Mm*CDO was produced using a codon-optimized *cdo1* gene in the pVP16 expression vector with the gene for an attached maltose binding protein (MBP) as a solubility tag. WT *Mm*CDO Prep 1 was purified from *Escherichia coli* Rosetta 2(DE3) cells containing the pVP16 expression vector with the *cdo-mbp* gene grown in LB medium at 37 °C and 225 rpm. At an OD_{600} of 1.0, FAS was added to a final concentration of 520 μM. At an OD_{600} of 2.73, expression was induced via the addition of 10 μM IPTG and the medium was supplemented with another addition of FAS to 50 μM along with 3 g/L lactose and 2 g/L Casamino acids. Cells were left to grow overnight at 25 °C and 225 rpm. Filtered cell lysate in 25 mM HEPES, 300 mM

NaCl, pH 7.9 buffer was applied to a TALON column. Fractions containing CDO-MBP as determined by SDS-PAGE were pooled. Approximately 1 mg of tobacco etch virus protease was added to the pooled fractions per 50 mg of protein to cleave MBP from CDO, and the solution was dialyzed overnight against 2 L of the HEPES/NaCl buffer. The cleaved and dialyzed protein was then again applied to a TALON column. Fractions containing purified CDO as determined by SDS-PAGE were pooled and the activity of the purified enzyme was confirmed qualitatively using thin-layer chromatography as previously described.[18] C93G *Mm*CDO was produced and purified following the same procedure as was used for WT *Mm*CDO Prep 1.

WT *Mm*CDO Prep 2 was purified from *Escherichia coli* Rosetta 2(DE3) cells containing the pVP-16 *cdo-mbp* expression vector grown in the modified TB+G medium described above at 37 °C and 250 rpm. At an OD_{600} of 8.0, expression was induced with the addition of IPTG to 1 mM and FAS was added to 500 μM. Cells were grown for an additional 4 hours after induction. Filtered cell lysate in the Tris IMAC A buffer described above was applied to a TALON column and fractions containing CDO-MBP as determined by SDS-PAGE were pooled. The rest of the purification procedure was identical to that used for WT *Mm*CDO Prep 1 described above.

SDS-PAGE Densitometry. Purified CDO samples were run on Criterion TGX Stain-Free Precast Gels at 200 V for 42 min. Gels were stained with Coomassie Brilliant Blue G-250 and then destained in a solution of 50%:40%:10% (v/v) methanol:H_2O:acetic acid. The gels were imaged using a photo scanner. ImageJ software was employed to quantify the intensities of the two SDS-PAGE gel bands in each sample, and percent crosslinking was calculated as the ratio of the intensities of the bottom and top bands.

Mass Spectrometry. Gel pieces were de-stained completely in 100 mM NH_4HCO_3 in 50%:50% (v/v) MeOH:H_2O and dehydrated in 25 mM NH_4HCO_3 in 50%:50% (v/v) CH_3CN:H_2O and then again in 100% CH_3CN. The samples were dried in a Speed-Vac, reduced with 25 mM dithiothreitol in 25 mM NH_4HCO_3 at 56 °C, alkylated with 55 mM chloroacetamide in 25 mM

NH_4HCO_3 in darkness at room temperature, washed once in H_2O, then dehydrated as above. Primary digestion was performed by rehydrating samples with 10 ng/µL trypsin in 25 mM NH_4HCO_3/0.01% (w/v) of ProteaseMAX™ from Promega Corp for 3 hours at 42 °C. Secondary digestion was performed for 1 hour at 37 °C using 20 ng/µL endoproteinase AspN from Roche Diagnostics in 25 mM NH_4HCO_3. Proteolysis was terminated by acidification with 2.5% trifluoroacetic acid (TFA) to 0.3% (v/v). Degraded ProteaseMAX™ was removed via centrifugation and the peptides were solid phase extracted on a C18 column (Pierce™ C18 SPE tips, 10 µl bed). Peptides were eluted with acetonitrile/H_2O/TFA (70%:30%:0.1% (v/v)), dried to minimum volume, and diluted to 20 µL total volume with 0.1% formic acid.

Peptides were analyzed by nanoLC-MS/MS using the Agilent 1100 nanoflow system connected to a hybrid linear ion trap-orbitrap mass spectrometer (LTQ-Orbitrap Elite™, Thermo Fisher Scientific) equipped with an EASY-Spray™ electrospray source (held at constant 35 °C). 2 uL of extracted peptides were loaded onto a capillary emitter column (PepMap® C18, 3µM, 100Å, 150 × 0.075mm, Thermo Fisher Scientific). A NanoHPLC system delivered solvents A (0.1% (v/v) formic acid) and B (99.9% (v/v) acetonitrile). Peptides were loaded at 100% A and eluted by gradually increasing percent B directly into the nano-electrospray. As peptides eluted from the HPLC-column/electrospray source, survey MS scans were acquired in the Orbitrap with a resolution of 120,000 followed by collision-induced dissociation (CID)-type MS/MS with 2.0 AMU isolation and 10 msec activation time with 35% normalized collision energy fragmentation of the 30 most intense peptides detected in the MS1 scan from 350 to 1800 m/z; redundancy was limited by dynamic exclusion. Monoisotopic precursor selection and charge state screening were enabled and +1 as well as undefined charge states were rejected.

Raw MS/MS data were searched against a user defined *Bacillus subtilis* amino acid sequence database (UP00001570 reference proteome, 02/19/2021 download plus CDOG82C construct of interest, 5,488 total sequences) or a Uniprot *Mus musculus* reference database (UP000000589, 06/16/2020 download, 63723 total sequences) plus a common lab

contaminants cRAP database (117 total entries) using the in-house Mascot search engine 2.7.0 (Matrix Science) with variable methionine oxidation, asparagine and glutamine deamidation, plus fixed Cys carbamidomethylation. Peptide mass tolerance was set at 10 ppm and fragment mass at 0.6 Da. Protein annotations, significance of identification, and spectral based quantification were done with the help of the Scaffold software (version 4.11.0, Proteome Software Inc., Portland, OR). Peptide identifications were accepted if they could be established at greater than 64.0% probability to achieve a false discovery rate (FDR) of less than 1.0% by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at greater than 14.0% probability to achieve an FDR of less than 1.0% and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm.[34] Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters.

Sample Preparation for Spectroscopy. Protein samples for EPR spectroscopy were prepared in a buffer of 20 mM Tris, 5 mM imidazole, 500 mM NaCl, pH 8.0. The Fe(II) and Fe(III) contents of the protein were determined via a colorimetric assay with the iron chelator tripyridyl triazine (TPTZ) and an ε_{595} of 22.1 mM^{-1} cm^{-1}.[35] To increase the iron content, protein samples were incubated anaerobically with a 2-fold molar excess of FAS to protein for 30 min. Chelex 100 was added to remove unbound iron from solution. Typical iron incorporation was ~60% after reconstitution. Ammonium hexachloroiridate was added in a 3-fold molar excess over iron-loaded protein to oxidize it to the Fe(III) state. Excess oxidant was removed using either a PD-10 desalting column or buffer exchange through a 10 kDa Centricon filter. EPR samples were prepared with 0.5 mM Fe(III)-loaded protein, a 10-fold molar excess of Cys, and a 15-fold molar excess of KCN.

Protein samples for MCD spectroscopy were prepared anaerobically in a glove box. Protein in a buffer of 20 mM Tris, pH 8.0, with 5 mM imidazole and 500 mM NaCl was

reconstituted with FAS as described above. Tris(2-carboxyethyl)phosphine (TCEP) was added in a 3-fold molar excess over iron-loaded protein to reduce it to the Fe(II) state. Samples were then concentrated to 1.4 mM Fe(II)-loaded protein and incubated with a 2-fold molar excess of Cys. 55% (v/v) glycerol was added as a glassing agent.

Spectroscopy. X-band EPR data were collected using a Bruker ELEXSYS E500 spectrometer. Sample temperature was maintained at 20 K by an Oxford ESR 900 continuous flow liquid He cryostat regulated by an Oxford ITC-503S temperature controller. All EPR spectra were obtained using the following experimental parameters: frequency = 9.381 GHz; microwave power = 12.62 mW; modulation amplitude = 3 G; and modulation frequency = 100 kHz. EPR spectral fits were performed using the SIMPOW program.[36]

Low-temperature MCD spectra were recorded with a Jasco J-715 spectropolarimeter in conjunction with an Oxford Instruments SpectromagPT 7 Tesla magnetocryostat. MCD spectra are presented as the difference between spectra obtained with the magnetic field aligned parallel and antiparallel to the light propagation axis to eliminate contributions from the circular dichroism background and glass strain.

Kinetic Assays. Quantitative kinetic assays were performed using a combination of two methods. The first method was adapted from Stipanuk and coworkers and Jameson and coworkers and employed an Intrada column from Imtakt on a Waters ultra-high performance liquid chromatography mass spectrometer (UPLC-MS).[30,37,38] Protein samples were thawed in hand and then exchanged into a buffer of 200 mM 2-(N-morpholino)ethanesulfonic acid (MES), 150 mM NaCl, 0.1 mM bathocuproine disulfonate (BCS), pH 6.1. The concentration of Fe(II)-loaded protein was determined via the TPTZ assay described above.[35] 100 µL reactions of 50 µM protein and 1-50 mM cysteine were run at 37 °C for 12 min. 20 µL reaction aliquots were quenched every 3 min by the addition of 180 µL of equal volumes of 1 M HCl and 1 mM asparagine, which served as an internal standard. Quenched reaction aliquots were centrifuged at 15,000 × g for 3 min and the supernatants were transferred to a 96-well plate. 2 µL of

samples were injected onto the UPLC-MS and applied to the Intrada column under 80% buffer A (acetonitrile with 0.1% (v/v) formic acid), 20% buffer B (100 mM ammonium formate) and species eluted during a gradual increase to 80% buffer B were analyzed with the mass detector. Enzyme activity was quantified using the ratio of CSA mass peak to asparagine mass peak in comparison with a calibration curve. Values for k_{cat} and K_M were determined using the concentration of Fe(II)-loaded enzyme and the Michaelis-Menten analysis reported by Johnson.[39]

The second method was adapted from Jameson and coworkers.[40] Protein was exchanged into a buffer of 100 mM sodium phosphate, 20 mM NaCl, 0.1 mM BCS, pH 7.5. The concentration of Fe(II)-loaded protein was determined using the colorimetric TPTZ assay described above[35] and used as the concentration of active enzyme. A 0.64 mM solution of 5,5'-dithio-bis-[2-nitrobenzoic acid] (DTNB, Ellman's reagent) in phosphate buffer was prepared and stored at 4 °C for up to a week. A 100 mM stock of Cys in phosphate buffer was prepared and adjusted to a pH of 7.5 using NaOH. For each protein sample, 5 reactions of 7 μM protein were run at 37 °C and 300 rpm. The reaction was initiated by addition of substrate Cys to protein to a final concentration of 1, 3, 5, 10, or 20 mM. Reaction aliquots of 2.5 μL were quenched every 2 min for 24 min by addition of the aliquot to a well in a 96-well plate containing 97.5 μL of 0.6 mM DTNB. Absorbance readings at 412 nm for each well were taken on a Tecan SPARK plate reader using 350 μL capacity BRAND clear flat-bottom 96 well plates. A standard curve of 0.1-20 mM Cys was prepared with each protein assay and used to calculate the concentration of Cys remaining at each reaction time point. Values for k_{cat} and K_M were again determined using the concentration of Fe(II)-loaded enzyme and the Michaelis-Menten analysis reported by Johnson.[39]

Sequence Similarity Network and Multiple Sequence Alignment. A sequence similarity network (SSN) of the CDO family was generated using the InterPro-defined Cysteine Dioxygenase Type I family (IPR010300) with the resources provided by the Enzyme Function

Initiative-Enzyme Similarity Tool (EFI-EST).[41] Upon retrieval of the sequences and primary network generation, the dataset was pared down to further refine the information provided and yield a suitable file size. Only sequences with lengths between 125 and 300 amino acids were used. A 65% representative node cutoff was used to accommodate the large number of sequences and an E-value of 10^{-18} was chosen as the minimum stringency cutoff. The SSN was visualized using Cytoscape 3.9.1.[42] Taxonomy labels for clusters were generated using the Taxonomy Name/ID Status Report function from the NIH Taxonomy Database website to identify the complete taxonomy lineage from the taxonomy ID for each sequence in the SSN cluster.[43] An in-house MATLAB R2018a script was then used to sort the taxonomy lineage identifiers in frequency order.

A multiple sequence alignment (MSA) of all sequences in the InterPro Cysteine Dioxygenase Type I family was generated using the Super5 algorithm in the Multiple Sequence Comparison by Log-Expectation (MUSCLE) v5 program.[44] The alignment was visualized in SnapGene Viewer 4.2.11 (www.snapgene.com) and the positions that aligned with C93 and Y157 in the *Mm*CDO sequence were identified. An in-house MATLAB R2018a script was used to identify all CDO sequences that contained analogous C and Y residues at these positions, and these sequences were labeled in the SSN as being putatively capable of crosslink formation. SnapGene was used to generate consensus sequences of subsets of CDOs using their aligned FASTA format sequences generated by MUSCLE. Residues were numbered by the *Mm*CDO residue with which they aligned. When comparing several consensus sequences, conserved sequence gaps found in all consensus sequences were removed for clarity.

2.3 Results

Production of Crosslinked G82C *Bs*CDO Variant. The goal of this study was to determine if a single G-to-C substitution in *Bs*CDO could lead to the formation of a C–Y crosslink analogous to that found in eukaryotic CDOs. For eukaryotic CDOs, a characteristic double band pattern is

observed on an SDS-PAGE gel, as the crosslinked and non-crosslinked forms travel as proteins with slightly different apparent molecular weights.[16] We found that while WT *Bs*CDO produced a single SDS-PAGE band, as expected, the G82C variant traveled as two bands (**Figure 2**), like eukaryotic CDOs. Crosslink formation in eukaryotic CDOs is known to be a side reaction of Cys oxidation and thus dependent on the presence of enzyme-bound iron and substrates Cys and O_2.[16] By varying the iron content of our growth media during the production of G82C *Bs*CDO, we were able to change the relative intensities of the two bands on the SDS-PAGE gel (**Figure 2**), further suggesting that the G82C *Bs*CDO variant is capable of crosslink formation. ImageJ software was used to quantify the intensities of the two SDS-PAGE gel bands in each sample, and the fraction of crosslinked protein was calculated as the ratio of the intensities of the bottom and top bands.[45] As expected, by increasing the concentration of iron in the growth medium, the fraction of crosslinked G82C *Bs*CDO variant increased (Table 1). While the difference in percentage of crosslinked protein determined for the three low-iron G82C *Bs*CDO preparations (Preps 2, 3, and 4) likely falls within the error margin associated with SDS-PAGE densitometry, a significantly higher fraction of crosslinked variant was obtained from the high-iron preparation (Prep 1).

Interestingly, because the G82C *Bs*CDO samples from the low iron Preps 3 and 4 were run at a lower concentration on the SDS-PAGE gel, a splitting of the lower band associated with crosslinked protein into a pair of two closely spaced bands could be discerned (**Figure 2**, lanes D and E). Such a splitting has not been observed for eukaryotic CDOs and may reflect two slightly different confirmations of the denatured, crosslinked forms of G82C *Bs*CDO.

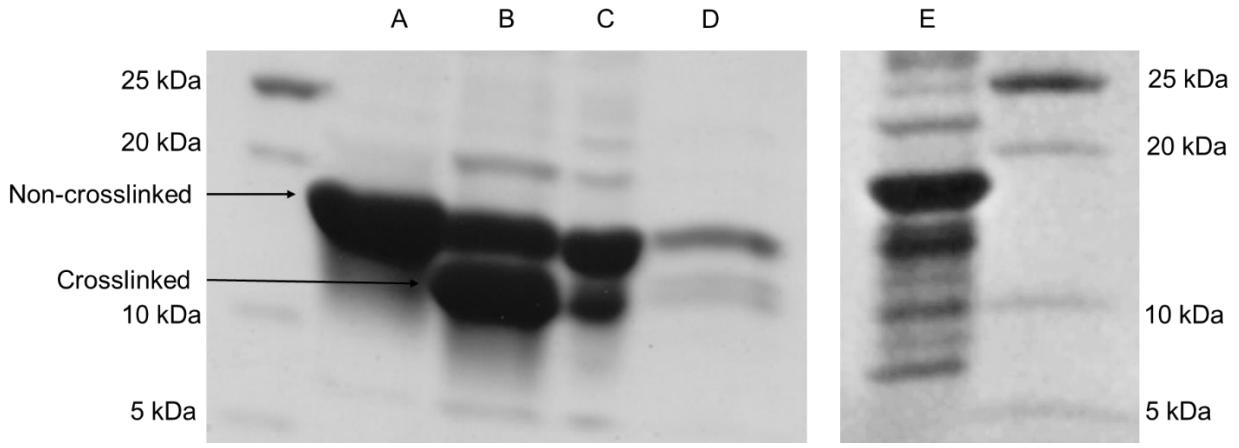


Figure 2.2 SDS-PAGE gels of (A) WT BsCDO, (B) G82C BsCDO Prep 1, (C) G82C BsCDO Prep 2, (D) G82C BsCDO Prep 3, and (E) G82C BsCDO Prep 4. The different conditions used for Preps 1-4 and percentage of crosslinked protein in each sample are provided in Table 1.

Table 2.1 Growth conditions and percentage of crosslinked WT BsCDO and different preparations of the G82C BsCDO variant as determined via integration of the SDS-PAGE gel band intensities in **Figure 2**

Lane	Sample	Growth Conditions	Crosslinked Fraction
A	WT *Bs*CDO	Modified TB medium, 500 μM FAS	0%
B	G82C *Bs*CDO High Iron Prep 1	Modified TB medium, 500 μM FAS	62%
C	G82C *Bs*CDO Low Iron Prep 2	Modified TB medium, 100 μM 1,10-phenanthroline	34%
D	G82C *Bs*CDO Low Iron Prep 3	LB medium, 100 μM, 1,10-phenanthroline, IPTG induction at $OD_{600} \approx 4$	42%
E	G82C *Bs*CDO Low Iron Prep 4	LB medium, 100 μM 1,10-phenanthroline, IPTG induction at $OD_{600} \approx 0.8$	38%

Mass Spectrometry. To confirm that the double band pattern observed for the different preparations of G82C *Bs*CDO on the SDS-PAGE gel in **Figure 2** was indeed due to the formation of a C82–Y141 crosslink in a subset of protein monomers, MS experiments were performed. The top and bottom bands of a high iron G82C *Bs*CDO sample (Prep 1) were excised from an SDS-PAGE gel, digested using a trypsin/AspN combination, and subjected to a MS analysis. Complete sequence coverage was achieved, which indicated with >95% certainty that the protein in both bands was G82C *Bs*CDO. The total ion chromatograms of the protein extracted from the top and bottom gel bands were identical except for the relative intensities of the two peaks associated with the peptides containing the non-crosslinked C82 and Y141 residues (**Figure 3**). The drastically reduced intensities of these peaks in the sample from the

lower gel band indicate that in this protein fraction the corresponding peptides are crosslinked. The newly observed middle band on the SDS-PAGE gel for G82C *Bs*CDO was also excised and analyzed using MS. For this protein fraction, the relative intensities of the two peptide peaks were found to be comparable to, albeit slightly more intense than, those displayed by the protein extracted from the bottom gel band (**Figure A.2.2**), in support of our hypothesis that the two lower gel bands reflect two slightly different confirmations of the denatured crosslinked enzyme.

Because no new species due to dipeptide formation could be observed by scanning raw data or by using crosslinking software, we performed an analogous MS experiment for WT *Mm*CDO (Prep 2, ~65% crosslinked). The top and bottom bands were excised from an SDS-PAGE gel and processed as described above for G82C *Bs*CDO. Again, while we observed different relative intensities of the two peaks associated with the peptides containing the non-crosslinked C and Y residues in the ion chromatograms of the protein extracted from the top and bottom gel bands (**Figure A.2.1**), we were unable to identify a new peptide fragment containing the C–Y crosslink. This result is identical to what Dominy et al. observed in their original MS analysis of the top and bottom bands of WT *Mm*CDO.[16] Thus, our MS experiments provide compelling evidence that the G82C *Bs*CDO variant is capable of C–Y crosslink formation.

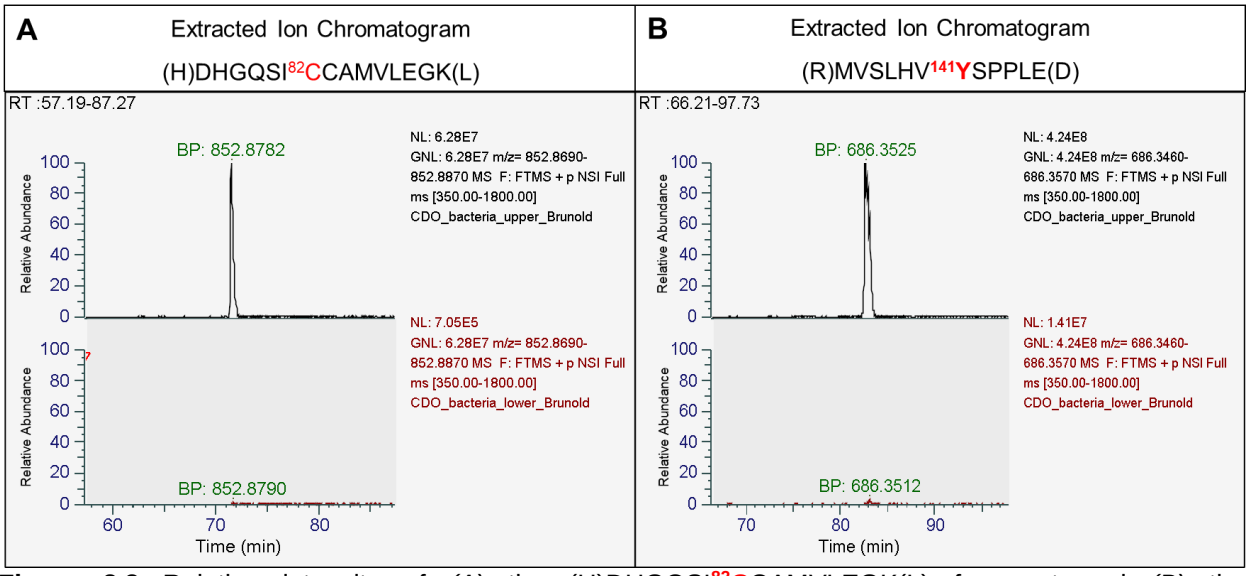


Figure 2.3 Relative intensity of *(A)* the (H)DHGQSI[82]**C**CAMVLEGK(L) fragment and *(B)* the (R)MVSLHV[141]**Y**SPPLE(D) fragment in the mass spectra of the protein extracted from the upper and lower SDS-PAGE gel bands of the high iron G82C *Bs*CDO Prep 1. Analogous results were obtained with WT *Mm*CDO (**Figure A.2.1**).

EPR spectra of (Cys/CN$^-$)-bound Fe(III)CDO complexes. Previous studies revealed that the S = ½ EPR spectrum displayed by (Cys/CN$^-$)-bound Fe(III)*Mm*CDO, which mimics the putative Fe(III)-superoxo intermediate, is sensitive to the absence or presence of the C–Y crosslink.[19] Crosslinked (Cys/CN$^-$)-bound Fe(III)*Mm*CDO was found to exhibit a larger spread of the EPR g-values than the fully non-crosslinked C93A variant and the non-crosslinked fraction of as-isolated *Mm*CDO.[19] We were able to replicate these results using as-isolated WT *Mm*CDO (Prep 1), which in our hands was mainly crosslinked, and its C93G variant, which is unable to form this crosslink (**Figure 2.4A and B**).

As expected, the EPR spectrum of (Cys/CN$^-$)-bound WT Fe(III)*Bs*CDO displays a single, rhombic EPR signal (**Figure 2.4C**) with g values that are similar to those exhibited by the (Cys/CN$^-$) adduct of the C93G *Mm*CDO variant (**Figure 2.4B**), which is also unable to form a C–Y crosslink. Alternatively, the EPR spectrum of (Cys/CN$^-$)-bound G82C *Bs*CDO (**Figure 2.4D**) is nearly identical to that of WT *Mm*CDO (**Figure 2.4A**), showing contributions from both crosslinked (major fraction) and non-crosslinked (minor fraction) protein. The g values obtained from fits of the EPR spectra in **Figure 2.4** are listed in Table 2.2. These fits also allowed us to

determine the relative contributions from the two different S = ½ signals for (Cys/CN^-)-bound WT *Mm*CDO and G82C *Bs*CDO, and thus the ratio of crosslinked to non-crosslinked Fe(III)-bound protein. It is important to note that this ratio is not expected to match the ratio of crosslinked to non-crosslinked protein determined using SDS-PAGE densitometry, because the former only depends on the fraction of CDO containing (Cys/CN^-)-bound Fe(III) centers, while the latter is determined by the complete CDO population regardless of Fe loading or oxidation state. To test the hypothesis that iron was required for crosslink formation, we additionally collected an EPR spectrum of G82C *Bs*CDO from cells grown in a low-iron environment (Prep 4). As expected, a higher relative contribution from the (Cys/CN^-) adduct of the non-crosslinked protein to the EPR spectrum of this sample was observed (**Figure 2.5**). The same trend was seen when comparing EPR spectra of (Cys/CN^-)-bound WT *Mm*CDO samples obtained using different growth and purification conditions (**Figure A.2.3**). Together with our SDS-PAGE gel and MS results, these EPR data provide compelling evidence that a C–Y crosslink forms in G82C *Bs*CDO.

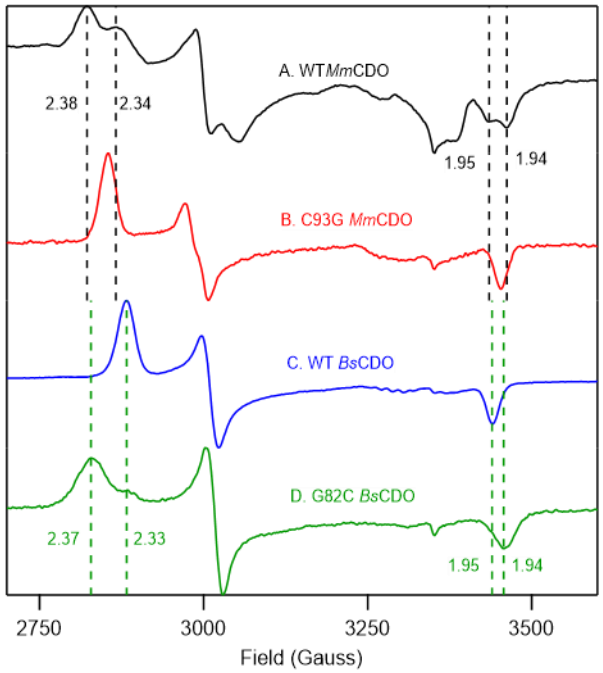


Figure 2.4 EPR spectra of the (Cys/CN^-)-Fe(III) adducts of (A) WT MmCDO Prep 1, (B) C93G MmCDO, (C) WT BsCDO, and (D) high iron G82C BsCDO Prep 1. Features between ~3250 and 3400 Gauss are due to a contaminant in the instrument cavity. The dashed black vertical lines indicate the g_1 and g_3 values for the crosslinked and non-crosslinked fractions of WT MmCDO, and the dashed green vertical lines indicate the g_1 and g_3 values for the crosslinked and non-crosslinked fractions of G82C BsCDO. All EPR g values for the different species obtained from spectral fits are provided in **Table 2.1**.

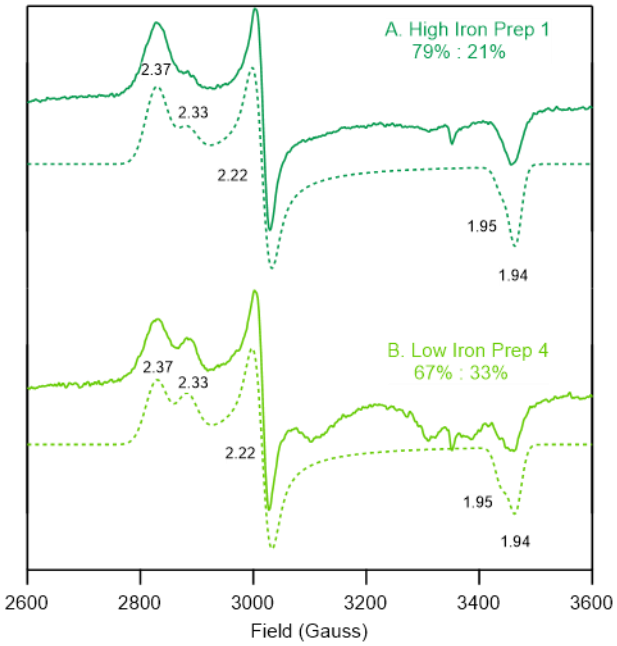


Figure 2.5 EPR spectra of the (Cys/CN^-)-Fe(III) adducts of (A) high iron G82C BsCDO Prep 1 and (B) low iron G82C BsCDO Prep 4. Spectra are shown as solid lines and fits as dashed lines. The ratios of crosslinked to non-crosslinked Fe(III)-bound active sites as determined via fitting of these spectra are included. Features between ~3250 and 3400 Gauss are due to a contaminant in the instrument cavity.

Table 2.2 g values and relative contributions from the crosslinked and non-crosslinked protein fractions obtained from fits of the EPR spectra shown in **Figures 2.4, 2.5, and A.2.3–A.2.5**

Sample	Species	% Contribution	g_1	g_2	g_3
A. WT *Mm*CDO Prep 1	Crosslinked	65	2.378	2.235	1.935
	Non-crosslinked	35	2.335	2.205	1.951
B. WT *Mm*CDO Prep 2	Crosslinked	88	2.378	2.235	1.935
	Non-crosslinked	12	2.335	2.205	1.951
C. C93G *Mm*CDO	Non-crosslinked	100	2.349	2.243	1.941
D. WT *Bs*CDO	Non-crosslinked	100	2.327	2.227	1.948
E. G82C *Bs*CDO High Iron Prep 1	Crosslinked	79	2.370	2.223	1.935
	Non-crosslinked	21	2.326	2.223	1.949
F. G82C *Bs*CDO Low Iron Prep 4	Crosslinked	67	2.370	2.223	1.939
	Non-crosslinked	33	2.326	2.223	1.949

MCD Spectra of Cys-bound WT and G82C Fe(II)*Bs*CDO. In previous studies of as-isolated Cys-bound WT *Mm*CDO, features at 32,000 and 15,700 cm^{-1} were shown to arise from $S_{Cys} \rightarrow$ Fe(II) and $S_{Cys} \rightarrow$ Fe(III) charge transfer (CT) transitions, respectively.[18] Interestingly, in MCD spectra of the H155A *Mm*CDO variant, the $S_{Cys} \rightarrow$ Fe(II) CT transitions were found to be markedly blue-shifted from their WT counterparts.[17] These shifts were attributed to the formation of a six-coordinate (H_2O/Cys)-Fe(II) complex in the variant, rather than a five-coordinate Cys-only Fe(II) adduct as observed for the WT enzyme. The ~100-fold decreased catalytic efficiency displayed by the H155A *Mm*CDO variant thus likely stems from the lack of an open coordination site for O_2 of the Cys-bound Fe(II) center. To ensure that the G82C substitution and partial crosslink formation did not cause the binding of a water molecule or any other drastic perturbation to the catalytically relevant Fe(II) form of *Bs*CDO, we collected MCD spectra of Cys-bound WT and G82C Fe(II)*Bs*CDO (**Figure 2.6**). Both spectra are dominated by the expected $S_{Cys} \rightarrow$ Fe(II) CT band at ~33,000 cm^{-1}. The close similarity of these spectra indicates that the G82C substitution and partial formation of the C–Y crosslink in *Bs*CDO causes no major geometric or electronic structural changes of the Fe(II) center. Thus, we can confidently attribute any major activity differences between WT and G82C *Bs*CDO to the absence or (partial) presence of the crosslink.

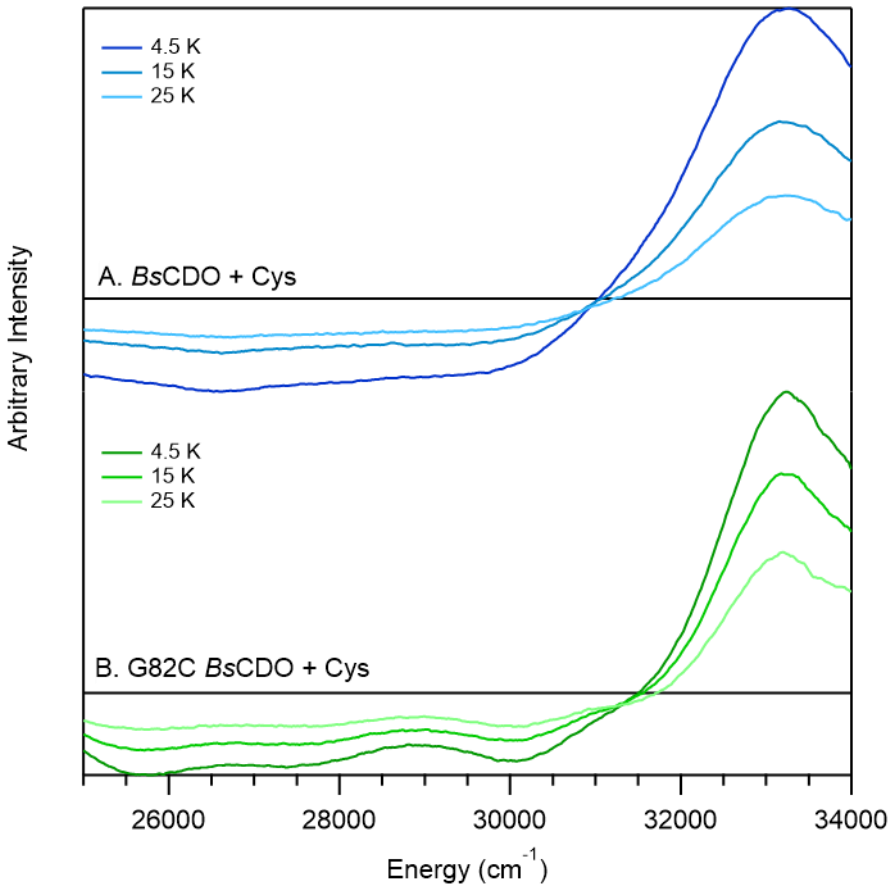


Figure 2.6 Variable temperature MCD spectra at 7 T of Cys-bound (A) WT Fe(II)BsCDO and (B) low iron G82C Fe(II)BsCDO Prep 2.

Kinetic Assays of WT and G82C *Bs*CDO with native substrate Cys. We recently adopted two protocols to determine K_M and k_{cat} values for previously uncharacterized CDO variants. The first, based on work by Stipanuk and coworkers and Jameson and coworkers, involves the use of UPLC-MS to monitor both substrate consumption and product formation.[30,37,38] The second, pioneered by Jameson and coworkers, utilizes Ellman's reagent to monitor substrate depletion spectrophotometrically.[40] To demonstrate that we successfully adopted these assays, we first incubated WT *Mm*CDO with various concentrations of Cys and found that the enzyme displayed Michaelis-Menten kinetics with a K_M of 5.5 mM and k_{cat} of 0.8 s^{-1} (Table 3), in good agreement with results reported by Stipanuk and coworkers (K_M of 4.5 mM and k_{cat} of 0.72 s^{-1}).[37] We then confirmed that WT *Bs*CDO has comparable activity to fully crosslinked eukaryotic CDO, despite

the absence of the C–Y crosslink (Table 3). G82C *Bs*CDO showed reduced catalytic efficiency compared to WT *Bs*CDO, and activity increased as the ratio of crosslinked to non-crosslinked enzyme increased. While in Table 3 we report the percentage of crosslinked enzyme for each sample as determined using SDS-PAGE densitometry, this method accounts for all protein in the preparation, whereas kinetic assays only probe the active, Fe(II)-bound form of the enzyme. Thus, we expect a qualitative, rather than a quantitative correlation between the percentage of crosslinked total protein and enzymatic activity of G82C *Bs*CDO. Michaelis-Menten fits of all kinetic data are provided in **Figure A.2.6**.

Our least crosslinked sample of G82C *Bs*CDO (Prep 2) showed a drastic (~150-fold) reduction in catalytic efficiency relative to WT *Bs*CDO. This result is somewhat unexpected, as one would anticipate a linear relationship between activity and percentage crosslink formation if only the crosslinked fraction were active. This drastic reduction in activity was persistent across all low-iron preparations of G82C *Bs*CDO (Preps 2, 3, and 4), regardless of whether or not the protein was reconstituted with iron before the activity assay (data not shown). Because we also observed very weak EPR signals for samples of the low-iron growths of G82C *Bs*CDO after reconstitution with Fe(II) and incubation with Cys and CN^-, and since previous studies revealed that different CDO variants incorporate different amounts of Zn(II), we hypothesize that G82C *Bs*CDO produced under low Fe conditions or in the presence of a chelator incorporates different metals in place of Fe.[46] Reconstitution with FAS could lead to non-specifically bound Fe that yields seemingly normal Fe-loading values in the TPTZ assay. In this scenario, inactive, mis-metallated protein and non-specifically bound Fe in G82C *Bs*CDO Prep 2 that increase the apparent G82C Fe(II)*Bs*CDO concentration could explain the unexpectedly large difference in k_{cat} between the high iron Prep 1 and low iron Prep 2 of G82C *Bs*CDO.

Table 2.3 Kinetic parameters of different MmCDO and BsCDO species as determined by activity assays

Sample	% Crosslinked In SDS-PAGE Gel	K_M (mM)	k_{cat} (sec^{-1})	k_{cat}/K_M ($mM^{-1}sec^{-1}$)
*Mm*CDO Prep 2	Not measured	5.5 ± 0.3	0.8 ± 0.1	0.15 ± 0.03
*Bs*CDO	0	6.6 ± 0.8	0.4 ± 0.1	0.05 ± 0.02
G82C *Bs*CDO High Iron Prep 1	62	10.7 ± 0.3	0.17 ± 0.02	0.016 ± 0.002
G82C *Bs*CDO Low Iron Prep 2	34	11.5 ± 0.1	0.0041 ± 0.0003	0.00036 ± 0.00003

2.4 Discussion

Since their discovery and initial kinetic and structural characterization seventeen years ago,[21,22] no further investigations of bacterial CDOs have been reported. In the present study, we took advantage of the fact that WT *Bs*CDO is unable to form the C–Y crosslink that is found in all eukaryotic CDOs. By substituting the glycine at position 82 with a cysteine, we were able to successfully create a variant of *Bs*CDO capable of crosslink formation, as confirmed by SDS-PAGE gel electrophoresis, mass spectrometry, and EPR spectroscopy. MCD data indicate that the Cys-bound (catalytically relevant) form of the G82C Fe(II)*Bs*CDO variant has essentially the same active site electronic structure as WT *Bs*CDO. Kinetic analyses revealed that C82 in *Bs*CDO has a similar effect on the rate of Cys oxidation as does the analogous C93 in *Mm*CDO. Introduction of the Cys residue leads to a drop in enzyme activity, but crosslink formation between C82 and Y141 recovers, at least partially, the activity of the WT enzyme. Because the population of crosslinked G82C *Bs*CDO correlates with the amount of iron present during protein production, it is reasonable to assume that C–Y crosslink formation in this variant also occurs as a result of an unproductive reaction with L-cysteine and dioxygen, as in the case of WT *Mm*CDO.[16]

In their initial study, Stipanuk and coworkers discovered only 38 putative bacterial CDOs in sequence databases, and all appeared incapable of crosslink formation.[21] With the exponential increase in available protein sequence data over recent years, we decided to conduct an updated and more comprehensive analysis of putative CDOs across all organisms.

As of July 2022, the InterPro Cysteine Dioxygenase Type I family contained 11,270 proteins, 9,159 of which were from bacteria, 1,823 from eukaryotes, and 187 from archaea.[47] It is clear from these numbers alone that CDOs appear to be much more prevalent in bacteria than originally thought, and that the sulfur metabolic pathways of bacteria may not yet be fully understood. **Figure 2.7** shows a sequence similarity network (SSN) of all proteins between 125 to 300 amino acids in length that belong to the InterPro Cysteine Dioxygenase Type I family (IPR010300).[47] Nodes represent groups of proteins with ≥ 65% sequence identity and edges are drawn between pairs of nodes for which the BLAST E-values are less than a user-defined, upper limit threshold, E.

At a low stringency value ($E=10^{-21}$, **Figure 2.7A**), a few large clusters appear. All eukaryotic CDOs are clustered together, along with a grouping of CDOs from the Fibrobacteres-Chlorobi-Bacteroidetes (FCB) clade of bacteria and a grouping from proteobacteria and planctomycetes bacteria, all of which appear capable of crosslink formation based on the presence of Cys and Tyr residues analogous to C93 and Y157 in *Mm*CDO. A large cluster of CDOs from actinobacteria and a small cluster of CDOs from bacilli appear incapable of crosslink formation because they lack the Cys residue corresponding to C93 in *Mm*CDO and instead typically possess a G at this position. A large cluster of CDOs from proteobacteria is split roughly down the middle between putatively crosslinked and non-crosslinked CDOs. Some putatively crosslinked CDOs from actinobacteria cluster with the putatively crosslinked half of CDOs from proteobacteria. Generally, clustering occurs based on superkingdom or phylum classification, while also depending on putative crosslinking ability. The majority of bacterial CDOs that appear capable of crosslink formation are from proteobacteria. Some CDOs from actinobacteria may also be capable of crosslink formation; however, they share greater sequence identity with the allegedly crosslinked proteobacterial CDOs than with the non-crosslinked actinobacterial CDOs.

At a high stringency value ($E=10^{-26}$, **Figure 2.7B**), smaller clusters emerge based on the organisms' class or order. We observe that the FCB clade CDOs remain closest to eukaryotic CDOs. The major types of proteobacterial CDOs putatively capable of crosslink formation are from beta- and gammaproteobacteria, specifically Burkholderia, Xanthomonadales, and Vibronales. Vibronales and planctomycetes CDOs cluster with archaeal CDOs. These are all gram-negative pathogenic bacteria, suggesting the possibility that crosslinked CDO was obtained through horizontal gene transfer from a eukaryotic host. In their discovery of non-crosslinked bacterial CDOs, Dominey et al. suggested that CDO may aid spore formulation in bacteria by decreasing Cys levels and thus promoting disulfide bond formation.[21] However, most of the putatively crosslinked bacterial CDOs in our SSN are from organisms that do not form spores. Streptomyces is the largest genus of actinobacteria that appears capable of CDO crosslink formation, and these CDOs have greater sequence identity with those from Burkholderia than with other, non-crosslinked actinobacterial CDOs. Interestingly, *Bs*CDO, which was used in this study, is not as similar to eukaryotic CDOs as some other non-crosslinked CDOs are.

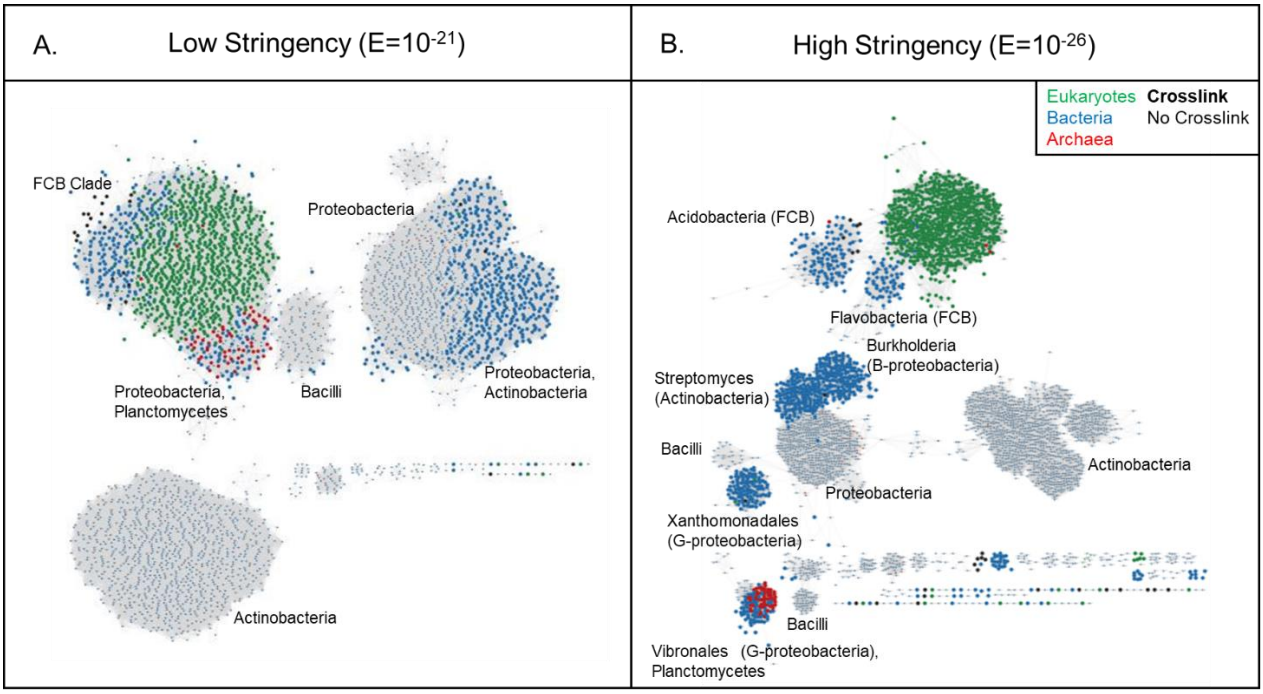


Figure 2.7 SSN of the InterPro CDOI family at *(A)* low stringency (E=10^{-21}) and *(B)* high stringency (E=10^{-26}). Eukaryotic CDOs are shown in green, bacterial CDOs in blue, and archaeal CDOs in red. CDOs that are putatively capable of crosslink formation based on the presence of residues analogous to both C93 and Y157 (*Mm*CDO numbering) are shown as larger dots. Each major cluster is labeled with the taxonomic category to which the majority of CDOs in that cluster belong.

To better understand the conservation of specific structurally and catalytically relevant residues across different clusters of CDOs, a multiple sequence alignment (MSA) was performed. **Figure 2.8** shows the aligned consensus sequences of eukaryotic, putatively crosslinked bacterial, and non-crosslinked bacterial CDOs. Residues are numbered according to the *Mm*CDO scheme and important residues are highlighted. The three Fe-binding His residues as well as the S-H-Y "catalytic triad" are conserved across all CDOs. However, several key differences can be noted. At position 60, non-crosslinked bacterial CDOs contain a Q, which is typical of MDOs, rather than the R typical of CDOs.[22] Interestingly, however, *Bs*CDO maintains the R residue here. Instead of a Y at position 58, which is thought to aid in substrate binding, non-crosslinked bacterial CDOs contain a conserved R.[15] Meanwhile, putatively crosslinked bacterial CDOs lack the *cis*-proline bond thought to position residue 164 and differentiate CDOs from MDOs.[23]

Thus, while the 3-His triad and S-H-Y triad are absolutely conserved, residues known to be involved in substrate binding in thiol dioxygenase enzymes are not uniformly conserved across (putatively) crosslinked and non-crosslinked CDOs. This may indicate differences in substrate specificity across clusters of CDOs. Considering the differences between the effect on enzyme activity of installing a crosslink into *Bs*CDO performed in this study and into *Pa*MDO as accomplished previously by Jameson and coworkers, this may also imply that these substrate binding residues and the absence or presence of crosslink cooperatively affect enzyme turnover rate.[32]

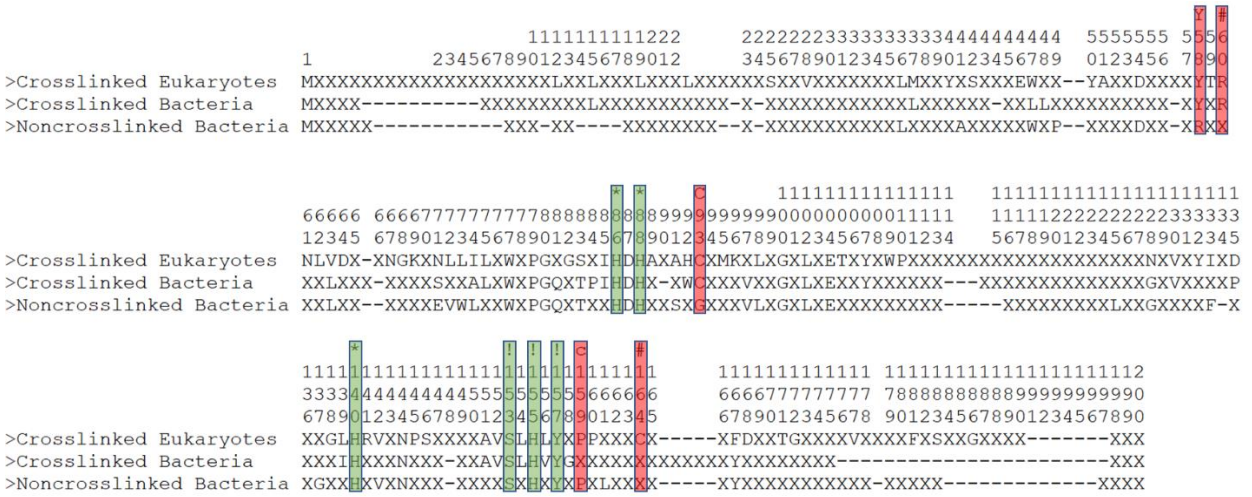


Figure 2.8 MSA comparing the consensus sequences of crosslinked eukaryotic, putatively crosslinked bacterial, and non-crosslinked bacterial CDOs. One letter amino acid codes are used, where "X" represents a position where no particular amino acid is conserved and "-" represents a gap where no amino acid exists in that consensus sequence. Residues are labeled vertically above each position with MmCDO numbering. Residues conserved across all three groups of CDOs are highlighted in green. Differences in conserved amino acid residues are highlighted in red. Legend: *, Fe(II)-binding His residues; !, S-H-Y "catalytic triad"; Y, additional substrate-binding Tyr in CDOs; #, active site R and Q residues distinguishing CDOs from MDOs and MSDOs; c, cis-Pro peptide bonds in CDOs.

2.5 Conclusions

Collectively, our data show that it is possible to build a C–Y crosslink into *Bs*CDO, further confirming previous observations that the spatial proximity of these two residues contributes substantially to their ability to crosslink.[32,48] We have demonstrated that as in eukaryotic CDO, the non-crosslinked isoform displays decreased catalytic efficiency.[16] Our bioinformatic analysis

of the CDO family revealed the existence of many putatively crosslinked bacterial CDOs, the majority of which are from gram-negative pathogenic bacteria. Additionally, we observed that residues involved in substrate binding are not conserved across eukaryotic, putatively crosslinked bacterial, and non-crosslinked bacterial CDOs, indicating potential relationships between substrate specificity, C–Y crosslinking, and enzyme activity.

Accession Codes

*Bs*CDO, UniProt O32085; *Mm*CDO, UniProt P60334

Funding

The authors are grateful for financial support from the National Institute of General Medical Sciences of the National Institutes of Health (Grant GM117120 to T.C.B.). This research was also supported by funds from the University of Wisconsin-Madison FY21 Kellett Mid-Career Award (Grant AAH8228 to T.C.B.).

Mass spectrometry data collection and analysis were performed by the Mass Spectrometry Core at the University of Wisconsin-Madison Biotechnology Center, which is supported by user fees and by the university.

Acknowledgement

The authors thank Dr. Andrew Buller for the use of his UPLC-MS instrument, Dr. Rebeca Fernandez for her development of the UPLC-MS kinetic assay, and Joshua Miller for providing EPR spectra of WT and C93G (Cys/CN^-)-Fe(III)*Mm*CDO.

Notes

The authors declare no competing financial interest.

2.6 References

(1) Lombardini, J. B.; Singer, T. P.; Boyer, P. D. Cysteine Oxygenase II. Studies on the Mechanism of the Reaction with 18Oxygen. *J. Biol. Chem.* **1969**, *244* (5), 1172–1175. https://doi.org/https://doi.org/10.1016/S0021-9258(18)91825-9.

(2) Stipanuk, M. H.; Dominy, J. E.; Lee, J.-I.; Coloso, R. M. Mammalian Cysteine Metabolism: New Insights into Regulation of Cysteine Metabolism. *J. Nutr.* **2006**, *136* (6), 1652S-1659S. https://doi.org/https://doi.org/10.1093/jn/136.6.1652S.

(3) Lehmann, A.; Hagberg, H.; Orwar, O.; Sandberg, M. Cysteine Sulphinate and Cysteate: Mediators of Cysteine Toxicity in the Neonatal Rat Brain? *Eur. J. Neurosci.* **1993**, *5* (10), 1398–1412. https://doi.org/10.1111/j.1460-9568.1993.tb00926.x.

(4) Montine, T. J.; Picklo, M. J.; Amarnath, V.; Whetsell, W. O.; Graham, D. G. Neurotoxicity of Endogenous Cysteinylcatechols. *Exp. Neurol.* **1997**, *148* (1), 26–33. https://doi.org/10.1006/exnr.1997.6662.

(5) Fjellstedt, E.; Harnevik, L.; Jeppsson, J.-O.; Tiselius, H.-G.; Söderkvist, P.; Denneberg, T. Urinary Excretion of Total Cystine and the Dibasic Amino Acids Arginine, Lysine and Ornithine in Relation to Genetic Findings in Patients with Cystinuria Treated with Sulfhydryl Compounds. *Urol. Res.* **2003**, *31* (6), 417–425. https://doi.org/10.1007/s00240-003-0366-6.

(6) Thomas Heafield, M.; Fearn, S.; Steventon, G. B.; Waring, R. H.; Williams, A. C.; Sturman, S. G. Plasma Cysteine and Sulphate Levels in Patients with Motor Neurone, Parkinson's and Alzheimer's Disease. *Neurosci. Lett.* **1990**, *110* (1–2), 216–220. https://doi.org/https://doi.org/10.1016/0304-3940(90)90814-P.

(7) Emery, P.; Salmon, M.; Bradley, H.; Wordsworth, P.; Tunn, E.; Bacon, P. A.; Waring, R. Genetically Determined Factors as Predictors of Radiological Change in Patients with Early Symmetrical Arthritis. *Br. Med. J.* **1992**, *305* (6866), 1387–1389.

https://doi.org/10.1136/bmj.305.6866.1387.

(8) Davies, M. H.; Ngong, J. M.; Pean, A.; Vickers, C. R.; Waring, R. H.; Elias, E. Sulphoxidation and Sulphation Capacity in Patients with Primary Biliary Cirrhosis. *J. Hepatol.* **1995**, *22* (5), 551–560. https://doi.org/10.1016/0168-8278(95)80450-1.

(9) Bradley, H.; Gough, A.; Sokhi, R. S.; Hassell, A.; Waring, R.; Emery, P. Sulfate Metabolism Is Abnormal in Patients with Rheumatoid Arthritis. Confirmation by in Vivo Biochemical Findings. *J. Rheumatol.* **1994**, *21* (7), 1192–1196. https://doi.org/10.1016/s0950-3579(05)80178-5.

(10) Finkelstein, J. D.; Kyle, W. E.; Harris, B. J.; Martin, J. J. Methionine Metabolism in Mammals: Concentration of Metabolites in Rat Tissues. *J. Nutr.* **1982**, *112* (5), 1011–1018. https://doi.org/10.1093/jn/112.5.1011.

(11) Stipanuk, M. H.; Ueki, I.; Stipanuk, M. H.; Ueki, I. Dealing with Methionine/Homocysteine Sulfur: Cysteine Metabolism to Taurine and Inorganic Sulfur. *J. Inherit. Metab. Dis.* **2011**, *34* (1), 17–32. https://doi.org/10.1007/s10545-009-9006-9.

(12) Lee, J. I.; Londono, M.; Hirschberger, L. L.; Stipanuk, M. H. Regulation of Cysteine Dioxygenase and γ-Glutamylcysteine Synthetase Is Associated with Hepatic Cysteine Level. *J. Nutr. Biochem.* **2004**, *15* (2), 112–122. https://doi.org/10.1016/j.jnutbio.2003.10.005.

(13) Dominy, J. E.; Hirschberger, L. L.; Coloso, R. M.; Stipanuk, M. H. In Vivo Regulation of Cysteine Dioxygenase via the Ubiquitin-26S Proteasome System. In *Advances in Experimental Medicine and Biology*; Springer New York, 2006; Vol. 583, pp 37–47. https://doi.org/10.1007/978-0-387-33504-9_4.

(14) Mccoy, J. G.; Bailey, L. J.; Bitto, E.; Bingman, C. A.; Aceti, D. J.; Fox, B. G.; Phillips, G. N. Structure and Mechanism of Mouse Cysteine Dioxygenase. *PNAS* **2006**, *103* (9), 3084–3089. https://doi.org/https://doi.org/10.1073/pnas.0509262103.

(15) Driggers, C. M.; Cooley, R. B.; Sankaran, B.; Hirschberger, L. L.; Stipanuk, M. H.;

Karplus, P. A. Cysteine Dioxygenase Structures from PH 4 to 9: Consistent Cys-Persulfenate Formation at Intermediate PH and a Cys-Bound Enzyme at Higher PH. *J. Mol. Biol.* **2013**, *425* (17), 3121–3136. https://doi.org/10.1016/j.jmb.2013.05.028.

(16) Dominy, J. E.; Hwang, J.; Guo, S.; Hirschberger, L. L.; Zhang, S.; Stipanuk, M. H. Synthesis of Amino Acid Cofactor in Cysteine Dioxygenase Is Regulated by Substrate and Represents a Novel Post-Translational Regulation of Activity. *J. Biol. Chem.* **2008**, *283* (18), 12188–12201. https://doi.org/10.1074/jbc.M800044200.

(17) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of the H155A Variant of Cysteine Dioxygenase: Geometric and Electronic Consequences of a Third-Sphere Amino Acid Substitution. *Biochemistry* **2015**, *54* (18), 47. https://doi.org/10.1021/acs.biochem.5b00171.

(18) Gardner, J. D.; Pierce, B. S.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Characterization of Substrate-Bound Mouse Cysteine Dioxygenase: Nature of the Ferrous and Ferric Cysteine Adducts and Mechanistic Implications. *Biochemistry* **2010**, *49* (29), 6033–6041. https://doi.org/10.1021/bi100189h.

(19) Li, W.; Blaesi, E. J.; Pecore, M. D.; Crowell, J. K.; Pierce, B. S. Second-Sphere Interactions between the C93−Y157 Cross-Link and the Substrate-Bound Fe Site Influence the O2 Coupling Efficiency in Mouse Cysteine Dioxygenase. *Biochemistry* **2013**, *52* (51), 9104–9119. https://doi.org/10.1021/bi4010232.

(20) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of Iron(III) Cysteine Dioxygenase: Implications for the Nature of the Putative Superoxo-Fe(III) Intermediate. *Biochemistry* **2014**, *53* (36), 5759–5770. https://doi.org/10.1021/bi500767x.

(21) Dominy, J. E.; Simmons, C. R.; Karplus, P. A.; Gehring, A. M.; Stipanuk, M. H. Identification and Characterization of Bacterial Cysteine Dioxygenases: A New Route of Cysteine Degradation for Eubacteria. *J. Bacteriol.* **2006**, *188* (15), 5561–5569.

https://doi.org/10.1128/JB.00291-06.

(22) Driggers, C. M.; Hartman, S. J.; Karplus, P. A. Structures of Arg- and Gln-Type Bacterial Cysteine Dioxygenase Homologs. *Protein Sci.* **2015**, *24* (1), 154–161. https://doi.org/10.1002/pro.2587.

(23) Aloi, S.; Davies, C. G.; Karplus, P. A.; Wilbanks, S. M.; Jameson, G. N. L. Substrate Specificity in Thiol Dioxygenases. *Biochemistry* **2019**, *58* (19), 2398--2407. https://doi.org/10.1021/acs.biochem.9b00079.

(24) Stipanuk, M. H.; Simmons, C. R.; Karplus, P. A.; Dominy, J. E. Thiol Dioxygenases: Unique Families of Cupin Proteins Iron-Dependent Enzymes. *Amino Acids* **2011**, *41* (1), 91–102. https://doi.org/10.1007/s00726-010-0518-2.

(25) Weits, D. A.; Giuntoli, B.; Kosmacz, M.; Parlanti, S.; Hubberten, H. M.; Riegler, H.; Hoefgen, R.; Perata, P.; Van Dongen, J. T.; Licausi, F. Plant Cysteine Oxidases Control the Oxygen-Dependent Branch of the N-End-Rule Pathway. *Nat. Commun.* **2014**, *5* (1), 1–10. https://doi.org/10.1038/ncomms4425.

(26) White, M. D.; Kamps, J. J. A. G.; East, S.; Taylor Kearney, L. J.; Flashman, E. The Plant Cysteine Oxidases from Arabidopsis Thaliana Are Kinetically Tailored to Act as Oxygen Sensors. *J. Biol. Chem.* **2018**, *293* (30), 11786–11795. https://doi.org/10.1074/jbc.RA118.003496.

(27) Li, J.; Griffith, W. P.; Davis, I.; Shin, I.; Wang, J.; Li, F.; Wang, Y.; Wherritt, D. J.; Liu, A. Cleavage of a Carbon–Fluorine Bond by an Engineered Cysteine Dioxygenase. *Nat. Chem. Biol.* **2018**, *14* (9), 853–860. https://doi.org/10.1038/s41589-018-0085-5.

(28) Ito, N.; Phillips, S. E. V.; Stevens, C.; Ogel, Z. B.; McPherson, M. J.; Keen, J. N.; Yadav, K. D. S.; Knowles, P. F. Novel Thioether Bond Revealed by a 1.7 Å Crystal Structure of Galactose Oxidase. *Nature* **1991**, *350* (6313), 87–90. https://doi.org/10.1038/350087a0.

(29) Schnell, R.; Sandalova, T.; Hellman, U.; Lindqvist, Y.; Schneider, G. Siroheme- and [Fe4-S4]-Dependent NirA from Mycobacterium Tuberculosis Is a Sulfite Reductase with a

Covalent Cys-Tyr Bond in the Active Site. *J. Biol. Chem.* **2005**, *280* (29), 27319–27328. https://doi.org/10.1074/jbc.M502560200.

(30) Davies, C. G.; Fellner, M.; Tchesnokov, E. P.; Wilbanks, S. M.; Jameson, G. N. L. The Cys-Tyr Cross-Link of Cysteine Dioxygenase Changes the Optimal PH of the Reaction without a Structural Change. *Biochemistry* **2014**, *53* (50), 7961--7968. https://doi.org/10.1021/bi501277a.

(31) Stipanuk, M. H.; Ueki, A. I.; Dominy, A. J. E.; Simmons, A. C. R.; Hirschberger, A. L. L. Cysteine Dioxygenase: A Robust System for Regulation of Cellular Cysteine Levels. *Amino Acids* **2009**, *37* (1), 55–63. https://doi.org/10.1007/s00726-008-0202-y.

(32) Fellner, M.; Aloi, S.; Tchesnokov, E. P.; Wilbanks, S. M.; Jameson, G. N. L. Substrate and PH-Dependent Kinetic Profile of 3-Mercaptopropionate Dioxygenase from Pseudomonas Aeruginosa. *Biochemistry* **2016**, *55* (9), 1362–1371. https://doi.org/10.1021/acs.biochem.5b01203.

(33) Njeri, C. W.; Ellis, H. R. Shifting Redox States of the Iron Center Partitions CDO between Crosslink Formation or Cysteine Oxidation. *Arch. Biochem. Biophys.* **2014**, *558*, 61–69. https://doi.org/10.1016/J.ABB.2014.06.001.

(34) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658. https://doi.org/https://doi.org/10.1021/ac0341261.

(35) Fischer, D. S.; Pricet, D. C. A Simple Serum Iron Method Using the New Sensitive Chromogen Tripyridyl-s-Triazine. *Clin. Chem.* **1964**, *10* (1), 21–31. https://doi.org/https://doi.org/10.1093/clinchem/10.1.21.

(36) Nilges, M. J. Electron Paramagnetic Resonance Studies of Low Symmetry Nickel(I) and Molybdenum(V) Complexes, University of Illinois at Urbana-Champaign, 1979.

(37) Simmons, C. R.; Hirschberger, L. L.; Machi, M. S.; Stipanuk, M. H. Expression, Purification, and Kinetic Characterization of Recombinant Rat Cysteine Dioxygenase, a

Non-Heme Metalloenzyme Necessary for Regulation of Cellular Cysteine Levels. *Protein Expr. Purif.* **2006**, *47* (1), 74–81. https://doi.org/10.1016/J.PEP.2005.10.025.

(38) Siakkou, E.; Wilbanks, S. M.; Jameson, G. N. L. Simplified Cysteine Dioxygenase Activity Assay Allows Simultaneous Quantitation of Both Substrate and Product. *Anal. Biochem.* **2010**, *405* (1), 127–131. https://doi.org/10.1016/j.ab.2010.06.013.

(39) Johnson, K. A. New Standards for Collecting and Fitting Steady State Kinetic Data. *Beilstein J. Org. Chem. 152* **2019**, *15* (1), 16–29. https://doi.org/10.3762/BJOC.15.2.

(40) Fellner, M.; Doughty, L. M.; Jameson, G. N. L.; Wilbanks, S. M. A Chromogenic Assay of Substrate Depletion by Thiol Dioxygenases. *Anal. Biochem.* **2014**, *459*, 56–60. https://doi.org/10.1016/j.ab.2014.05.008.

(41) Zallot, R.; Oberg, N.; Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, *58* (41), 4169–4182. https://doi.org/https://doi.org/10.1021/acs.biochem.9b00735.

(42) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(43) Federhen, S. The NCBI Taxonomy Database. *Nucleic Acids Res.* **2012**, *40* (D1), D136–D143. https://doi.org/10.1093/NAR/GKR1178.

(44) Edgar, R. C. High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny. *bioRxiv* **2022**. https://doi.org/10.1101/2021.06.20.449169.

(45) Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W. NIH Image to ImageJ: 25 Years of Image Analysis. *Nat. Methods 2012 97* **2012**, *9* (7), 671–675. https://doi.org/10.1038/nmeth.2089.

(46) Ye, S.; Wu, A.; Wei, L.; Tang, D.; Sun, P.; Bartlam, M.; Rao, Z. An Insight into the Mechanism of Human Cysteine Dioxygenase: Key Roles of the Thioether-Bonded Tyrosine-Cysteine Cofactor. *J. Biol. Chem.* **2007**, *282* (5), 3391–3402. https://doi.org/10.1074/jbc.M609337200.

(47) Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Azaro Pinto, B. L. ´; Salazar, G. A.; Bileschi, M. L.; Bridge, A.; Colwell, L.; Gough, J.; Haft, D. H.; Letuní, I.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Orengo, C. A.; Pandurangan, A. P.; Rivoire, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A. InterPro in 2022. *Nucleic Acids Res.* **2022**. https://doi.org/10.1093/nar/gkac993.

(48) Martinie, R. J.; Godakumbura, P. I.; Porter, E. G.; Divakaran, A.; Burkhart, B. J.; Wertz, J. T.; Benson, D. E. Identifying Proteins That Can Form Tyrosine-Cysteine Crosslinks. *Metallomics* **2012**, *4* (10), 1037–1042. https://doi.org/10.1039/C2MT20093G.

CHAPTER 3

Active Site Flexibility as a Determining Factor in Substrate Specificity in Cysteine Dioxygenase Homologs

The following data for this chapter was collected previously by Dr. Stephanie Dillon: MCD spectra of substrate (analogue)-bound *Mm*CDO, EPR spectra of substrate (analogue)-bound *Mm*CDO with and without azide (excluding 3-MPA).

Chapter 3 : Active Site Flexibility as a Determining Factor in Substrate Specificity in Cysteine Dioxygenase Homologs

3.1 Introduction

Cysteine dioxygenase (CDO) is a mononuclear non-heme iron enzyme that catalyzes the conversion of cysteine (Cys) to cysteine sulfinic acid (CSA).[1] Tight regulation of intracellular Cys levels is incredibly important in eukaryotic organisms, as excess free Cys has been linked to a variety of autoimmune and neurogenerative disorders.[2–5] CDO plays a vital role in this regulation, as it catalyzes the first step in multiple pathways of Cys catabolism. Under conditions of high Cys concentration, CDO is tagged for proteolytic degradation by the 26S proteosome system.[6] Under low Cys concentrations, an active-site C-Y covalent crosslink forms in the CDO active site, which increases catalytic efficiency by approximately 10-fold.[7] CDO is also incredibly specific for its native substrate L-Cys, meaning regulation of CDO activity has a very direct and immediate effect on free Cys concentration.[8] However, the basis of this extreme substrate specificity is not fully understood.

CDO has been shown to catalyze the oxidation of only two other known substrates (D-cysteine and cysteamine) with measurable, albeit significantly diminished, activity.[8,9] Even homocysteine (Hcys) and selenocysteine (Sec), which are closely related to Cys, are competitive inhibitors of the enzyme.[10–13] A crystal structure of *Rattus norvegicus* CDO revealed that binding of the native substrate Cys occurs in a bidentate fashion, involving coordination from the sulfur (S_{Cys}) and nitrogen (N_{Cys}) (Figure 3.1).[14,15] Spectroscopic studies have provided evidence that Sec is capable of binding to the catalytically relevant Fe(II)-containing CDO active site, as well as the mechanistically important Fe(III)-CDO active site, in a similar orientation as the native substrate, indicating that the inability of CDO to oxidize Sec stems from a step later in the catalytic cycle.[11] Indeed, computational data indicate that a Sec-bound CDO intermediate is higher in energy than its Cys-bound counterpart. Alternatively, a crystallographic study of

mammalian CDO demonstrated that HCys binds to Fe in a monodentate fashion, where only the S coordinates directly to the metal ion, providing clues as to why HCys is a competitive inhibitor of the enzyme and not an alternative substrate.[16]

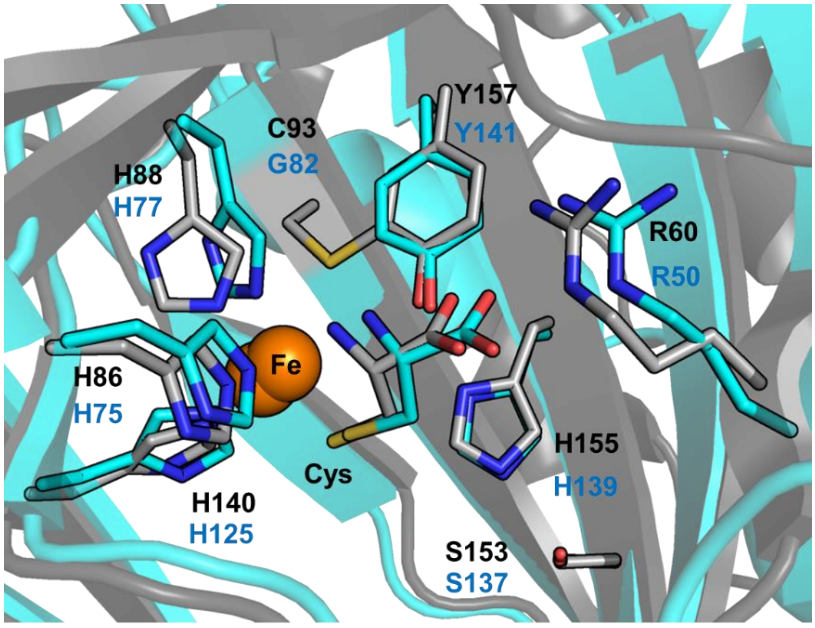


Figure 3.1 Overlay of the active site regions of Cys-bound MmCDO (gray, PDB ID 4JTO) and BsCDO (cyan, PDB ID 4QM9).

Several key active site residues have been identified as playing an important role in substrate positioning and binding. Residue Y157 (*Mm*CDO numbering) forms a hydrogen bond with the carboxylate group of substrate Cys, and removing its hydroxyl group in the Y157F *Mm*CDO variant completely kills activity.[9] A hydrogen-bonding network between S153, H155, and Y157 likely helps position Y157 correctly for interaction with the Cys carboxylate, as the H155A variant has significantly reduced activity.[17] The Cys carboxylate group is also held in place by a salt bridge to the guanidyl group of R60. The enzyme 3-mercaptopropionate dioxygenase (MDO) is missing a residue analogous to R60, and instead contains an R on the opposite side of the active site, so that the salt bridge it forms with the carboxylate moiety of its substrate 3-mercaptopropionic acid (3-MPA) correctly positions the it for coordination to the enzyme's Fe center.[18]

Eukaryotic forms of CDO also contain the crosslink between C93 and Y157 that is responsible for increasing catalytic efficiency, although it is not known if this crosslink plays any role in facilitating substrate binding.[7] Many bacterial CDOs exist which do not possess this crosslink due to a G in the position analogous to C93 in *Mm*CDO, including the CDO for *Bacillus subtilis* (*Bs*CDO) (Figure 3.1).[19] Remarkably, the lack of crosslink does not make *Bs*CDO any less active than fully crosslinked *Mm*CDO, likely because any non-crosslinked C93 in *Mm*CDO is stabilizing a solvent molecule in the O_2 binding site.[9,19]

In previous studies, azide was employed as a possible mimic of superoxide to investigate the nature of a proposed Cys-bound Fe(III)superoxo intermediate.[20] Spectroscopic and computational work indicated that N_3^- does not coordinate to the metal ion of Cys-bound Fe(III)CDO in the large majority of species likely because of the occupation of the putative O_2 binding site opposite of H86 by a solvent molecule.[12] This was corroborated by an X-ray crystallographic study, which did not find evidence for the formation of a Cys- and azide-bound Fe(III)CDO complex. Interestingly, a crystal structure of azide-bound Fe(III)CDO showed that N_3^- coordinates to the Fe(III) center of substrate-free CDO not at the putative superoxo-binding site opposite of H86, but rather in a similar position as S_{Cys} in Cys-bound Fe(III)CDO.[16] EPR spectra of Cys and N_3^--bound Fe(III)CDO show a slightly perturbed S=5/2 signal relative to that of Cys-bound only Fe(III)CDO, indicating that while N_3^- cannot displace native substrate Cys from its coordination site, it still binds somewhere in the active site, most likely near R60.[20]

In this study, we compare WT *Mm*CDO, the naturally non-crosslinked WT *Bs*CDO, as well as a G82C *Bs*CDO variant that we have previously demonstrated to be capable of crosslink formation. Using magnetic circular dichroism (MCD) and electron paramagnetic resonance (EPR) spectroscopies, we probe the ability of these enzymes to bind various analogues of the substrate Cys (Figure 3.2). We also take advantage of azide's ability to potentially displace a substrate analogue from the active site and bind in its place, or else to bind nearby in the active site in a way that is sensitive to EPR spectroscopy. Finally, we evaluated the ability of these

CDOs to turnover the analogues at both high and low concentrations using UPLC-MS and NMR to detect the depletion of substrate and the appearance of product.

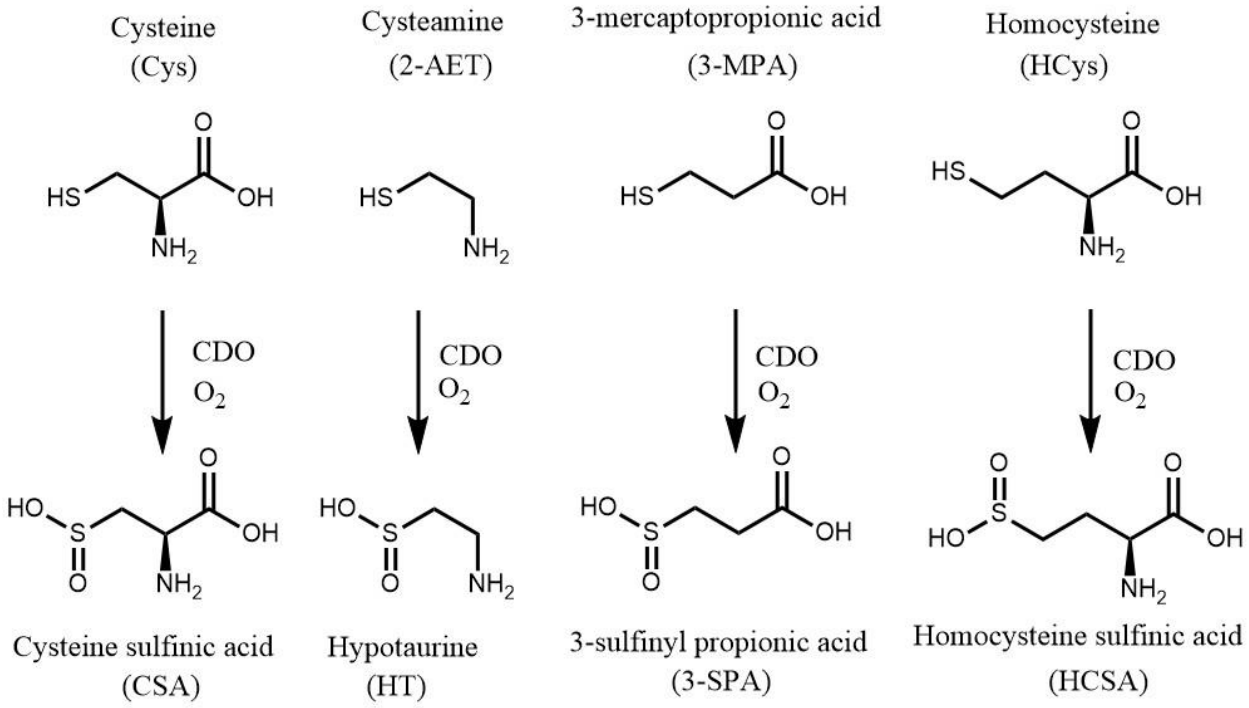


Figure 3.2 Substrate analogues used in this study and their (potential) conversions by CDO.

3.2 Materials and Methods

Recombinant Gene Expression and Protein Purification. Gene expression and protein purification of WT and G82C *Bs*CDO were conducted as described previously. In brief, *E. coli* Rosetta 2(DE3) cells containing the pQE-30 vector with the WT or G82C *cdoA* gene were grown in TB+G medium (12 g/L tryptone, 24 g/L yeast extract, 8 mL/L glycerol with 100 mL/L 0.17 M KH_2PO_4/0.72 M K_2HPO_4) at 37 °C and 250 rpm, and expression was induced with addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at an OD_{600} of ~8.0. Ferrous ammonium sulfate (FAS) was added to a final concentration of 500 μM at the time of induction to increase Fe incorporation into the CDO active site. Cells grew for an additional four hours after induction. Filtered cell lysate in IMAC A buffer (20 mM Tris, 5 mM imidazole, 500 mM NaCl, pH 8.0) was applied to an immobilized metal affinity chromatography column and eluted with an increasing

gradient of IMAC B (20 mM Tris, 500 mM imidazole, 500 mM NaCl, pH 8.0) buffer. Fractions containing CDO as determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) were pooled and activity of the purified enzyme was confirmed qualitatively using thin-layer chromatography as previously described.[12]

WT *Mm*CDO used for EPR, NMR and UPLC-MS was produced using a codon-optimized *cdo1* gene in the pVP16 expression vector with the gene for an attached maltose binding protein (MBP) as a solubility tag. *E. coli* Rosetta 2(DE3) cells containing the pVP16 expression vector with the *cdo-mbp* gene were grown in supplemented Terrific Broth (24 g/L yeast extract, 12 g/L tryptone, 89 mM phosphate, 2 mM $MgSO_4$, 108.6 mM glycerol, and 28.17 mM aspartate) containing chloramphenicol and carbenicillin (34 μg/mL and 100 μg/mL, respectively) at 37 °C and 250 RPM. When the cell cultures reached OD_{600}≈4.5, ferrous ammonium sulfate [$(NH_4)_2Fe(SO_4)_2$] was added to a final concentration of 95 μM. Once cell culture density met OD_{600}≈9, the shaker temperature was reduced to 25 °C, and *Cdo1* gene overexpression was induced by adding isopropyl β-D-1 thiogalactopyranoside (IPTG), D-lactose, Casamino acids and additional ferrous ammonium sulfate to final concentrations of 165 μM, 8 mM, 0.20% (w/v), and 190 μM, respectively. Filtered cell lysate in 25 mM HEPES, 300 mM NaCl, pH 7.9 buffer was applied to a TALON column. Fractions containing CDO-MBP as determined by SDS-PAGE were pooled. Approximately 1 mg of tobacco etch virus protease was added to the pooled fractions per 50 mg of protein to cleave MBP from CDO, and the solution was dialyzed overnight against 2 L of the HEPES/NaCl buffer. The cleaved and dialyzed protein was then again applied to a TALON column. Fractions containing purified CDO as determined by SDS-PAGE were pooled and the activity of the purified enzyme was confirmed qualitatively using thin-layer chromatography as previously described.[12] C93G *Mm*CDO was produced and purified following the same procedure as was used above for WT *Mm*CDO.

WT *Mm*CDO used for MCD spectroscopy was produced from a non-codon optimized *cdo1* gene in the same vector. It was produced and purified the same way, except that a diethylaminoethyl (DEAE) cellulose anion exchange column used to purify MBP-CDO fusion, and an amylose affinity column was used to purify cleaved CDO.

Sample Preparation for Spectroscopy. Protein samples for EPR spectroscopy were prepared in a buffer of 20 mM Tris, 5 mM imidazole, 500 mM NaCl, pH 8.0. The Fe(II) and Fe(III) contents of the protein were determined via a colorimetric assay with the iron chelator tripyridyl triazine (TPTZ) and an ε_{595} of 22.1 mM^{-1} cm^{-1}.[21] To increase the iron content, protein samples were incubated anaerobically with a 2-fold molar excess of FAS to protein for 30 min. Chelex 100 was added to remove unbound iron from solution. Typical iron incorporation was ~60% after reconstitution. Ammonium hexachloroiridate was added in a 3-fold molar excess over iron-loaded protein to oxidize it to the Fe(III) state. Excess oxidant was removed using either a PD-10 desalting column or buffer exchange through a 10 kDa Centricon filter. *Bs*CDO EPR samples were prepared with 0.5 mM Fe(III)-loaded protein, a 10-fold molar excess of Cys or a 15-fold molar excess of Cys analogue, and a 100-fold molar excess of azide. *Mm*CDO samples were prepared with a ~45-60 molar excess of Cys (analogue) over total iron content, and a ~10 molar excess azide over Cys (analogue).

Protein samples for MCD spectroscopy were prepared anaerobically via use of a Schlenk line or a glove box. Protein in a buffer of 20 mM Tris, pH 8.0, with 5 mM imidazole and 500 mM NaCl was reconstituted with FAS as described above. Tris(2-carboxyethyl)phosphine (TCEP) was added in a 3-fold molar excess over iron-loaded protein to reduce it to the Fe(II) state. Samples were then concentrated to 1.4 mM Fe(II)-loaded protein. *Bs*CDO samples were incubated anaerobically with a 10-fold molar excess of Cys or a 15-fold molar excess of Cys analogue, and a 100-fold molar excess of azide over total iron content. *Mm*CDO samples were incubated anaerobically with an ~5-25 molar excess of substrate (analogue) over total iron

content and at least a ~15-35 molar excess azide over substrate (analogue) content. 55% (v/v) glycerol was added as a glassing agent.

Spectroscopy. X-band EPR data were collected using a Bruker ELEXSYS E500 spectrometer. Sample temperature was maintained at 20 K by an Oxford ESR 900 continuous flow liquid He cryostat regulated by an Oxford ITC-503S temperature controller. All EPR spectra were obtained using the following experimental parameters: frequency = 9.381 GHz; microwave power = 12.62 mW; modulation amplitude = 3 G; and modulation frequency = 100 kHz.

Low-temperature MCD spectra were recorded with a Jasco J-715 spectropolarimeter in conjunction with an Oxford Instruments SpectromagPT 7 Tesla magnetocryostat. MCD spectra are presented as the difference between spectra obtained with the magnetic field aligned parallel and antiparallel to the light propagation axis to eliminate contributions from the circular dichroism background and glass strain.

UPLC-MS Assays. Product formation under high substrate concentrations was measured using a Waters ultra-high performance liquid chromatography mass spectrometer (UPLC-MS).[22–24] All samples were run on was collected on an Intrada column from Imtakt, except for 3-MPA samples, which were run on a C18 column. Protein samples were thawed in hand and then exchanged into a buffer of 100 mM sodium phosphate, 20 mM NaCl, 0.1 mM BCS, pH 7.5. The concentration of Fe(II)-loaded protein was determined via the TPTZ assay described above.[21] 200 µL reactions of ~10 µM protein and 200 mM Cys (analogue) were run at 37 °C for 2 hrs. Reactions with HCys were run at 150 mM HCys. 50 µL reaction aliquots were quenched every 3 min by the addition of 50 µL of equal volumes of 1 M HCl and 1 mM asparagine, which served as an internal standard. Quenched reaction aliquots were centrifuged at 15,000 × g for 3 min and the supernatants were transferred to a 96-well plate. 2 µL of samples were injected onto the UPLC-MS. Samples from Cys, 2-AET, and HCys reactions were applied to the Intrada column under 80% buffer A (acetonitrile with 0.1% (v/v) formic acid), 20% buffer B (100 mM ammonium formate) and species eluted during a gradual increase to 80% buffer B were analyzed with the

mass detector. Samples from 3-MPA reactions were applied to the C18 column under 99% buffer C (water with 0.1% (v/v) formic acid), 1% buffer D (acetonitrile with 0.1% (v/v/) formic acid) and species eluted during a gradual increase to 99% buffer D were analyzed with the mass detector. Graphs of intensity at a specified mass over time are reported as the difference between the sample and a 2 hour no enzyme control to account for any autooxidation that may occur in solution. CSA was detected at 152 negative, HT was detected at 110 positive, 3-SPA was detected at 137 negative and HCys was detected at 168 positive.

^{1}H-NMR Assays. NMR spectra were acquired using a Bruker Ascend 500 spectrometer. Reactions for analysis by NMR were run in the same way as UPLC-MS reactions, except that 20 mM substrate (analogue) was used, and the reaction was not quenched. Instead, a 145 μL aliquot was pulled from the reaction for each timepoint and 5 μL of a trimethylsilylpropanoic acid (TSP) solution was added as an internal standard. Because the reactions were not quenched, reported timepoints are approximate.

3.3 Results

MCD Spectroscopy of Substrate (Analogue)-Bound Fe(II)CDOs

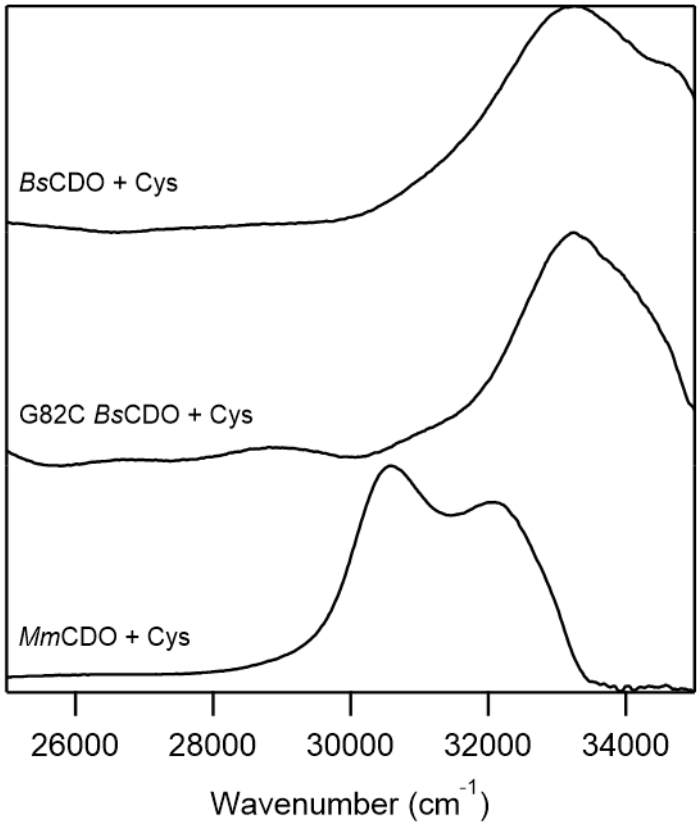


Figure 3.3 MCD spectra of Cys-Fe(II)CDOs. G82C *Bs*CDO is approximately 32% crosslinked as determined by SDS-PAGE densitometry.

In this study, we have compared various substrate (analogue)-bound forms of WT *Mm*CDO, WT *Bs*CDO, and the G82C *Bs*CDO variant, which we have shown forms a C-Y crosslink analogous to that found in WT *Mm*CDO. Low temperature (LT) MCD spectra of the three reduced, Fe(II)-bound CDOs incubated with the native substrate Cys are shown in Figure 3.3. Fe(II) is the catalytically relevant state of CDO. For WT *Mm*CDO, the pair of positively-signed features (> 30,000 cm^{-1}) were previously attributed to S→Fe(II) CT transitions.[12] In WT and G82C *Bs*CDO, the analogous features are blue-shifted by ~2000 cm^{-1}. A previous study of the H155A variant of *Mm*CDO, in which the active site hydrogen bonding network is disrupted and the positional freedom of the C-Y crosslink increased, revealed a similar blue shift that was

computationally shown to arise from the coordination of a water molecule at the site where molecular oxygen binds during Cys turnover, leading to formation of a 6-coordinate Fe(II) center.[17] By analogy, we attribute the blue-shift of the S→Fe(II) CT transitions in WT and G82C *Bs*CDO to the binding of a water molecule to create a 6-coordinate Fe(II) center. In non-crosslinked WT *Mm*CDO, untethered C93 is hypothesized to stabilize this water molecule, increasing the activation energy for O_2 binding and thus lowering the catalytic efficiency of enzymatic Cys turnover.[17] In WT *Bs*CDO, there is a G at the position analogous to C93 in *Mm*CDO, and thus the bound water is not stabilized and we observe no decrease in the catalytic efficiency relative to fully crosslinked WT *Mm*CDO. The crosslink-capable G82C *Bs*CDO variant exhibits decreased catalytic efficiency and the efficiency increases as the fraction of crosslinked protein increases, suggesting that the newly introduced C residue stabilizes the bound water molecule in the non-crosslinked form as in WT *Mm*CDO. However, because the S→Fe(II) CT transitions of 34%-crosslinked G82C *Bs*CDO and WT *Bs*CDO are similarly blue shifted from those of WT *Mm*CDO, the crosslink in G82C *Bs*CDO likely has more conformational freedom than the analogous crosslink in WT *Mm*CDO does.

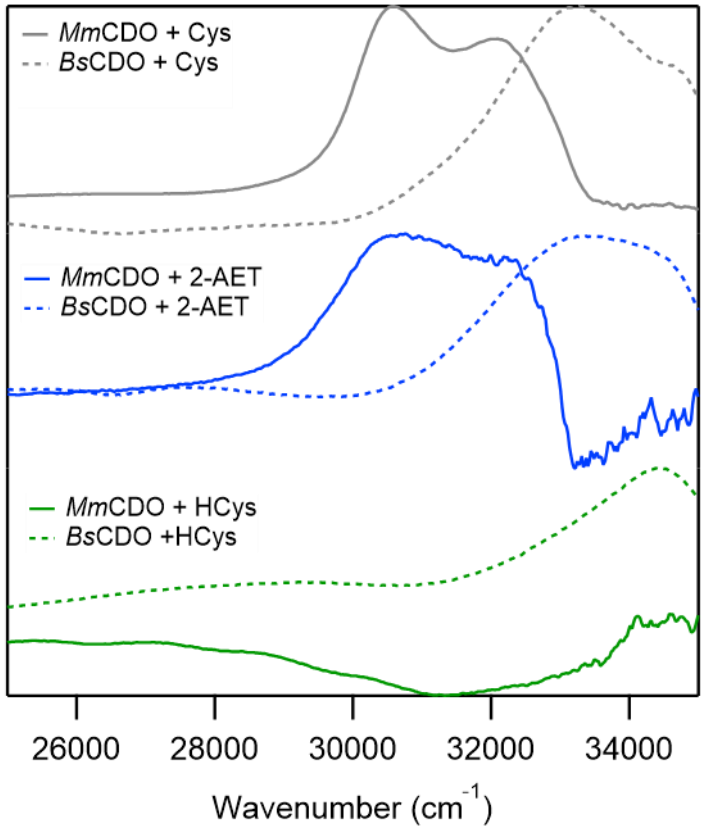


Figure 3.4 MCD spectra of WT Fe(II)MmCDO (solid line) and Fe(II)BsCDO (dashed line) incubated with Cys, 2-AET, and HCys.

We also collected MCD spectra of Fe(II)-bound WT *Mm*CDO and *Bs*CDO incubated with substrate analogues (Figure 3.4). The LT MCD spectrum of *Mm*CDO incubated with 2-AET shows a positive feature at ~30,500 cm^{-1} and a higher energy shoulder at ~32,000 cm^{-1}, similar to Cys-bound *Mm*CDO. These features are therefore also assigned as S→Fe(II) CT transitions, which indicates that 2-AET directly coordinates to Fe(II)CDO in a similar fashion as Cys. In contrast, no temperature dependent features are observed above 30,000 cm^{-1} in the MCD spectrum of WT *Mm*CDO incubated with HCys. The lack of features in this region suggests that under the conditions used to collect this spectrum, Hcy does not directly bind to the majority of Fe(II) centers of WT *Mm*CDO. Upon addition of Cys to this sample, the S→Fe(II) CT features characteristic of Cys-bound Fe(II)*Mm*CDO appeared in the corresponding MCD spectrum (not show), providing further support that Fe(II)*Mm*CDO has a very low affinity for HCys. However, it

should be noted that a previous Mössbauer study afforded evidence for HCys binding to Fe(II)*Mm*CDO, so it is possible that the presence of glycerol, required to obtain MCD spectra, may affect the ability of Fe(II)*Mm*CDO to bind HCys.

The LT MCD spectrum of WT *Bs*CDO incubated with 2-AET closely resembles that obtained in the presence of Cys. Thus, 2-AET likely binds to WT *Bs*CDO in a similar manner to Cys with a water molecule occupying the sixth coordination site of the Fe(II) ion. Interestingly, HCys also appears capable of binding to WT Fe(II)*Bs*CDO based on the appearance of a positively signed, temperature-dependent feature around 34,500 cm^{-1} in the MCD spectrum of WT *Bs*CDO inclubated with HCys. The difference between *Bs*CDO and *Mm*CDO in their ability to bind the bulky Hcy substrate analogue under the same conditions indicates that the size or structure of the *Bs*CDO active site is likely more accommodating to larger molecules. This is reinforced by the observation that Cys-bound *Bs*CDO binds a water molecule even in the presence of a C–Y crosslink (G82C *Bs*CDO, Figure 3.2).

Abs Spectroscopy of Substrate (Analogue)- and Azide-Bound Fe(III)CDOs

Although only Fe(II)CDO can turnover Cys, there is precedence for an Fe(III)-superoxo catalytic intermediate.[12] A previous spectroscopic study demonstrated that azide is capable of binding to Fe(III)*Mm*CDO only in the absence of substrate, and a subsequent X-ray crystallographic study showed that azide and S_{Cys} compete for the same binding site.[20] When azide is added to Cys-bound Fe(III)CDO, the Cys-Fe coordinate bonds are strong enough that azide does not displace Cys, instead binding elsewhere in the active site and causing minor perturbations to the complex's S=5/2 EPR signal.[20] Thus, we have added azide to Cys analogue-bound Fe(III)CDO as a probe of the analogue-Fe(III) coordinate bonds.

Fe(III)*Mm*CDO is yellow in color. When substrate binds it turns blue and displays an Abs feature at ~15,875 cm^{-1}. Azide-bound Fe(III)*Mm*CDO turns red and displays an Abs feature at ~24,000 cm^{-1}.[20] Thus, we would expect that if azide displaces the substrate from its coordination

site to the Fe(III), the enzyme complex would also turn red. Cys-Fe(III)*Mm*CDO incubated with excess azide looks nearly identical to Cys-Fe(III)*Mm*CDO without azide, with only a modest red-shift from ~15, 875 cm^{-1}, as azide is incapable of displacing the Cys from its coordination site.[20]

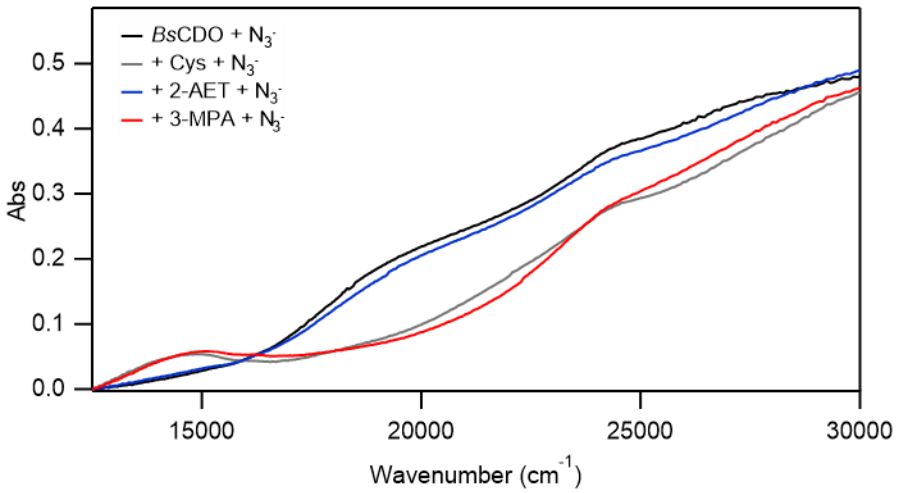


Figure 3.5 RT Abs spectra of WT BsCDO incubated with azide in the absence or presence of substrate (analogue).

RT Abs spectra of substrate (analogue)-bound *Bs*CDO display a similar pattern (Figure 3.5). The spectrum of *Bs*CDO incubated with azide exhibits a broad feature centered ~22,500 cm^{-1}. As with *Mm*CDO, Cys-Fe(III)*Bs*CDO incubated with excess azide instead has a lower energy feature, here at ~14,500 cm^{-1}, indicating that azide is not capable of displacing Cys from its position coordinating to the Fe(III). The Abs spectrum of 3-MPA-bound Fe(III)*Bs*CDO in the presence of azide contains the same low energy peak, indicating that azide also cannot displace this susbtrate analogue. In contrast, 2-AET-bound Fe(III)*Bs*CDO incubated with azide, however, has an almost identical Abs spectrum to azide-bound Fe(III)*Bs*CDO, and 2-AET is therefore displaced from its coordination site by azide.

EPR Spectroscopy of Substrate (Analogue)- and Azide-Bound Fe(III)CDOs

EPR spectroscopy has been used frequently to gain insight into both the primary- and secondary- coordination environment of the Fe(III) center of CDO. A previous study of Cys-

bound Fe(III)*Mm*CDO incubated with excess azide found only a very weak high field S=1/2 signal produced by the 6-coordinate complex with both substrate and azide bound directly to the Fe(III) center.[20] The S=5/2 signal seen centered at ~1500 Gauss for Cys-bound Fe(III)*Mm*CDO is retained upon incubation with azide, with only minor perturbation. Thus azide does not displace Cys from the active site and bind directly to the Fe(III) center, but rather binds somewhere in the secondary coordination sphere, likely near R60 (*Mm*CDO numbering).[20]

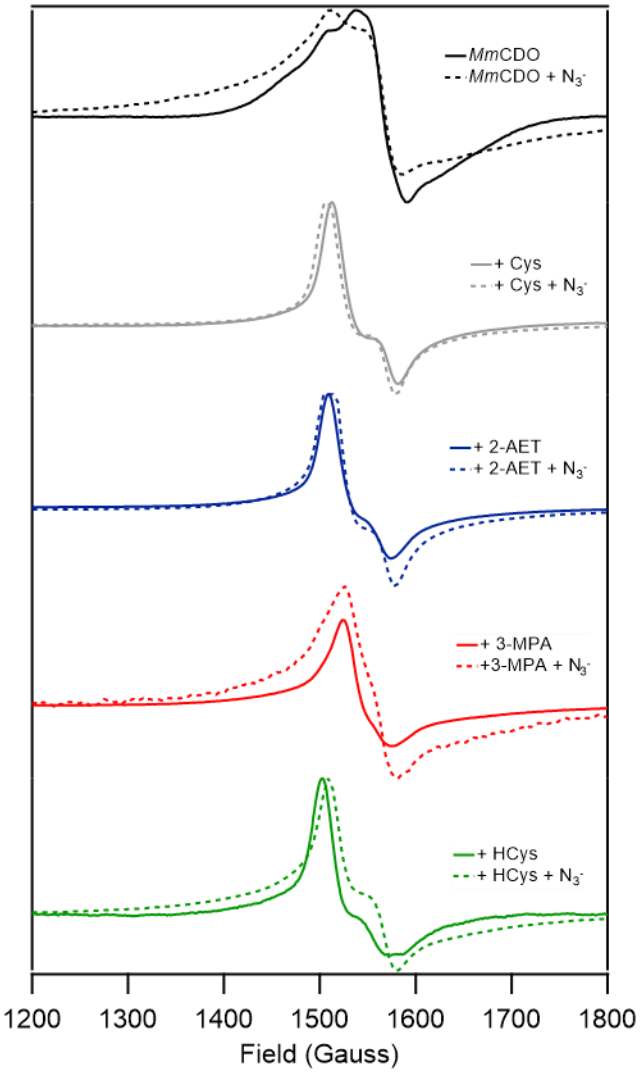


Figure 3.6 EPR spectra of substrate (analogue)-bound Fe(III)*Mm*CDO with (solid line) and without (dashed line) the addition of a 100-fold excess of azide. .

In this study, we employed EPR spectroscopy to further probe the ability of azide to either displace Fe(III)-coordinated substrate (analogue) or else to perturb the electronic structure of the active site in various analogue-bound Fe(III)CDO complexes. Figure 3.6 shows the substrate (analogue)-bound complexes of Fe(III)*Mm*CDO with and without addition of a 100-fold excess of azide. All samples of substrate-bound *Mm*CDO in the presence of azide exhibit S=5/2 EPR signals that are markedly different from that of N_3^--Fe(III)*Mm*CDO, indicating that azide is not capable of displacing any of the substrate (analogues) under these conditions. The 3-MPA-Fe(III)*Mm*CDO signal shows a large increase in g-spread after azide addition, likely because 3-MPA lacks the amine group that allows Cys to form one of two coordinate bonds with the Fe(III) center. The extent of broadening of the HCys-Fe(III)*Mm*CDO signal upon azide addition is intermediate between that of the Cys and 2-AET-bound samples and the 3-MPA-bound sample. HCys has also been shown to coordinate the Fe(III) monodentate, but it's bulkiness likely limits its conformational freedom inside the active site relative to 3-MPA.

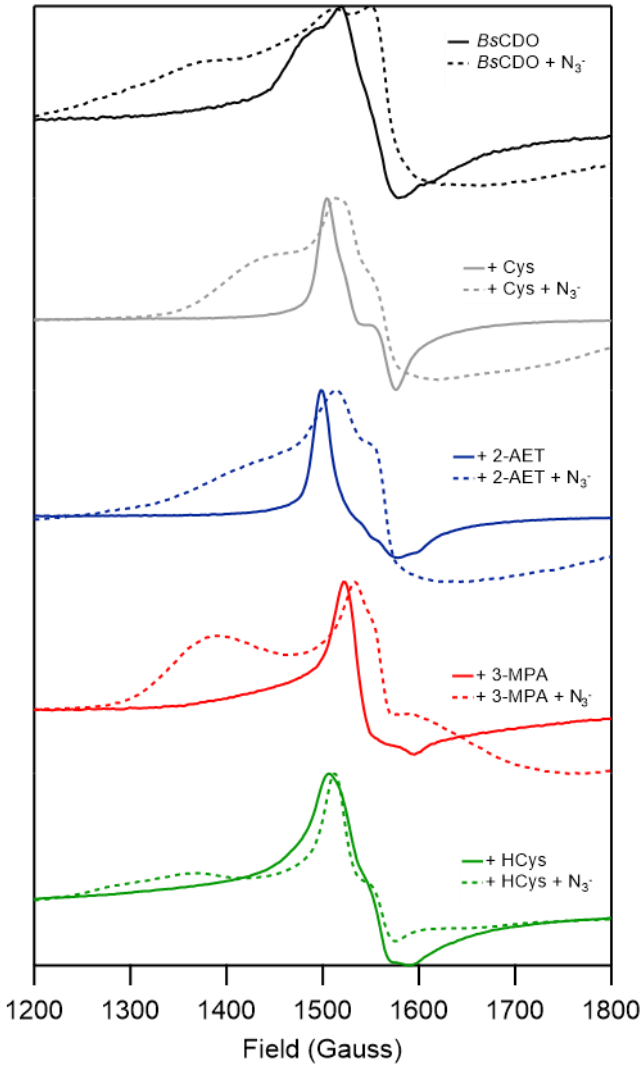


Figure 3.7 EPR spectra of substrate (analogue)-bound Fe(III)BsCDO with (solid line) and without (dashed line) the addition of a 100-fold excess of azide.

Figure 3.7 shows the substrate (analogue)-bound complexes of Fe(III)*Bs*CDO with and without addition of a 100-fold excess of azide. In comparison to WT *Mm*CDO, all samples of substrate-bound *Bs*CDO in the presence of azide exhibit EPR spectra with considerably broader S=5/2 signals. This implies a greater level of conformational freedom in the *Bs*CDO active site than in the *Mm*CDO active site. The 2-AET-Fe(III)*Bs*CDO spectrum after azide addition is nearly identical to that of N_3^--Fe(III)*Bs*CDO, indicating that azide displaces 2-AET to coordinate directly

to the majority of Fe(III) centers. This matches what we observe in our Abs spectra of substrate (analogue)-bound Fe(III)*Bs*CDO. As with *Mm*CDO, the monodentate binding of 3-MPA affords it the largest amount of conformational freedom, and the bulky nature of HCys limits its movement within the active site.

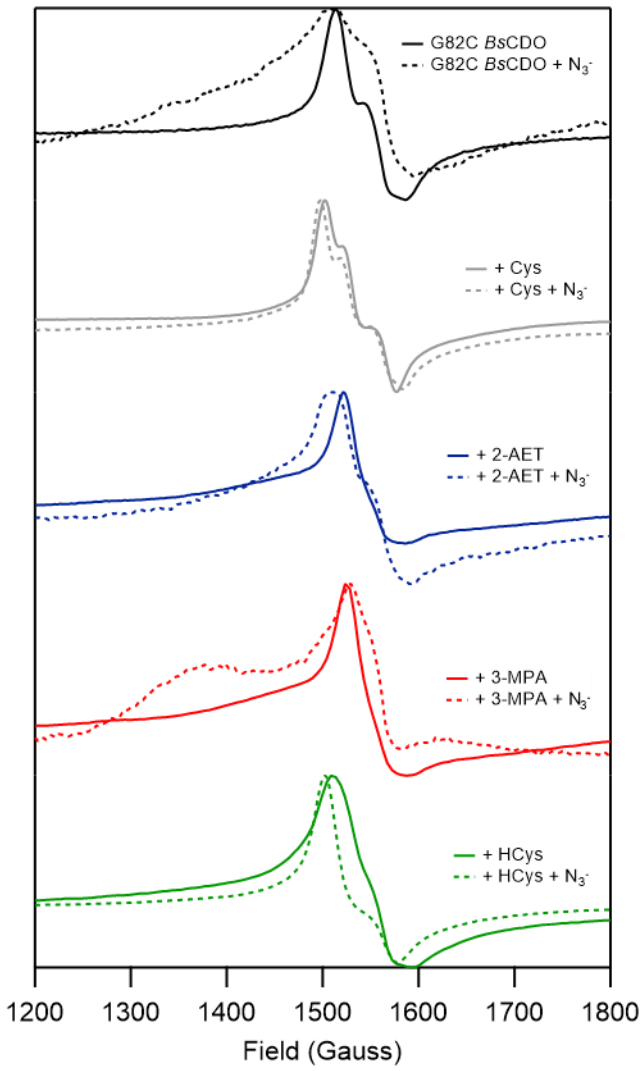


Figure 3.8 EPR spectra of ~62%-crosslinked substrate (analogue)-bound G82C Fe(III)*Bs*CDO with (solid line) and without (dashed line) the addition of a 100-fold excess of azide. .

Figure 3.8 shows the EPR spectra of substrate (analogue)-bound G82C Fe(III)*Bs*CDO with and without addition of a 100-fold excess of azide. Interestingly, all samples excluding 3-MPA exhibit less broadening in response to azide addition compared to WT *Bs*CDO. In fact,

these spectra look very similar to those exhibited by azide and substrate (analogue)-bound *Mm*CDO complexes. This indicates that formation of the C-Y active site crosslink (~62% in this G82C *Bs*CDO prep as determined by SDS-PAGE gel densitometry) plays a significant role in limiting conformational freedom of bound substrate in the active site of Fe(III)CDO. The 3-MPA-Fe(III)G82C *Bs*CDO spectrum, however, exhibits a similar level of inhomogeneous broadening as observed for the WT *Bs*CDO version of the complex upon addition of azide. This suggests that the *Bs*CDO active site is less constrained or compact in a way that cannot be attributed simply to the lack of the active-site crosslink. In particular, there seems to be more room for substrate flexibility near the R60-containing pocket of the active site, as the 3-MPA-Fe(III)G82C *Bs*CDO complex is afforded a much greater level of conformational freedom than the other azide-incubated substrate(analogue)-Fe(III)G82C *Bs*CDO complexes. The crystal structures of Cys-bound WT *Mm*CDO (4JTO) and WT *Bs*CDO (4QM9) show R60 0.2 Å further away from the Cys carboxylate group in *Bs*CDO, indicating a weaker salt bridge as well as a slightly less constrained active site pocket around this end of the substrate.

EPR Spectroscopy of Substrate (Analogue)- and Cyanide-Bound Fe(III)CDOs

Cyanide has also been used as a superoxide surrogate in previous studies of Fe(III)*Mm*CDO to assess the importance of second-sphere interactions.[9] Cyanide binds to the same iron coordination site as molecular oxygen, and like oxygen cannot bind unless substrate is already coordinated to the iron center. The strong π-accepting character of cyanide produces a low-spin cyano/substrate-bound Fe(III)*Mm*CDO complex with a rhombic EPR signal that is sensitive to the presence or absence of the C-Y crosslink.[9]

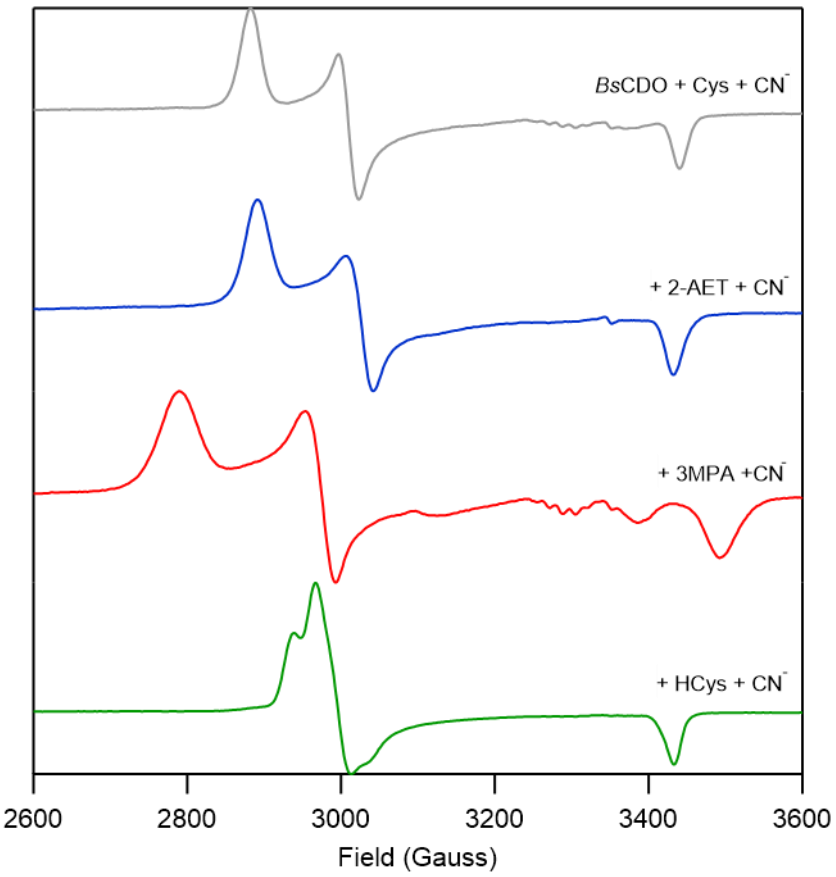


Figure 3.9 EPR spectra of substrate (analogue)- and cyanide-bound Fe(III)*Bs*CDO. Only the S=1/2 signal corresponding to binding of both substrate (analogue) and cyanide to the Fe(III) center is shown. Variable contributions from the S=5/2 signal of only substrate (analogue)-bound complex is observed at lower field.

We collected EPR spectra of cyano/substrate (analogue)-bound WT and G82C Fe(III)*Bs*CDO complexes to assess the effect of crosslink formation on substrate binding and positioning in *Bs*CDO. Figure 3.9 shows the EPR spectra of (CN^-/substrate(analogue))-Fe(III)*Bs*CDO complexes. We previously published the (CN^-/Cys)-Fe(III)*Bs*CDO EPR spectrum, which only exhibits a single S=1/2 signal with g values that align closely with those of non-crosslinked (CN^-/Cys)-Fe(III)*Mm*CDO. The (CN^-/2-AET)-Fe(III)*Bs*CDO spectrum is very similar to that of the (CN^-/Cys)-bound complex, indicating that removal of the carboxylate group from the substrate has little to no effect on the positioning of the cyano and substrate ligands in the naturally non-crosslinked WT *Bs*CDO. In contrast, the (CN^-/3-MPA)-Fe(III)*Bs*CDO signal exhibits a much wider g spread. As with the azide-incubated 3-MPA-Fe(III)CDO complexes, this

can likely be attributed to 3-MPAs inability to coordinate to the Fe(III) center through an amine group unlike Cys. The (CN^-/HCys)-Fe(III)*Bs*CDO EPR spectrum, which displays a markedly narrower g spread than those of the analogous Cys and 2-AET complexes, differs significantly from any other (CN^-/substrate(analogue))-Fe(III) CDO EPR spectrum reported to date. Efforts are underway to explore the origin of this unusual EPR signal.

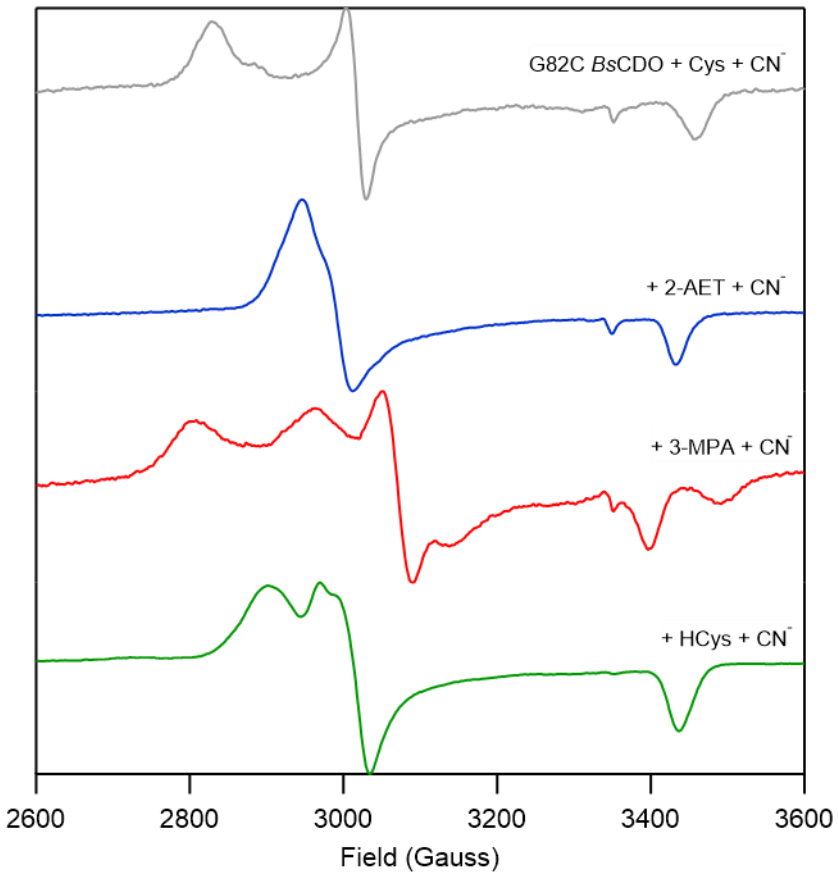


Figure 3.10 EPR spectra of substrate (analogue)- and cyanide-bound Fe(III)G82C *Bs*CDO. Only the S=1/2 signal corresponding to binding of both substrate (analogue) and cyanide to the Fe(III) center is shown. Some S=5/2 signal of only substrate (analogue)-bound complex is observed at lower field.

Figure 3.10 shows the EPR spectra of (CN^-/substrate(analogue))-bound G82C Fe(III)*Bs*CDO complexes. We previously published the spectrum of (CN^-/Cys)-bound G82C Fe(III)*Bs*CDO, which is capable of crosslink formation and thus displayed two distinct S=1/2 signals. The signal with the larger g spread was attributed to the crosslinked form of the enzyme complex, and the signal with the smaller g spread was attributed to the non-crosslinked

complex. The Fe(III)-C-N moiety in the (CN^-/Cys)-Fe(III)*Mm*CDO complex has been shown to be nearly linear in the absence of crosslink. In response to crosslink formation, the Fe(III)-C-N angle increases slightly. The crosslink also shifts the position of Y157, in effect moving the carboxylate moiety of substrate Cys so that both of its oxygen atoms are positioned closer to the guanidyl group of R60. This leads to a lengthening of the Fe-S bond by ~0.02 Å.[9]

Interestingly, the EPR spectrum of (CN^-/2-AET)-bound)G82C Fe(III)*Bs*CDO shows a single S=1/2 signal, which displays a substantially smaller g spread than the analogous WT *Bs*CDO complex and much more resembles that of (CN^-/HCys)-Fe(III)*Bs*CDO. The sensitivity of the (CN^-/Cys)-Fe(III)*Bs*CDO complexes, then, may be due in large part to the distinct positioning of the Cys carboxylate moiety in the presence and absence of crosslink. The G82C (CN^-/3-MPA)-Fe(III)*Bs*CDO EPR spectrum clearly exhibits two distinct signals; however, the relative intensities of the two signals do not match the relative intensities of the two signals in the EPR spectrum of G82C (CN^-/Cys)-Fe(III) *Bs*CDO. This could indicate a higher affinity of the non-crosslinked form of 3-MPA-bound G82C Fe(III)*Bs*CDO for cyanide than the crosslinked form. Just as in WT *Bs*CDO, the G82C (CN^-/3-MPA)-Fe(III)*Bs*CDO EPR signal is overall broader than those of the other CN^-/substrate(analogue) complexes, likely due to the expected increased conformational freedom from the lack of a substrate(analogue) amine-Fe(III) bond. The G82C (CN^-/HCys)-Fe(III)*Bs*CDO spectrum appears to show two distinct signals, but the difference in g-spread between the signals is small compared to that exhibited by the G82C (CN^-/Cys)-Fe(III)*Bs*CDO counterpart While the differences between the EPR spectra presented in Figures 3.9 and 3.10 are much larger than anticipated, more work is needed to interpret these exciting results.

UPLC-MS Assay of High Substrate (Analogue) Concentration CDO Reactions

To assess the ability of WT *Mm*CDO and *Bs*CDO to turn over substrate analogues, we used UPLC-MS. To minimize the effect of differences in binding affinity between the various

analogues, we ran these reactions at 150-200 mM substrate (analogue) and ~10 μM enzyme. Although not physiologically relevant, this allowed us, to some level of approximation, to observe the effect of the analogues only on the turnover rate of the enzyme. Because the various Cys analogues and their oxygenated products did not ionize well on the MS under standard running conditions, 200 mM substrate (analogue) gave mass peaks of reasonable intensities for integration.

Table 3.1 shows the integration of the MS peak corresponding to the mass of the sulfinic acid product of each substrate (analogue) at three time points throughout its reaction with either *Mm*CDO or *Bs*CDO. Each MS peak had the integrated area of a 2 hour no-enzyme control subtracted to account for any auto-oxidation that may be occurring in solution. Thus, a negative area indicates that the enzyme produced less product in the given amount of time compared to the auto-oxidized reaction in solution over 2 hours. Graphs of the background-subtracted MS peaks are shown in Figure A.3.1. The same data is displayed graphically in Figure 3.11, which reports the MS peak area for each oxygenated product of each substrate analogue produced in two hours relative to the amount of CSA produced in 2 hours by the same enzyme.

Table 3.1 Areas of the MS peaks corresponding to the product of each reaction of substrate (analogue) with *Mm*CDO and *Bs*CDO. For each reported peak value, the area of any product peak that appeared in a no-enzyme control was subtracted

		30 min	1 hour	2 hour
Cys	*Mm*CDO	3.09×10^4	4.76×10^4	6.85×10^4
	*Bs*CDO	3.21×10^3	1.33×10^4	2.89×10^4
2-AET	*Mm*CDO	-3.01×10^4	-6.35×10^3	1.60×10^4
	*Bs*CDO	1.06×10^5	1.72×10^5	1.30×10^5
3-MPA	*Mm*CDO	1.23×10^4	1.23×10^4	1.62×10^4
	*Bs*CDO	1.22×10^3	4.16×10^3	8.76×10^3
HCys	*Mm*CDO	1.45×10^5	1.43×10^5	1.84×10^5
	*Bs*CDO	5.83×10^4	2.60×10^5	3.04×10^5

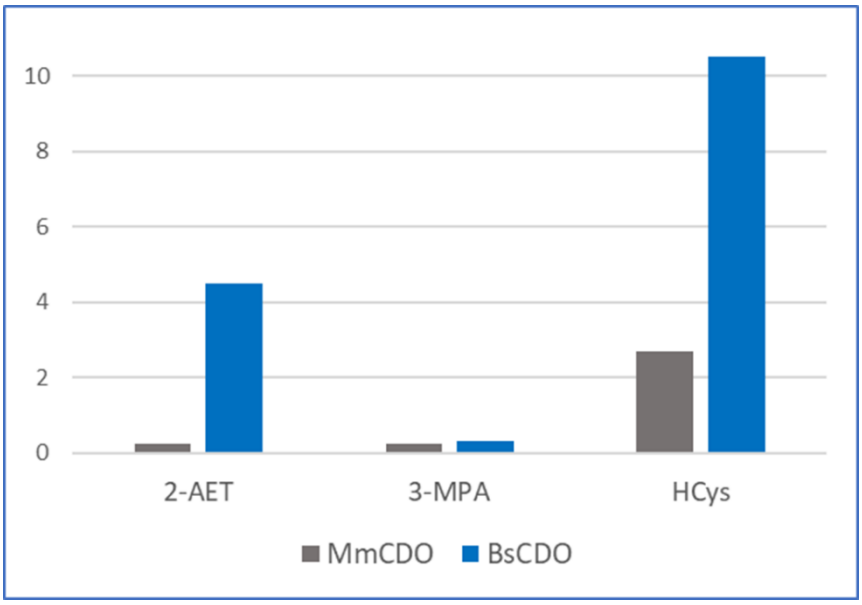


Figure 3.11 Histogram of the ratio of MS peak areas for oxygenated product of each substrate analogue relative to CSA produced in 2 hours by *Mm*CDO (gray) and *Bs*CDO (blue). Ratios are calculated from the values given for the 2-hour time point in Table 3.1

As expected, both *Mm*CDO and *Bs*CDO turn over native substrate Cys. The larger product peak areas for *Mm*CDO compared to *Bs*CDO with Cys indicate that *Mm*CDO was at a slightly higher concentration in reaction, as the two CDOs are reported to have the same catalytic efficiencies.

The no-enzyme control of 2-AET showed a significant amount of auto-oxidation after 2 hours. The product peak area for *Mm*CDO , after subtracting the no-enzyme control, is negative at 30 min, less negative at 1 hr, and positive after 2 hr. This indicates that *Mm*CDO is able to turn over 2-AET but at a rate only slightly faster than auto oxidation. *Bs*CDO, on the other hand, has increasingly larger product peak areas over time and thus turns over 2-AET at a higher rate than *Mm*CDO. The product peak area decreases slightly from 1 hour to 2 hours in the *Bs*CDO reaction, but with the rate of auto oxidation so high we may have simply proceeded past the linear regime of the enzyme's 2-AET turnover kinetics.

Both *Mm*CDO and *Bs*CDO can turn over 3-MPA to roughly the same extent. Both enzymes are also capable of HCys turnover. Although we did not obtain any evidence of HCys binding to Fe(II)*Mm*CDO by MCD spectroscopy, we clearly see turnover (and thus binding) in

the absence of glycerol when HCys is present in extreme excess. *Bs*CDO turns HCys over significantly faster than *Mm*CDO.

NMR Assay of Low Substrate (Analogue) Concentration CDO Reactions

To assay substrate analogue turnover under physiologically more relevant conditions, as well as to compare the relative binding affinities of *Mm*CDO and *Bs*CDO for these analogues, we turned to ^{1}H nuclear magnetic resonance (^{1}H-NMR) spectroscopy. Reactions of 20 mM substrate (analogue) were run with ~10 μM enzyme. Enzyme was kept at a concentration low enough relative to that of substrate and product that it did not contribute to the NMR spectra. Again, *Bs*CDO was at a slightly lower concentration than *Mm*CDO. ^{1}H-NMR spectra were collected at two different time points throughout the reactions. Spectra were also taken of standard solutions of Cys, cystine, CSA, 2-AET, hypotaurine (HT), 3-MPA, and HCys, but not of 3-sulfinopropionic acid (3-SPA) and homocysteine sulfinic acid (HCSA) as these compounds are not readily available. Differences in resolution between spectra are due to differences in shimming quality between samples, which can be inconsistent for complex samples in aqueous solvents. Minor shifts of peaks from the same proton signal can occur between samples as well, usually due to small changes in pH or salt concentration.

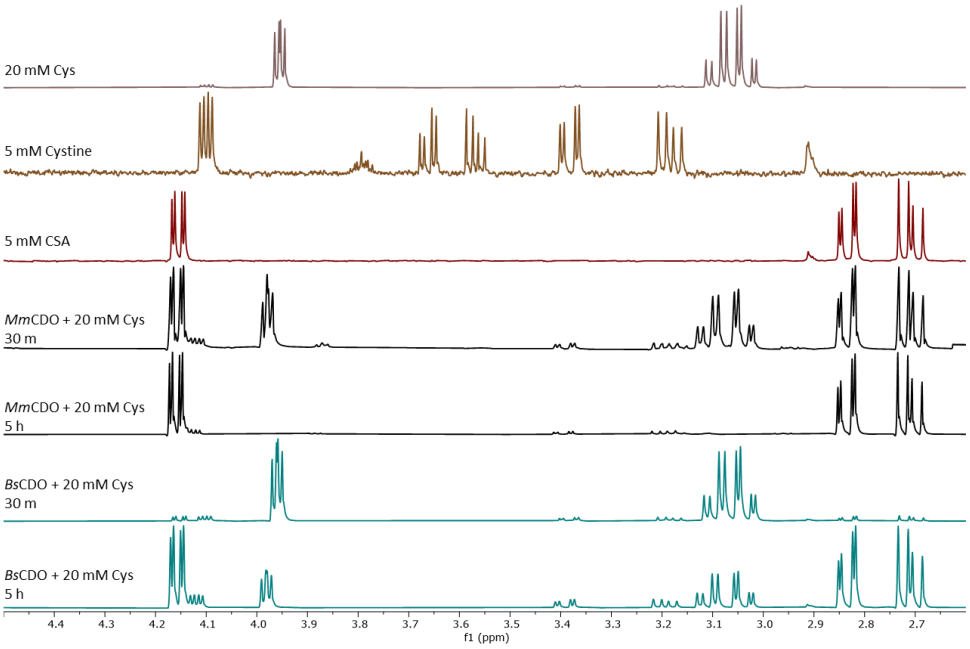


Figure 3.12 ^{1}H-NMR spectra collected at two time points after incubating *Mm*CDO (black) and *Bs*CDO (teal) with native substrate Cys. Spectra of standard solutions of reactant Cys, disulfide cystine, and product CSA are shown at the top.

Table 3.2 Relative peak areas of product and reactants from the NMR spectra in **Figure 3.12**

		Cys	**Cystine**	**CSA**
Standard	20 mM Cys	1.00	0.05	0.00
***Mm*CDO**	20 mM Cys 30 m	1.00	0.10	1.36
	20 mM Cys 5 h	1.00	2.47	36.04
***Bs*CDO**	20 mM Cys 30 m	1.00	0.06	0.07
	20 mM Cys 5 h	1.00	0.31	2.41

Figure 3.12 shows the ^{1}H NMR spectra of reaction mixtures of *Mm*CDO and *Bs*CDO 30 min and 5 hours after incubation with Cys, along with those of Cys, cystine, and CSA standards. The ^{1}H-NMR spectra of both the *Mm*CDO and *Bs*CDO reactions lose the peaks associated with Cys and gain peaks associated with CSA. A small amount of cystine also appears over time. Table 3.2 shows the ratio of peak areas from the beta protons of each compound in the different samples, where the area of the Cys beta proton signal is set to one.

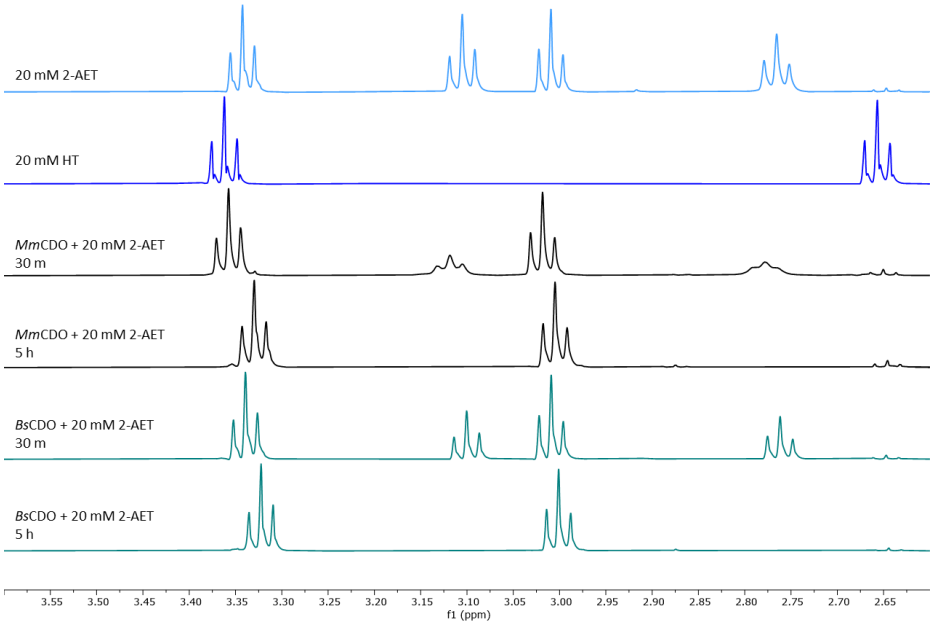


Figure 3.13 ^{1}H-NMR spectra collected at two time points after incubating MmCDO (black) and BsCDO (teal) with substrate analogue 2-AET. Spectra of standard solutions of reactant 2-AETand product HT are shown at the top.

Table 3.3 Relative peak areas of product and reactants from the NMR spectra in **Figure 3.13**

		2-AET	**Cystamine**	**HT**
Standard	20 mM 2-AET	1.00	0.92	0.05
***Mm*CDO**	20 mM 2-AET 30 m	1.00	2.09	0.18
	20 mM 2-AET 5 h	1.00	30.99	2.43
***Bs*CDO**	20 mM 2-AET 30 m	1.00	1.70	0.08
	20 mM 2-AET 5 h	1.00	38.49	1.19

Figure 3.13 shows the ^{1}H NMR spectra of reaction mixtures of *Mm*CDO and *Bs*CDO 30 min and 5 hours after incubation with 2-AET, along with those of 2-AET and HT standards. The spectrum of the 2-AET standard displays four signals: the signals centered at 3.11 and 2.76 ppm arise from the alpha and beta protons of 2-AET, respectively, while the set of signals at 3.36 and 3.01 are due to the analogous protons in cystamine, the disulfide of 2-AET. There is clearly a significant amount of disulfide in the 2-AET starting material, as well as a small amount of HT formed in the absence of enzyme. The ^{1}H-NMR spectra of both the *Mm*CDO and *Bs*CDO

reactions lose the 2-AET signal over time, but only a very small HT signal appears. Thus, the vast majority of 2-AET appears to undergo conversion to cystamine. Because a large amount of auto oxidation of 2-AET to HT was observed in UPLC-MS, a no-enzyme standard of 2-AET needs to be run under the same reaction conditions to determine if the formation of HT and cystamine are catalyzed by the enzymes. Table 3.3 shows the ratio of peak areas from the beta protons of each compound in the different samples, where the area of the 2-AET beta proton signal is set to one.

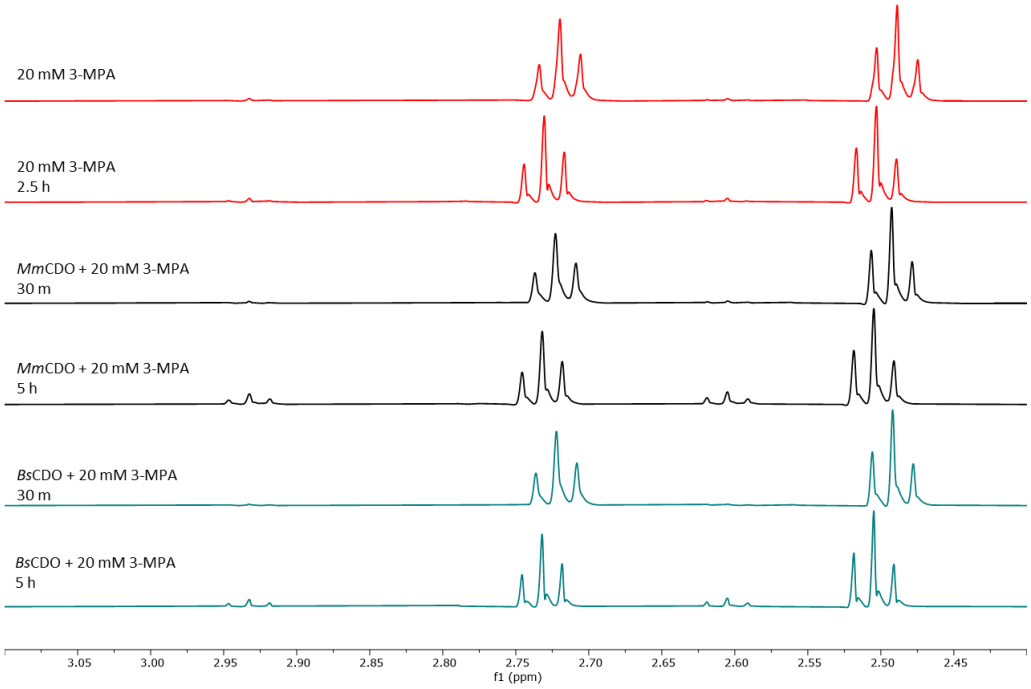


Figure 3.14 ^{1}H-NMR spectra collected at two time points after incubating MmCDO (black) and BsCDO (teal) with substrate analogue 3-MPA. Spectra of standard solutions of reactant 3-MPA are shown at the top.

Table 3.4 Relative peak areas of product and reactants from the NMR spectra in **Figure 3.14**

		3-MPA	**Disulfide**	**3-SPA**
Standard	20 mM 3-MPA	1.00	0.02	0.04
	20 mM 3-MPA 2.5 h	1.00	0.01	0.07
***Mm*CDO**	20 mM 3-MPA 30 m	1.00	0.03	0.03
	20 mM 3-MPA 5 h	1.00	0.01	0.14
***Bs*CDO**	20 mM 3-MPA 30 m	1.00	0.02	0.03
	20 mM 3-MPA 5 h	1.00	0.02	0.11

Figure 3.14 shows the ^{1}H NMR spectra of reaction mixtures of *Mm*CDO and *Bs*CDO 30 min and 5 hours after incubation with 3-MPA at, along with those of a 3-MPA standard and a no-enzyme 3-MPA control run for 2.5 hours. The weak pair of signals that grow in at 2.61 and 2.94 ppm in the ^{1}H NMR spectra of both the *Mm*CDO and *Bs*CDO reactions are tentatively attributed to the product 3-SPA. The presence of these signals in the 3-MPA standard and no-enzyme control is expected based on the observation of auto oxidation of 3-MPA to 3-SPA in the UPLC-MS assay. The disulfide exhibits signals between those of 3-MPA and 3-SPA; they are extremely weak in all spectra but can be seen in Figure A.3.5. Table 3.4 shows the ratio of peak areas from the beta protons of each compound in the different samples, where the area of the 3-MPA beta proton signal is set to one.

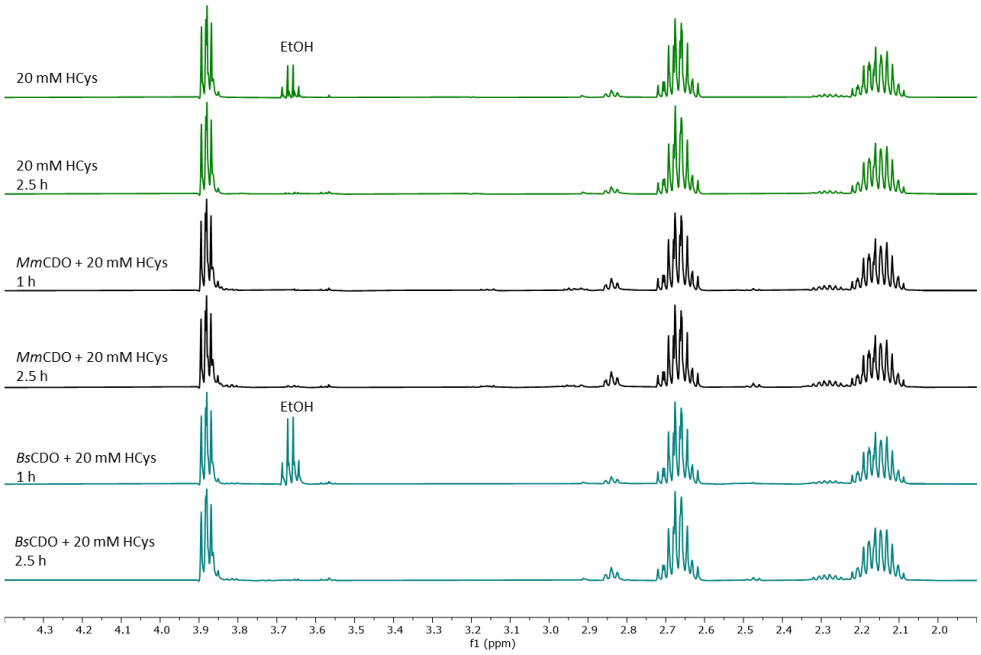


Figure 3.15 ^{1}H-NMR spectra collected at two time points after incubating *Mm*CDO (black) and *Bs*CDO (teal) with substrate analogue HCys. Spectra of standard solutions of reactant HCys are shown at the top.

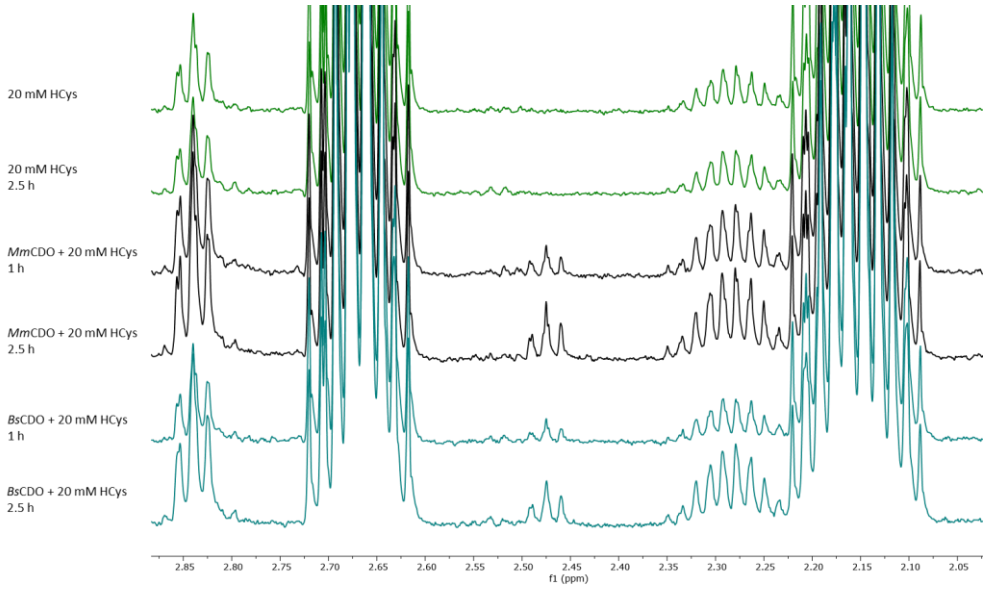


Figure 3.16 Closer view of the weak signals centered at 2.47 ppm likely arising from product HCSA that appear in ^{1}H-NMR spectra of reaction mixtures of both MmCDO and BsCDO after incubation with 20 mM HCys.

Table 3.5 Relative peak areas of product and reactants from the NMR spectra in **Figure 3.15 and 3.16**

		HCys	Disulfide	HCSA
Standard	20 mM Cys	1.00	0.07	0.00
	20 mM Cys 2.5 h	1.00	0.07	0.00
***Mm*CDO**	20 mM Cys 30 m	1.00	0.11	0.02
	20 mM Cys 5 h	1.00	0.14	0.03
***Bs*CDO**	20 mM Cys 30 m	1.00	0.07	0.01
	20 mM Cys 5 h	1.00	0.11	0.02

Figure 3.15 shows the ^{1}H NMR spectra of reaction mixtures of *Mm*CDO and *Bs*CDO 30 min and 2.5 hours after incubation with HCys, along with those of a HCys standard and a no-enzyme HCys control run for 2.5 hours. Two of these spectra contain a signal around 3.65 ppm from an ethanol impurity. The three signals around 2.15, 2.66, and 3.87 ppm can be attributed to the beta, gamma, and alpha protons of HCys, respectively. The weak pair of peaks at 2.88 and 2.84 ppm are tentatively assigned to the beta and gamma protons of the disulfide. Additional, very weak features appear near 2.47 ppmin the spectra of *Mm*CDO and *Bs*CDO incubated with HCys (Figure 3.16), which may be due to HCSA formation. Table 3.5 shows the ratio of peak areas from the beta protons of HCys and the disulfide, and the signal at 2.47 ppm in all samples, where the area of the HCys beta proton signal is set to one.

3.4 Discussion

Our MCD spectra provide evidence that both Fe(II)*Mm*CDO and Fe(II)*Bs*CDO are capable of binding Cys and 2-AET, but only Fe(II)*Bs*CDO is able to accommodate the bulkier HCys analogue under the experimental conditions used to collect these spectra. The S→Fe(II) CT features are blueshifted in the MCD spectra of substrate(analogue)-bound *Bs*CDO from their

counterparts in the spectra of the analogous *Mm*CDO complexes, implying a 6-coordinate, water-bound Fe(II) center in the bacterial enzyme that differs from the 5-coordinate complex observed in the crystal structures of substrate-bound *Mm*CDO. Similarly blueshifted S→Fe(II) CT features are displayed by the G82C *Bs*CDO variant in the presence of substrate (analogue), which forms the C-Y crosslink natively found in the *Mm*CDO active site. These results suggest that the active site of *Bs*CDO is more open or flexible in a way that allows a water molecule to bind even when crosslink is present.

Further support for a more open active site in *Bs*CDO compared to *Mm*CDO is provided by our EPR data. The S=5/2 signals in the EPR spectra of substrate(analogue)-Fe(III)*Bs*CDO complexes with and without azide display much greater inhomogeneous broadening than those of their Fe(III)*Mm*CDO counterparts, a sign of greater conformation freedom experienced by the former. The analogous G82C Fe(III)*Bs*CDO complexes exhibit S=5/2 EPR signals that more closely resemble those of Fe(III)*Mm*CDO, indicating that C-Y crosslink formation leads to a more constrained active site. However, 3-MPA-bound G82C Fe(III)*Bs*CDO incubated with azide displays a similarly broad S=5/2 EPR signal as its WT *Bs*CDO counterpart. Thus, while C-Y crosslink formation in response to the G82C substitution does reduce the active site flexibility in *Bs*CDO, additional differences in the active site region must exist between WT *Bs*CDO and WT *Mm*CDO, likely in the region nearest the carboxylate and amine moieties of substrate Cys. This hypothesis is consistent with the EPR spectra of the cyano/substrate(analogue)-bound *Mm*CDO, WT *Bs*CDO, and G82C *Bs*CDO complexes. This may, at least in part, be due to the positioning of R60 0.2 Å further away from the substrate carboxylate group in *Bs*CDO as opposed to *Mm*CDO, as can be observed in alignments of the two Cys-bound crystal structures, and which would lead to a weaker salt bridge between the Cys carboxylate and R60 guanidyl groups.

UPLC-MS and ^{1}H-NMR experiments have allowed us to assess the degree to which *Mm*CDO and *Bs*CDO are able to turn over substrate analogues. Despite *Mm*CDO's ability to

bind 2-AET in both the Fe(II) and Fe(III) states and without displacement by azide, the enzyme cannot oxidize this Cys analogue to any significant degree, even at high 2-AET concentrations. In comparison, *Bs*CDO binds 2-AET much more weakly, as evidenced by the fact that it can be readily displaced by azide. However, at high enough concentrations, *Bs*CDO can turn over 2-AET at a faster rate than *Mm*CDO. Thus, it appears that in *Mm*CDO, the lack of the substrate carboxylate moiety in 2-AET has a larger negative effect on the turnover rate than on the binding affinity, while in *Bs*CDO it negatively affects binding more so than it does turnover. The lower 2-AET binding affinity of *Bs*CDO compared to *Mm*CDO may, again, reflect a more open active site in the former enzyme. Because the *Mm*CDO active site is more constrained, the coordinated 2-AET substrate analogue may be forced to adopt conformations that are unfavorable for attack by the transiently formed Fe(III)-bound superoxide moiety, thus leading to a dramatic decrease in turnover rate.

Both *Mm*CDO and *Bs*CDO can bind 3-MPA even in the presence of an excess azide and display some turnover to 3-SPA. This result is quite surprising, considering that our EPR data provide evidence for a large conformational flexibility displayed by all 3-MPA-bound CDO complexes included in this study. Thus, we can conclude that while the Cys amine is not essential for substrate binding or catalytic turnover, it does increase the likelihood of correct substrate positioning and thus efficient binding and turnover.

As stated above, HCys appears to bind more tightly to the Fe(II) center in *Bs*CDO than in *Mm*CDO. The more open or flexible active site of *Bs*CDO is likely better suited for accommodating larger substrate analogues, such as HCys. Both *Mm*CDO and *Bs*CDO can turn over a very small amount of HCys. At high HCys concentrations, *Bs*CDO turns over HCys at a faster rate compared to *Mm*CDO. Because binding affinity should play a minimal role in the amount of product produced at high substrate concentrations (likely exceeding K_M), we can conclude that in the case of bulky substrate analogues, the increased flexibility of the *Bs*CDO active site is also responsible for the larger turnover rate.

From a comparison of the crystal structures of *Mm*CDO (4JTO) and *Bs*CDO (4QM9), the larger overall size of *Mm*CDO relative to *Bs*CDO is in part due to the approximately 30 additional residues at the C-terminal end of the former protein. These residues form a two-strand beta sheet positioned above the wall of the active site nearest the amine group of substrate Cys (Figure 3.16). In both *Mm*CDO and *Bs*CDO, this active site wall is composed only of a flexible amino acid backbone with no secondary structural features. The two-strand beta sheet in *Mm*CDO likely adds structural rigidity to this wall. Because *Bs*CDO lacks this beta sheet, the active site wall is granted much more flexibility in the region surrounding the amine moiety of coordinated substrate. Additionally, the extra beta sheet in *Mm*CDO closes off part of the substrate tunnel into the active site. This difference likely helps account for the wider substrate scope of *Bs*CDO.

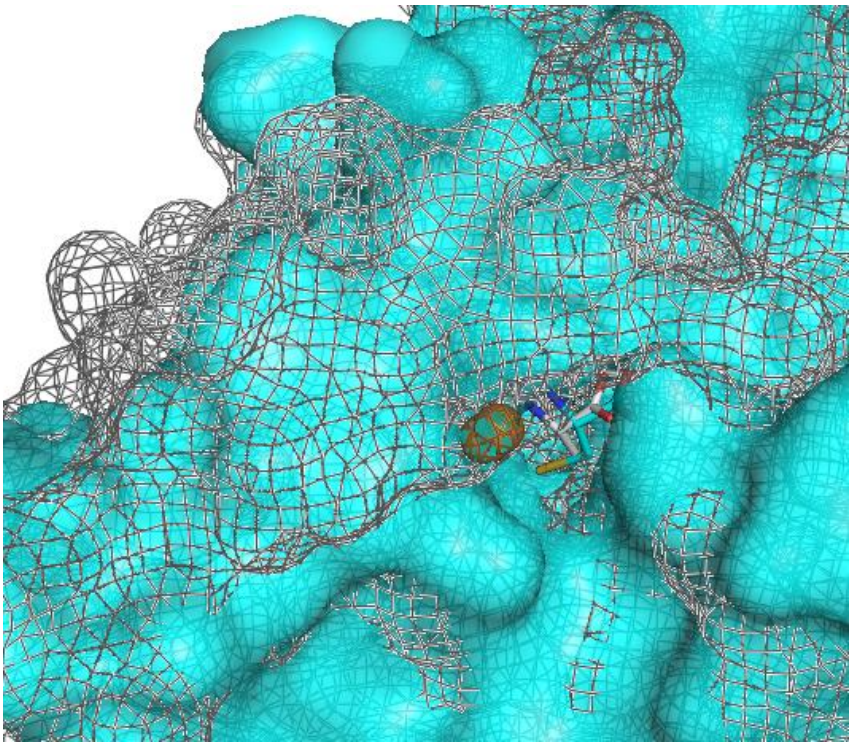


Figure 3.17 Overlay of the crystal structures of Cys-bound *Mm*CDO (gray mesh, PDB ID: 4JTO) and *Bs*CDO (cyan surface, PDB ID: 4QM9). The additional beta sheet above the otherwise flexible active site wall in *Mm*CDO likely gives added structural rigidity to the enzyme that increases specificity for native substrate Cys.

3.5 Conclusion

The goal of this study was to determine the origin of the extreme substrate specificity displayed by *Mm*CDO, as well as to identify any differences in substrate binding and positioning between *Mm*CDO and *Bs*CDO. To this end, we characterized various substrate (analogue)-bound complexes of WT *Mm*CDO, WT *Bs*CDO, and G82C *Bs*CDO. MCD, Abs, and EPR spectroscopies were used to evaluate the relative binding affinities of substrate (analogues), active site structures and conformational flexibility, and the role of the substrate Cys amine, thiol, and carboxylate moieties with regard to both binding and turnover. UPLC-MS and ^{1}H-NMR activity assays were performed to evaluate the extent of substrate (analogue) turnover at different time points and concentrations. We found that the *Bs*CDO active site is much more flexible and thus more accommodating of substrate analogues than the *Mm*CDO active site. The increased conformational freedom experienced by substrate (analogues) bound to the *Bs*CDO active site are not only due to the lack of a C-Y crosslink that is important for the function of *Mm*CDO, but may also be due to the absence of a small two-strand beta sheet above the flexible active site wall nearest the substrate amino group.

3.6 References

(1) Lombardini, J. B.; Singer, T. P.; Boyer, P. D. Cysteine Oxygenase II. Studies on the Mechanism of the Reaction with 18Oxygen. *J. Biol. Chem.* **1969**, *244* (5), 1172–1175. https://doi.org/https://doi.org/10.1016/S0021-9258(18)91825-9.

(2) Thomas Heafield, M.; Fearn, S.; Steventon, G. B.; Waring, R. H.; Williams, A. C.; Sturman, S. G. Plasma Cysteine and Sulphate Levels in Patients with Motor Neurone, Parkinson's and Alzheimer's Disease. *Neurosci. Lett.* **1990**, *110* (1–2), 216–220. https://doi.org/https://doi.org/10.1016/0304-3940(90)90814-P.

(3) Emery, P.; Salmon, M.; Bradley, H.; Wordsworth, P.; Tunn, E.; Bacon, P. A.; Waring, R. Genetically Determined Factors as Predictors of Radiological Change in Patients with

Early Symmetrical Arthritis. *Br. Med. J.* **1992**, *305* (6866), 1387–1389. https://doi.org/10.1136/bmj.305.6866.1387.

(4) Bradley, H.; Gough, A.; Sokhi, R. S.; Hassell, A.; Waring, R.; Emery, P. Sulfate Metabolism Is Abnormal in Patients with Rheumatoid Arthritis. Confirmation by in Vivo Biochemical Findings. *J. Rheumatol.* **1994**, *21* (7), 1192–1196. https://doi.org/10.1016/s0950-3579(05)80178-5.

(5) Davies, M. H.; Ngong, J. M.; Pean, A.; Vickers, C. R.; Waring, R. H.; Elias, E. Sulphoxidation and Sulphation Capacity in Patients with Primary Biliary Cirrhosis. *J. Hepatol.* **1995**, *22* (5), 551–560. https://doi.org/10.1016/0168-8278(95)80450-1.

(6) Dominy, J. E.; Hirschberger, L. L.; Coloso, R. M.; Stipanuk, M. H. In Vivo Regulation of Cysteine Dioxygenase via the Ubiquitin-26S Proteasome System. In *Advances in Experimental Medicine and Biology*; Springer New York, 2006; Vol. 583, pp 37–47. https://doi.org/10.1007/978-0-387-33504-9_4.

(7) Dominy, J. E.; Hwang, J.; Guo, S.; Hirschberger, L. L.; Zhang, S.; Stipanuk, M. H. Synthesis of Amino Acid Cofactor in Cysteine Dioxygenase Is Regulated by Substrate and Represents a Novel Post-Translational Regulation of Activity. *J. Biol. Chem.* **2008**, *283* (18), 12188–12201. https://doi.org/10.1074/jbc.M800044200.

(8) Li, W.; Pierce, B. S. Steady-State Substrate Specificity and O2-Coupling Efficiency of Mouse Cysteine Dioxygenase. *Arch. Biochem. Biophys.* **2015**, *565*, 49–56. https://doi.org/10.1016/J.ABB.2014.11.004.

(9) Li, W.; Blaesi, E. J.; Pecore, M. D.; Crowell, J. K.; Pierce, B. S. Second-Sphere Interactions between the C93−Y157 Cross-Link and the Substrate-Bound Fe Site Influence the O2 Coupling Efficiency in Mouse Cysteine Dioxygenase. *Biochemistry* **2013**, *52* (51), 9104–9119. https://doi.org/10.1021/bi4010232.

(10) Chai, S. C.; Jerkins, A. A.; Banik, J. J.; Shalev, I.; Pinkham, J. L.; Uden, P. C.; Maroney, M. J. Heterologous Expression, Purification, and Characterization of Recombinant Rat

Cysteine Dioxygenase*. **2004**. https://doi.org/10.1074/jbc.M413733200.

(11) Blaesi, E. J.; Gardner, J. D.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Characterization of the NO Adduct of Substrate-Bound Fe(II) Cysteine Dioxygenase: Insights into the Mechanism of O 2 Activation. *Biochemistry* **2013**, *52* (35), 6040–6051. https://doi.org/10.1021/bi400825c.

(12) Gardner, J. D.; Pierce, B. S.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Characterization of Substrate-Bound Mouse Cysteine Dioxygenase: Nature of the Ferrous and Ferric Cysteine Adducts and Mechanistic Implications. *Biochemistry* **2010**, *49* (29), 6033–6041. https://doi.org/10.1021/bi100189h.

(13) Tchesnokov, E. P.; Wilbanks, S. M.; Jameson, G. N. L. A Strongly Bound High-Spin Iron(II) Coordinates Cysteine and Homocysteine in Cysteine Dioxygenase. *Biochemistry* **2012**, *51* (1), 257–264. https://doi.org/10.1021/bi201597w.

(14) Driggers, C. M.; Cooley, R. B.; Sankaran, B.; Hirschberger, L. L.; Stipanuk, M. H.; Karplus, P. A. Cysteine Dioxygenase Structures from PH 4 to 9: Consistent Cys-Persulfenate Formation at Intermediate PH and a Cys-Bound Enzyme at Higher PH. *J. Mol. Biol.* **2013**, *425* (17), 3121–3136. https://doi.org/10.1016/j.jmb.2013.05.028.

(15) Simmons, C. R.; Liu, Q.; Huang, Q.; Hao, Q.; Begley, T. P.; Karplus, P. A.; Stipanuk, M. H. Crystal Structure of Mammalian Cysteine Dioxygenase: A NOVEL MONONUCLEAR IRON CENTER FOR CYSTEINE THIOL OXIDATION. *J. Biol. Chem.* **2006**, *281* (27), 18723–18733. https://doi.org/10.1074/JBC.M601555200.

(16) Driggers, C. M.; Kean, K. M.; Hirschberger, L. L.; Cooley, R. B.; Stipanuk, M. H.; Karplus, P. A. Structure-Based Insights into the Role of the Cys-Tyr Crosslink and Inhibitor Recognition by Mammalian Cysteine Dioxygenase Correspondence To. **2016**. https://doi.org/10.1016/j.jmb.2016.07.012.

(17) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of the H155A Variant of Cysteine Dioxygenase: Geometric and Electronic Consequences of

a Third-Sphere Amino Acid Substitution. *Biochemistry* **2015**, *54* (18), 47. https://doi.org/10.1021/acs.biochem.5b00171.

(18) Aloi, S.; Davies, C. G.; Karplus, P. A.; Wilbanks, S. M.; Jameson, G. N. L. Substrate Specificity in Thiol Dioxygenases. *Biochemistry* **2019**, *58* (19), 2398--2407. https://doi.org/10.1021/acs.biochem.9b00079.

(19) Dominy, J. E.; Simmons, C. R.; Karplus, P. A.; Gehring, A. M.; Stipanuk, M. H. Identification and Characterization of Bacterial Cysteine Dioxygenases: A New Route of Cysteine Degradation for Eubacteria. *J. Bacteriol.* **2006**, *188* (15), 5561–5569. https://doi.org/10.1128/JB.00291-06.

(20) Blaesi, E. J.; Fox, B. G.; Brunold, T. C. Spectroscopic and Computational Investigation of Iron(III) Cysteine Dioxygenase: Implications for the Nature of the Putative Superoxo-Fe(III) Intermediate. *Biochemistry* **2014**, *53* (36), 5759–5770. https://doi.org/10.1021/bi500767x.

(21) Fischer, D. S.; Pricet, D. C. A Simple Serum Iron Method Using the New Sensitive Chromogen Tripyridyl-s-Triazine. *Clin. Chem.* **1964**, *10* (1), 21–31. https://doi.org/https://doi.org/10.1093/clinchem/10.1.21.

(22) Simmons, C. R.; Hirschberger, L. L.; Machi, M. S.; Stipanuk, M. H. Expression, Purification, and Kinetic Characterization of Recombinant Rat Cysteine Dioxygenase, a Non-Heme Metalloenzyme Necessary for Regulation of Cellular Cysteine Levels. *Protein Expr. Purif.* **2006**, *47* (1), 74–81. https://doi.org/10.1016/J.PEP.2005.10.025.

(23) Siakkou, E.; Wilbanks, S. M.; Jameson, G. N. L. Simplified Cysteine Dioxygenase Activity Assay Allows Simultaneous Quantitation of Both Substrate and Product. *Anal. Biochem.* **2010**, *405* (1), 127–131. https://doi.org/10.1016/j.ab.2010.06.013.

(24) Davies, C. G.; Fellner, M.; Tchesnokov, E. P.; Wilbanks, S. M.; Jameson, G. N. L. The Cys-Tyr Cross-Link of Cysteine Dioxygenase Changes the Optimal PH of the Reaction without a Structural Change. *Biochemistry* **2014**, *53* (50), 7961--7968.

https://doi.org/10.1021/bi501277a.

Chapter 4

Bioinformatic Tools for Bioinorganic Chemists:

A How-To Guide

Chapter 4 : Bioinformatic Tools for Bioinorganic Chemists: A How-To Guide

Bioinformatics is a subdiscipline of biology and biochemistry that allows for the processing and analysis of large and complex sets of biological data. With modern advances in technology, we have been able to generate huge databases of DNA and protein sequences, structures, functions, metabolic pathways, taxonomic and evolutionary information, and much more. The trick is being able to efficiently parse through all this data and interpret it in a meaningful way. Luckily, there is an entire scientific field of biologists and computer scientists that specialize in creating user-friendly tools to do just that. By making the process approachable and easily accessible, they allow all of us—even those who don't consider themselves experts in the area—to enhance and contextualize our research findings within the larger framework on life on earth. In this chapter, I will provide a "how-to" user guide for the bioinformatic programs I utilized throughout my graduate research on CDO, with the aim of demonstrating how future students can apply similar methodologies to their own research.

4.1 Sequence Similarity Networks

Overview

Sequence similarity networks (SSNs) were developed by Dr. Patricia Babbitt's group at the University of Illinois-Urbana Champaign. They are a way of visualizing relationships among protein sequences. Proteins (or groupings of very similar proteins) are shown as colored dots, called "nodes", and lines, called "edges", are drawn between the nodes when they have a level of sequence similarity greater than or equal to some user defined stringency value. As the stringency value is increased, nodes will begin to separate into groupings called "clusters"; the higher the stringency value, the more similar the sequences within a single cluster are to one another. A helpful analogy is to imagine a phylogenetic tree turned end on; although SSNs aren't quite as rigorous as trees, they are faster to make and easier to interpret, which makes them great tools for those of us who aren't trained evolutionary biologists and just need a quick way to look at the relationships between members of a protein family.

There are two main applications of SSNs. The first is to provide an overview of the sequence-function relationships within a protein family, e.g., to understand the diversity within a group of proteins, or to discover new protein functions in a particular protein family. The second is to understand a particular protein's place within a larger set of sequences, e.g., a protein of known function may indicate the function of its entire cluster, or the pattern by which clusters fraction off from one another may indicate a common partial reaction but different specificity. SSNs have become an increasingly more popular tool over the past few years, and a simple literature database search for the term will give you plenty of examples of how they have been employed across a wide variety of research contexts.

To create an SSN, you will first need to visit the Enzyme Functionality Institute-Enzyme Similarity Tool (EFI-EST) website (https://efi.igb.illinois.edu/efi-est/) and input some information about the specific protein sequence or protein family you are interested in, as well as define some parameters on how you would like the data worked up. The program will calculate a set of initial results and email you when the job is completed. You will then need to review the initial results and make a few more decisions on how to narrow down the data set. The final network file can then be downloaded and visualized in the software Cytoscape (https://cytoscape.org/).

In the following paragraphs, I will walk through in more detail how to fill out the job input fields on the EFI-EST website and work up the results in Cystoscape, using the CDO family as an example. The CDO family is particularly large, so I include some helpful suggestions on working with such large file sizes. You can find a much more in-depth tutorial on the EFI-EST website (https://efi.igb.illinois.edu/efi-est/tutorial_startscreen.php) that explains every parameter and all the possible ways in which you could change them to suit your needs; here I synthesize the information down to only what is necessary to know for the most common applications.

Generating an SSN

When you open the EFI-EST website (https://efi.igb.illinois.edu/efi-est/), you will see five tabs, representing the five different methods by which you can generate an SSN.

(1) **“Sequence Blast”**: input the sequence of a particular protein of interest, e.g. the sequence of *Mm*CDO. A BLAST search is performed and the most related sequences pulled for inclusion in the network. The default number of nearest sequences is 1000, but up to 10,000 can be used. In a large family such as the CDO family, the nearest 10,000 sequences to *Mm*CDO are all eukaryotic CDOs, so this may not be the best option if you are trying to contextualize your protein of interest within a large or diverse family. However, it is useful if you are looking to determine an unknown enzyme functionality or don’t know which family the enzyme belongs to.

(2) “Families”: input the name of a protein family as defined by InterPro or Pfam. InterPro families typically include more sequences than the Pfam version, so it is typically best to use the InterPro-defined family. Both the InterPro (https://www.ebi.ac.uk/interpro/) and Pfam databases (http://pfam.xfam.org/) allow you to search for a family by protein sequence, keyword, or structural feature. For example, if you search the human CDO fasta sequence on InterPro, you find that it is a member of the cysteine dioxygenase type 1 family, the jellyroll family, and the cupin superfamily

(3) “FASTA”: Upload or input a list of sequences in FASTA format. This is a good option when you only want to align a particular subset of sequences in a family (perhaps as a result of a previously generated SSN), or when you are working with a set of sequences that don’t all belong to the same family (perhaps you’ve pulled sequences of enzymes that all appear to catalyze the same reaction but are evolutionarily distinct). Do not attempt to use this input option to manually decrease the size of your input sequence list—if needed, there are options to randomly eliminate sequences without skewing your data (read more below).

(4) "Accession IDs": this option is essentially the same as the "FASTA" option, but instead you can input UniProt IDs for individual proteins instead of sequences in FASTA format.

(5) "SSN Utilities": this option is new since I last worked with SSNs but appears to be a way to analyze an SSN you have already generated. Much of the analysis I had to perform manually in Matlab can now be calculated for you with a user-friendly interface. These options are discussed in further detail below.

Input Parameters

Once you have chosen the most suitable method (1-4) for generating your initial SSN and provided the required sequences or family names, you will need to define a few other input parameters before submitting the job for analysis.

(1) "BLAST Retrieval Options": these settings help determine the size of the sequence database that the calculations will be performed on.

- **UniProt BLAST query e-value**: sets a cut off for how similar a protein must be to your input protein in order to be included. A good starting value is 5, and if needed you can adjust this number and resubmit after seeing the initial calculation results.
- **Maximum number of sequences retrieved**: default is 1,000, maximum is 10,000. If you have some sense of the size of your protein family you can change this to include a larger or smaller subset of the family. You can also adjust this number and resubmit after seeing the variation in proteins in the resulting SSN.
- **Sequence database**: for very large families (tens of thousands of sequences and above), you can minimize the duration of your calculations and size of your resulting files by using the "UniRef90" or "UniRef50" options. This will cluster sequences with ≥ 90% or

≥ 50% sequence similarity, respectively, before performing the all-by-all sequence BLAST. However, there are similar options for reducing file size that can be applied after the calculations are complete, so I recommend only using UniRef90 or UniRef50 for very large families or superfamilies where the calculation runtime becomes prohibitive.

(2) **"Fragment Option"**: protein sequence databases always contain a significant number of protein fragment sequences. I always check this box—it will exclude anything too short from your input file. There will be other options later to further exclude anything that appears to be a fragment**.**

(3) **"Filter by Taxonomy"**: include only proteins from a selected taxonomic groups.

(4) **"SSN Edge Calculation Option"**: sets the initial stringency level for how similar two proteins must be for an edge to be draw between them. A good starting value is 5. Ideally, your initial SSN will be one giant cluster at this starting value, and you will increase it in the Cytoscape software and observe clustering patterns. If your initial SSN already has several smaller clusters, you will need to repeat the calculation with a smaller (less stringent) value. If your initial SSN remains as one large cluster even as you increase the stringency value a few times, you could decrease your file size by recalculating with a larger initial value.

(5) **"Protein Family Addition Options"**: add on an entire family to your list of input sequences. You can minimize the size of the addition by using UniRef90

or UniRef50, or by using the “fraction” feature to only include every N^{th} sequence in the family.

(6) “Family Domain Boundary Option”: include sequences from other families, but trim them to include only the domain defined by the input family.

Name your job and submit, and you will receive an email when the initial calculation is complete. For reference, a family of ~12,000 proteins took about 24 hours to run.

Analyzing Initial Results

Follow the link in the email to view your results. The “Dataset Summary” tab details the input parameters used to generate the dataset. The “Dataset Analysis” page provides several plots which give insight into the proteins included in the initial dataset, and you will need to look through them to determine suitable sequence length cutoffs and an alignment score threshold to apply to the dataset before visualizing it as an SSN.

(1) Length Histogram: plots the length of a protein sequence vs the number of sequences with that length. Proteins will tend to cluster into groupings based on the number of domains they have. Protein fragments will appear at shorter legnths. Use this plot to determine a minimum and maximum length of sequences to include in your final dataset. You will likely only want to include those proteins which have the same number of domains as your protein(s) of interest.

(2) Number of Edges Histogram: plots alignment score (how similar two proteins are) vs. the number of protein pairs with that score. This plot gives insight into the functional diversity of a dataset. If the histogram is clustered toward lower alignment scores, you are dealing with a divergent superfamily. If the histogram is clustered toward higher alignment scores, you have a

highly conserved isofunctional family. Depending on what information you are looking to get out of your SSN, you may want to repeat the calculation with a larger or smaller set of sequences to get more meaningful information.

(3) Alignment Length Quartile Plot: plots the alignment score (how similar two sequences are) vs the alignment length (how many residues in a pair of sequences align with one another). This provides another method for identifying single and multi-domain proteins: each grouping of same domain number proteins will appear as a roughly horizontal line (approximately the same length, but variable alignment scores). Use this plot in conjunction with the ”Length Histogram” to select appropriate minimum and maximum length cutoffs for your final dataset.

(4) “Percent Identity Quartile Plot”: plots alignment sore vs percent identity. Percent identity is a measure of how many aligned residues are exactly identical to one another, whereas alignment score accounts for residues that are similar to one another. Each positively sloped line on the plot represents a group of same domain number proteins (i.e., alignment score and percent identity are positively correlated; once another domain is added, alignment score continues to increase but there is an initial drop in percent identity). We can cut anything with very low alignment scores out of our final network, as these relationships are very distant and likely not meaningful. A good rule of thumb is to find the alignment score that corresponds to 35 percent identity on the plot and use this as your alignment score threshold. The value you choose will be used as the lower limit for which an edge is drawn in your

SSN. Similar to the “SSN Edge Calculation Option” parameter, you can adjust this value depending on the amount of clustering you see in your initial SSN.

Finalizing the Dataset

Move to the “SSN Finalization” tab and input an appropriate “Alignment Score Threshold” value (usually the alignment score corresponding to 35% identity in the Percent Identity Quartile Plot) as well as a Minimum Sequence Length Restriction (use the Length Histogram and Alignment Length Quartile Plot and pick a minimum length that will exclude any fragments; you could also exclude single domain proteins, e.g., if you are only interested in multi-domain proteins) and a Maximum Sequence Length (leave blank if you want to include multi-domain proteins, or select a value that will exclude them). Name your network and submit. You will receive another email when the network finalization is complete (typically only takes an hour or so).

When you receive the email, follow the link to download the network file. You will have the option to download either the full network or a “representative node network,” in which each node of in the network represents a collection of proteins that have all have a certain percent identity to one another. E.g., in an 80% representative node network, each node is a group of proteins that are 80% or more identical to one another. You will want to use the largest file your computer memory can handle. For reference, a 4 GB RAM can handle an SSN with ~500,000 edges and a 64 GB RAM can handle an SSN with ~5,000,000 edges.

Filtering and Styling the Network

Make sure you have download the free Cytoscape software, as well as Cytoscape’s “yFiles Layout Algorithms” (https://apps.cytoscape.org/apps/yfileslayoutalgorithms). Open the network file in Cytoscape and apply the “organic layout” to the network by selecting Layout → yFiles → Organic. Ideally, your network should appear as one large cluster of nodes. If there are already several clusters, you may want to regenerate your network with new parameters so that

you don't miss information on the initial clustering patterns. First, try decreasing the "Alignment Score Threshold" in the SSN finalization and downloading the newly finalized network file. If you still are seeing too many initial clusters, redo you initial calculation with the "SSN Edge Calculation Option" set to a smaller value.

You can try using other layout styles, but I have found that organic works very well. You can also use the panel on the left to change styles, colors, node size, etc. Try to change as much of these style preferences now as possible, as they will carry over to any new versions of the network you create as you change the stringency value. The table near the bottom of the screen will allow you to view nodes and edges and their characteristics as given by the database from which they were pulled. Be aware—not all proteins in a database are verified to have the function listed. Entries labeled as "Swiss-Prot" have been reviewed and are confirmed to have the listed activity. Entries labeled as "TrEMBL" are computationally predicted to have the listed function but are unreviewed. It can helpful to identify and label the Swiss-Prot verified proteins in your network; in a functionally diverse network, the positioning of these proteins may help you to predict the function of the entire clusters. There are many other traits you could use to style your network: organism/taxonomic category, length, number of domains, possession of a specific structural feature, etc. Some of these traits are already listed in your node table because they were included in the protein database. Others you may need to add manually. The easiest way to add and label by a manual category is as follows:

(1) In Cytoscape, top bar
    a. "Select", "nodes", "from ID list file", choose your file name.
(2) In Cytoscape, table
    a. Plus sign "create new column" and name it, "integer", click on one square in the new column and type "1", right click the "1" and select "apply to selected nodes"
(3) In Cystoscape, control panel

a. Style, e.g. “Shape”, Column: “name of column you created”, Mapping type: “Discrete Mapping”, next to “1” choose the shape you want

Next, progressively increase the stringency value and recalculate the layout of your network. Follow these steps:

(1) Open the network in Cytoscape and make a copy of the network: File → New → Network → Clone Current Network

(2) Open the Select tab of the Control Panel

(3) Delete previous filters by clicking the "X" to the left of the filter criteria

(4) Add a new filter by clicking the "plus sign" button, then select Column Filter

(5) Configure filter by selecting "Edge: alignment_score" from the dropdown menu

(6) Set the upper and lower limits of the filter by moving the left and right arrows along the scroll bar - note: you cannot get any less stringent than the e-value at which the network was originally calculated (i.e., leave the lower limit), but you can achieve greater stringency by increasing the upper limit.

(7) If the Apply Automatically option is selected, the edges that do not satisfy the e-value stringency will turn red.

(8) Delete red edges: Edit → Delete Selected Nodes and Edges

(9) Re-do the layout: Layout → yFiles → Organic

The “memory” symbol in the bottom right corner will indicate with a green, yellow, or red circle the amount of computer memory being used by the network file. You will also notice things running slower as you make the file bigger. To help free up space, I recommend saving a new copy of the Cytoscape file every few iterations as the memory symbol turns yellow or red. Open the new file and delete all the previous iterations besides the latest ones, then continue working.

Analyzing the Network

There is a lot of useful information you can gain from analyzing your SSN. A lot of the analysis I did for the SSN in chapter 2 needed to be manually performed by running my own large sequence alignments, referencing online databases, and writing Matlab programs. These methods will be explained here. Recently, however, the EFI-EST added an “SSN Utilities” tool to their webpage that will do much of the same analysis for you.

As a first step, try to see what general information and trends you pick up on as you filter your network. If you haven’t already, label by color, shape, etc. any traits that may be interesting and label proteins with Swiss-Prot verified functions. Do all the proteins in a cluster seem to have some unifying trait? As you increase the stringency value, do they form even smaller clusters based on another trait? Which closer are grouped closer or farther from one another? Does this make sense in the context of function or evolutionary history? Are there any patterns in which clusters conserve a specific domain or structural feature?

Multiple Sequence Alignments

A helpful next step can be to determine the consensus sequences of different clusters of proteins. This is now a feature that can be performed by uploading your finalized SSN file to the EFI-EST website, but I did this manually. By performing them myself, I was also able to employ a technique that allowed me to directly compare consensus sequences of clusters to one another.

If you are only aligning two sequences or a handful of sequences, you can use a pairwise alignment (https://www.ebi.ac.uk/Tools/psa/) or simple multiple sequence alignment (MSA) webserver such as Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), respectively. For alignments of large groups of sequences, however, you need to use a more robust alignment program. I recommend using MUSCLE. There is a webserver version of the program that can handle up to 500 sequences (https://www.ebi.ac.uk/Tools/msa/muscle/). For larger alignments, you will need to download the software to your desktop

(https://www.drive5.com/muscle/). The desktop version of MUSCLE can be run using the “align” command, which works well for groupings of up to a few thousand sequences. For larger groupings, you will need to use the “super5” command, which drastically shortens the run time by first making smaller alignments and then joining them together.

Create a MUSCLE folder on your computer where you will store all your files. Create a shortcut to a command prompt and set the starting location as your muscle folder. You will only need to know a few basic command lines.

(1) In your SSN, create column and label nodes by which cluster they belong to (I give each cluster a “Cluster #”)
(2) Download Cytoscape Table
(3) Sort by Cluster # → delete all others
(4) Format and copy/paste “List of IDs” into “Uniprot Retrieve/ID Mapping” site, convert to UniprotKB (https://www.uniprot.org/uploadlists/)
(5) Download FASTA uncompressed
(6) Rename to something you’ll remember and move to MUSCLE folder
(7) Command prompt: “muscle -in sequences.fasta -out aligned.fasta” or “muscle -super5 sequences.fasta -output aligned.fasta”

    where “sequences.fasta” is the name of the fasta file containing your unaligned sequences and “aligned.fasta” is the name of the file your aligned sequences will be saved as
(8) Open output .fasta file in Snapgene (https://www.snapgene.com/snapgene-viewer)
(9) Right panel lets you adjust consensus sequence parameters
(10) Right click “consensus sequence” to download in Snapgene or Fasta format.

Determining Putative C-Y Crosslinking from MSA

To determine whether each CDO in the SSN shown in chapter 2 was capable of crosslink formation, I first generated an MSA of all ~13,000 proteins in the CDO family using the

super5 MUSCLE algorithm as described above. I then opened the resulting fasta file in Snapgene and found residue C93 and residue Y157 in the *Mm*CDO sequence. I copied and pasted the list of which residue was at this position in all sequences into a text file ("AllCDOs Aligned Pos Y157.txt" and "AllCDOs Aligned Pos C93.txt"). The fasta file of all aligned CDOs was saved as "AllCDOsAlignedText.txt". The Matlab codes given in the Appendix as Matlab Code A.4.1 and Matlab Code A.4.2 read these files to determine which sequences contain a "C" and a "Y" at the positions aligned to C993 and Y157, respectively, in *Mm*CDO. The generated lists of sequences were then uploaded into the SSN file and the method used in the "Filtering and Styling the Network" section above was used to label any nodes that contained both an analogous C93 and Y157 residue.

Although I didn't end up using this data in chapter 2 or the resulting publication, Matlab Code A.4.4 allows you to identify in non-crosslinked CDOs which residue typically appears instead at position 93 or 157. This code could be adapted for similar analysis of two residues that appear to be conserved together.

Aligning Multiple Consensus Sequences

While analyzing my SSN, I wanted to be able to directly compare consensus sequences of individual clusters or smaller subsets of proteins (e.g., compare the consensus sequence of all non-crosslinked bacterial CDOs to the consensus sequences of all putatively crosslinked CDOs). This allowed me to look for patterns in conservation of specific residues across different categories of CDO enzymes. In order to do this, I first generated the MSA of all ~13,000 proteins in the CDO family using the super5 MUSCLE algorithm as described above. I think created a text file containing the identifiers of all proteins belonging to the specific category I was interested in. These could be searched for and downloaded from the node table in the Cytoscape SSN file. I then used the ran the code given in Matlab Code A.4.3 to generate a list in FASTA format of all the sequences in this subset, where each sequence was now the length of the longest CDO in the entire CDO family because dashes had been added where needed.

The output could then be run again through the MUSCLE software to create a consensus sequence specifically for this subset of proteins. Using this methodology ensured that all the consensus sequences generated would be the same length and residues would align with one another even when they were aligned with MUSCLE as different subsets.

Taxonomy

. The EFI-EST website now has a Taxonomy Tool (https://efi.igb.illinois.edu/taxonomy/) that does a similar analysis for you and visualizes the information is a much more intuitive way. This was unavailable while I was analyzing my network, so I used a different methodology which I describe briefly here. The online protein databases you download your SSN dataset from contain some taxonomic information about each protein in your network, but there is usually a fair amount of missing information that makes it difficult to easily analyze. Therefore, I used the NIH Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy). On the “Name/ID Status” search page, I uploaded a text file of the identifiers for whichever subgroup of CDOs I was interested in (e.g., all non-crosslinked bacterial CDOs). The webpage returns a list of numeric taxonomy identifiers, each of which corresponds to a specific kingdom/phylum/class/etc. I then used the codes given in Matlab Code A.4.5, A.4.6, and A.4.7 to format and sort these taxonomy IDs as well as create lists of the unique members of each taxonomy category (e.g., all the unique phyla) listed in order of frequency.

Chapter 5

Conclusions and Outlook

Chapter 5 : Conclusions and Outlook

The work presented in this thesis describes the first in-depth characterization of the CDO enzyme from *Bacillus subtilis*. By performing our investigation of *Bs*CDO in direct comparison with the more well-studied eukaryotic version of the enzyme (*Mm*CDO), specific structural features of CDO and their role in determining the substrate scope and reactivity of the enzyme could be better understood. In particular, we were able to create a G82C variant of the naturally non-crosslinked *Bs*CDO and demonstrate that a single DNA point mutation is all that is required to enable crosslink formation in the enzyme active site. The *Bs*CDO C-Y crosslink affects the kinetics of Cys turnover in much the same way as the natural crosslink does in *Mm*CDO, as introduction of a C residue at position 82 slows the turnover rate, but an increase in percent crosslinking between C82 and Y141 brings the rate closer to WT *Bs*CDO levels. Additionally, by comparing the ability of WT *Mm*CDO, WT *Bs*CDO, and G82C *Bs*CDO to bind and turn over various analogues of native substrate Cys, we discovered that WT *Bs*CDO has a much more flexible active site that is more accommodating of substrate analogues. This difference in flexibility is only partially removed by addition of the crosslink via G82C mutation. We propose that the presence of a small, exterior C-terminal beta sheet in WT *Mm*CDO adds additional rigidity to an otherwise flexible wall of the CDO active site, thus increasing specificity of the enzyme for native substrate Cys.

Studying eukaryotic CDO alongside a bacterial CDO homologue has proven to be a productive methodology for better understanding the role of unique structural features and the origin of regulatory mechanisms. This methodology can be expanded to address several lingering questions, outlined in the following project proposals:

(1) **Characterization of Putatively Crosslinked Bacterial CDOs.** In our bioinformatic analysis of the CDO family, we discovered a significant number of bacterial CDOs that appear capable of crosslink formation on the basis of their containing C and Y residues aligned with C93 and Y157 in *Mm*CDO. These CDOs are distributed across

6 main taxonomic groupings of bacteria, as seen in the SSN in Figure 2.7: actinobacteria, flavobacteria, burkholderia, streptomyces, xanthomondales, and vibronales/planctomycetes. Using fDETECT, we selected the most promising candidate for expression, purification, and crystallization from each group. The codon-optimized genes for three of these six CDOs have been purchased cloned into the pET-14b vector and expression in *E. coli* is currently underway. To confirm that these bacterial CDOs are indeed capable of crosslink formation, we can repeat many of the same experiments presented from the publication in Chapter 2 of this thesis (MS, substrate/CN^- EPR). The effect of the crosslink on the kinetics of Cys turnover in these enzymes should also be evaluated. Chapter 2 details UPLC-MS and Ellman's reagent-based assays for determining the Michaelis-Menten kinetic parameters for CDO, but the NMR methodology described in Chapter 3 could also be expanded in a similar way to that used by Jameson and would likely be easier to perform. Crystallization of these putatively crosslinked bacterial CDOs should also be pursued to compare the positioning of the crosslink in eukaryotic vs bacterial CDOs.

(2) **Mechanism of Substrate Analogue Binding in CDOs.** Previous work by Dr. Stephanie Dillon in Chapter 3 of her dissertation, which was expanded upon in Chapter 3 of this work, used MCD spectroscopy of substrate analogue-bound *Mm*CDO with and without addition of excess azide and computational modeling to investigate putative reaction intermediates in the CDO catalytic cycle. I have collected some MCD spectra of substrate analogue and azide bound *Bs*CDO (see Appendix Figures A.5.1 and A.5.2), but these samples had very low Fe(III) loading and thus could not be conclusively interpreted. The cyanide and substrate(analogue)-bound CDO complexes in Chapter 3 are incredibly interesting and somewhat unusual. Further study into the origin of their shapes would also likely give insight into the binding modes of the analogues. EPR data of the analogous

cyanide and substrate (analogue) bound *Mm*CDO will need to be collected. Stephanie collected an MCD spectrum of Cys and CN^- -bound *Mm*CDO, but analogous spectra will need to be taken to see if this is an interesting route to pursue. Because WT *Mm*CDO, WT *Bs*CDO, and G82C *Bs*CDO demonstrate differences in their ability to bind various substrate analogues, azide MCD spectroscopy, cyanide EPR and MCD spectroscopy, and computational modeling could be used to investigate the mechanism by which the different substrate analogues bind in the active sites of these CDOs, and thus deduce the role the C-Y crosslink plays in substrate recognition and binding.

(3) Movement of Putatively Flexible CDO Active Site Wall. A study should be conducted to confirm whether the C-terminal beta sheet identified in Chapter 3 does indeed provide an extra level of rigidity to the *Mm*CDO active site over that of *Bs*CDO. This could be done by attaching a fluorescent or spin label probe to this region of both enzymes and observing the level of movement upon binding of a bulky substrate analogue such as HCys. If a difference in movement of this region is observed, a variant of *Bs*CDO could be made that attaches the sequence of the beta sheet to it's C-terminal end. Movement of the same region in the variant could then be measured, as well as EPR experiments analogous to those performed in Chapter three on the substrate (analogue) and azide bound Fe(III)CDO complexes.

APPENDIX

Appendix

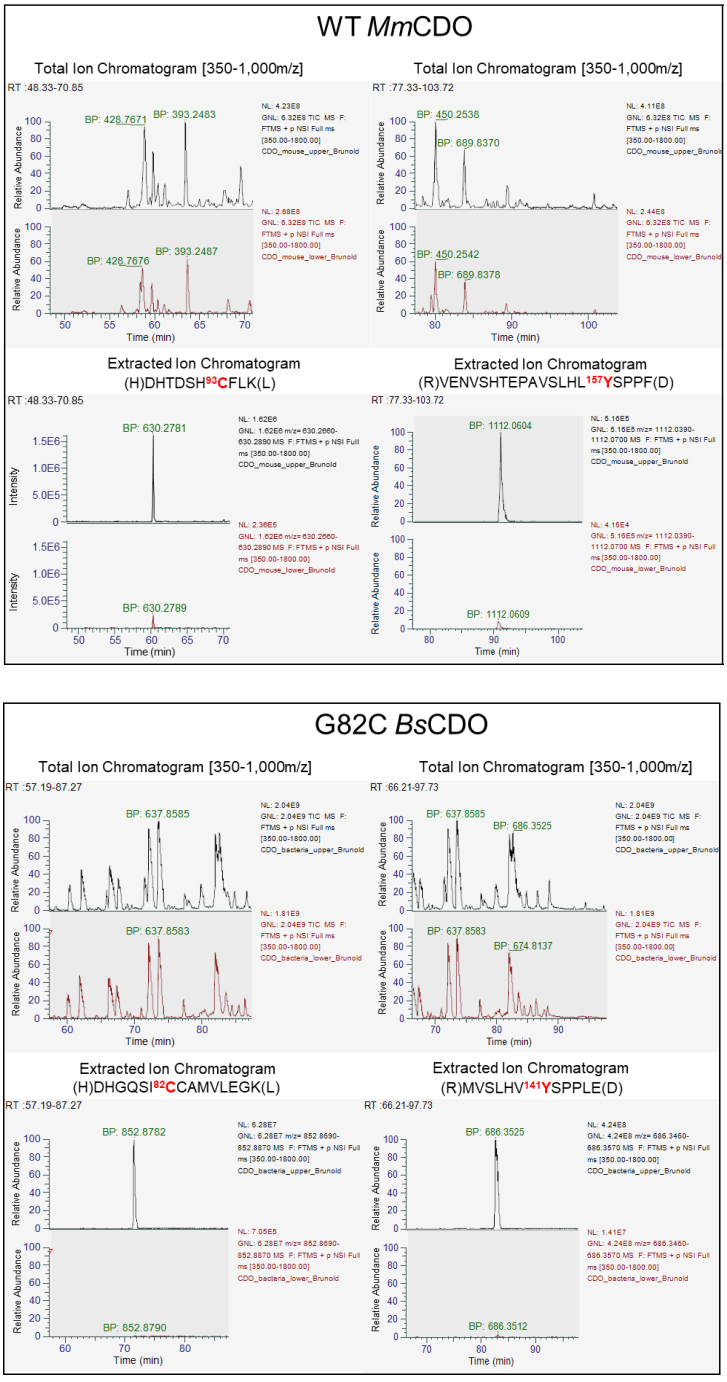


Figure A.2.1 Total ion chromatograms and extracted ion chromatograms of dual-digested WT MmCDO (top) and G82C BsCDO (bottom). For WT MmCDO, extracted ion chromatograms show the relative intensity of the (H)DHTDSH^{93}**C**FLK(L) fragment (left) and the (R)VENVSHTEPAVSLHL^{157}**Y**SPPF(D) fragment (right) in the mass spectra of the protein extracted from the upper and lower SDS-PAGE gel bands. For G82C BsCDO, extracted ion chromatograms show the relative intensity of the

(H)DHGQSI[82]**C**CAMVLEGK(L) fragment (left) and the (R)MVSLHV[141]**Y**SPPLE(D) fragment (right) in the mass spectra of the protein extracted from the upper and lower SDS-PAGE gel bands.

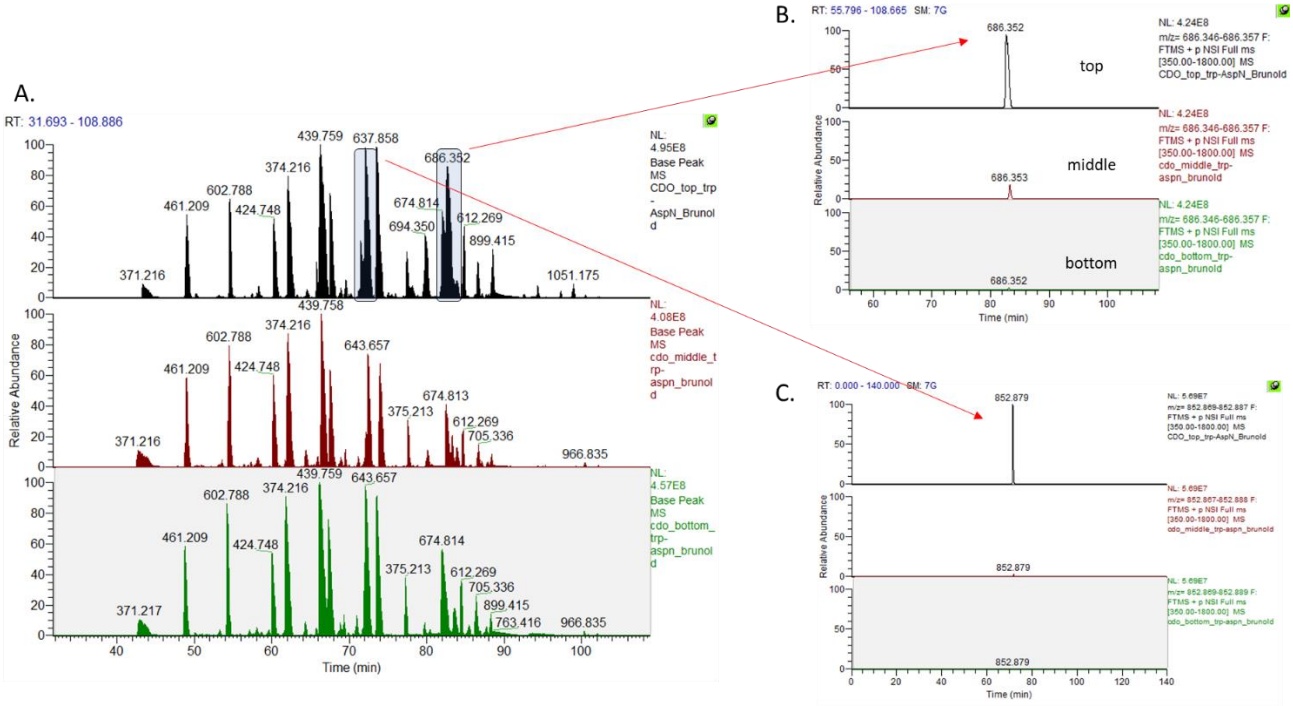


Figure A.2.2 *(A)* Total ion chromatograms and *(B,C)* extracted ion chromatograms of dual digested G82C *Bs*CDO from the top, middle, and bottom bands of the SDS-PAGE gel. *(B)* shows the relative intensities of the (H)DHGQSI[82]**C**CAMVLEGK(L) fragment and *(C)* shows the relative intensities of the (R)MVSLHV[141]**Y**SPPLE(D) fragment.

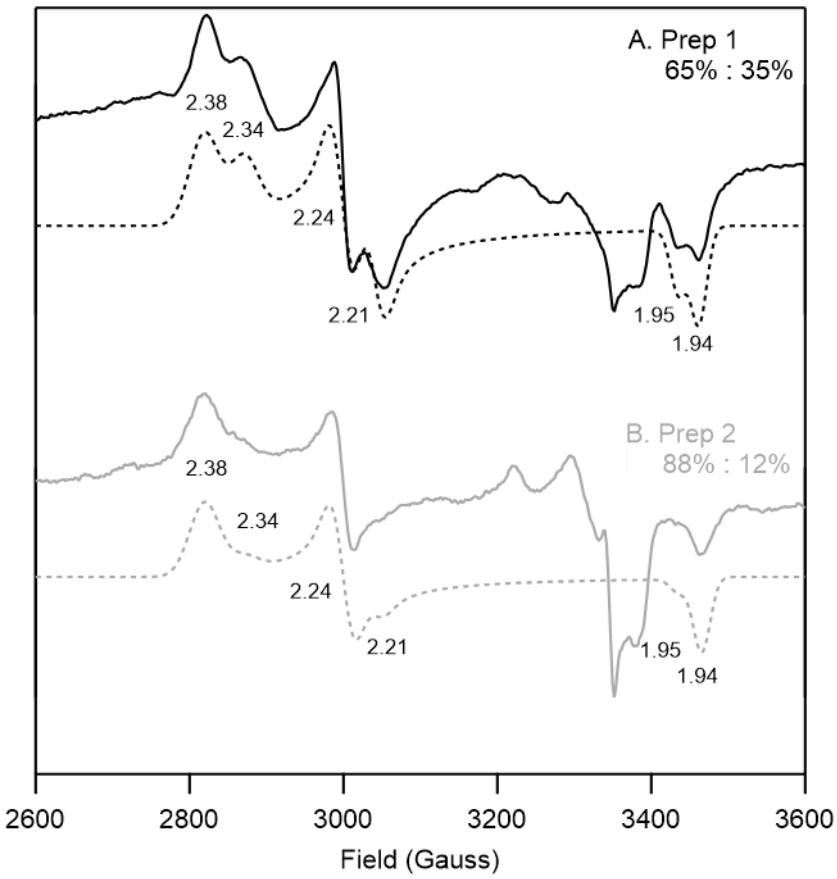


Figure A.2.3 EPR spectra of the (Cys/CN^-)-Fe(III) adducts of *(A)* WT *Mm*CDO Prep 1 and *(B)* WT *Mm*CDO Prep 2. Spectra are shown as solid lines and fits as dashed lines. The ratios of crosslinked to non-crosslinked Fe(III)-bound active sites as determined via fitting of these spectra are included. Features between ~3250 and 3400 Gauss are due to a contaminant in the instrument cavity.

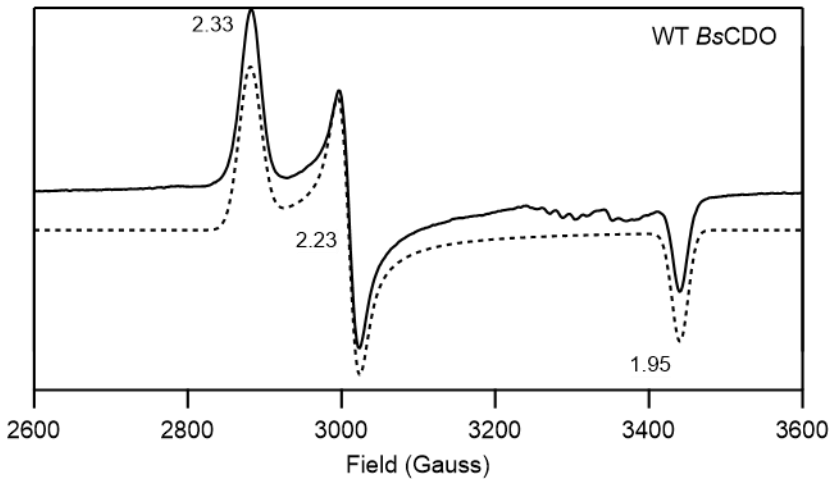


Figure A.2.4. EPR spectrum of the (Cys/CN^-)-Fe(III) adduct of WT *Bs*CDO. The spectrum is shown as solid line and the fit as dashed line. Features between ~3250 and 3400 Gauss are due to a contaminant in the instrument cavity.

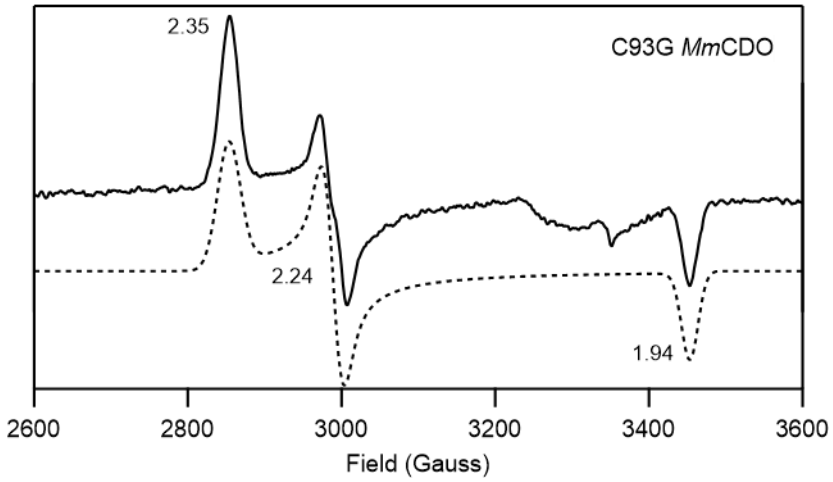


Figure A.2.5. EPR spectrum of the (Cys/CN^-)-Fe(III) adduct of C93G *Mm*CDO. The spectrum is shown as solid line and the fit as dashed line. Features between ~3250 and 3400 Gauss are due to a contaminant in the instrument cavity.

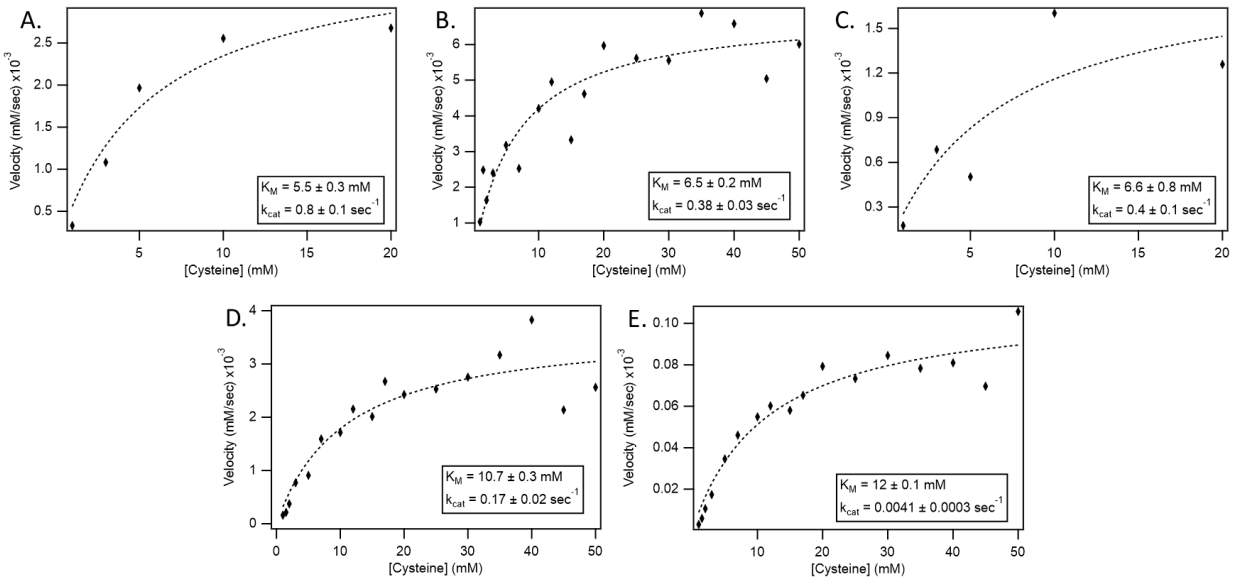


Figure A.2.6. Michaelis-Menten fits of kinetic assay data for *(A)* WT *Mm*CDO with Cys as measured via Ellman's reagent assay, *(B)* WT *Bs*CDO with Cys as measured via UPLC-MS assay, *(C)* WT *Bs*CDO with Cys as measured via Ellman's reagent assay, *(D)* high iron G82C *Bs*CDO Prep 1 with Cys as measured via UPLC-MS assay, and *(E)* low iron G82C *Bs*CDO Prep 2 with Cys as measured via UPLC-MS assay.

Table A.2.1. Extended fitting parameters for the EPR spectra of the (Cys/CN^-)-Fe(III) adducts of WT *Mm*CDO Preps 1 and 2 **(Figure A.2.3)** and G82C *Bs*CDO Preps 1 and 4 (**Figure 4)**

Sample	System	System Weight	g_1 line width (MHz)	g_2 line width (MHz)	g_3 line width (MHz)
WT *Mm*CDO Prep 1	Crosslinked	0.65	8	8	18
	Non-crosslinked	0.35	8	8	18
WT *Mm*CDO Prep 2	Crosslinked	0.88	8	8	18
	Non-crosslinked	0.12	8	8	18
G82C *Bs*CDO High Iron Prep 1	Crosslinked	0.79	8	8	19
	Non-crosslinked	0.21	8	8	19
G82C *Bs*CDO Low Iron Prep 4	Crosslinked	0.67	8	8	19
	Non-crosslinked	0.33	8	8	19

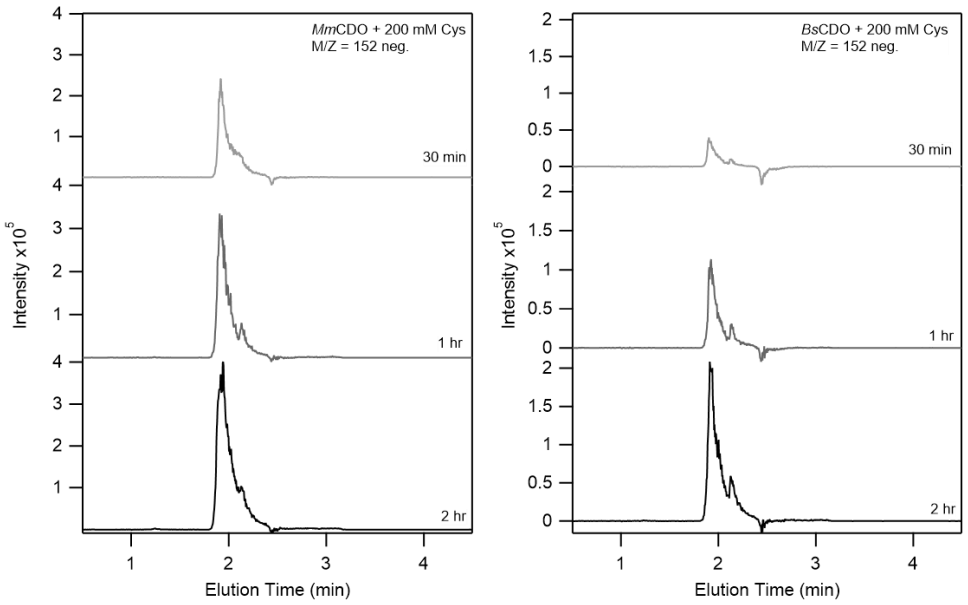


Figure A.3.1 Mass spectra at the M/Z corresponding to CSA of quenched reaction mixtures of *Mm*CDO and *Bs*CDO with Cys at three time points. A 2 hour no-enzyme Cys control has been background subtracted from each spectrum.

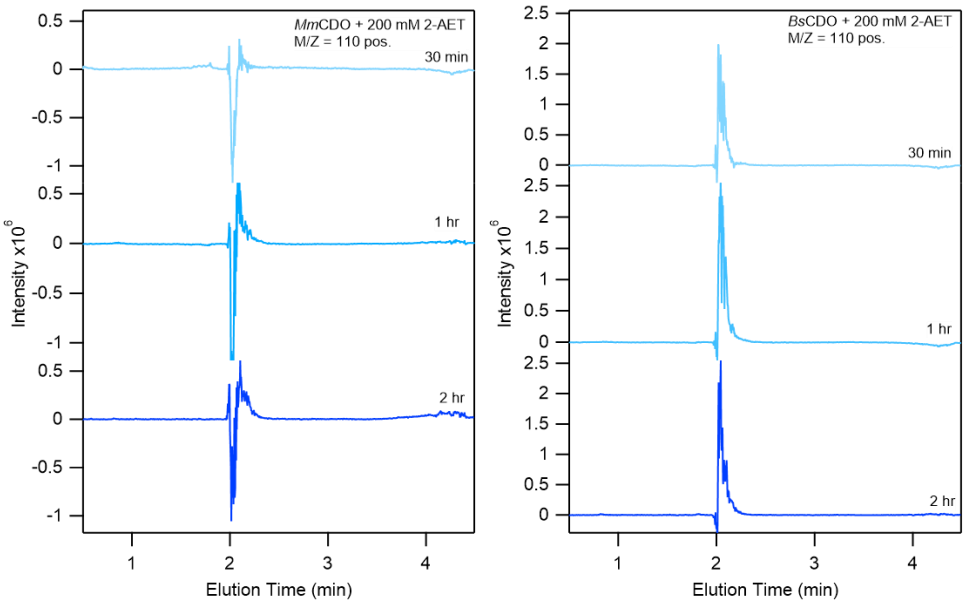


Figure A.3.2 Mass spectra at the M/Z corresponding to HT of quenched reaction mixtures of *Mm*CDO and *Bs*CDO with 2-AET at three time points. A 2 hour no-enzyme 2-AET control has been background subtracted from each spectrum.

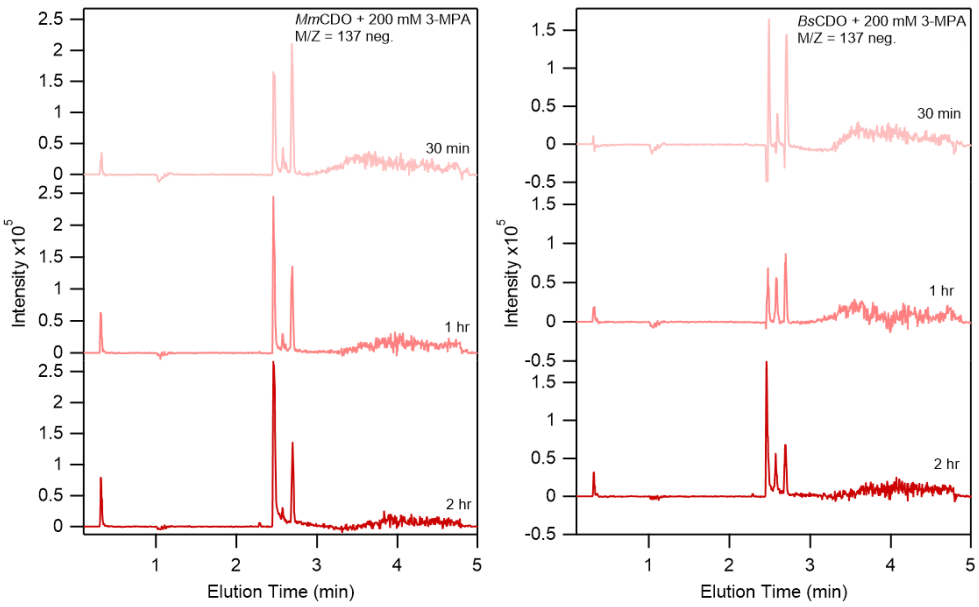


Figure A.3.3 Mass spectra at the M/Z corresponding to 3-SPA of quenched reaction mixtures of *Mm*CDO and *Bs*CDO with 3-MPA at three time points. A 2 hour no-enzyme 3-MPA control has been background subtracted from each spectrum.

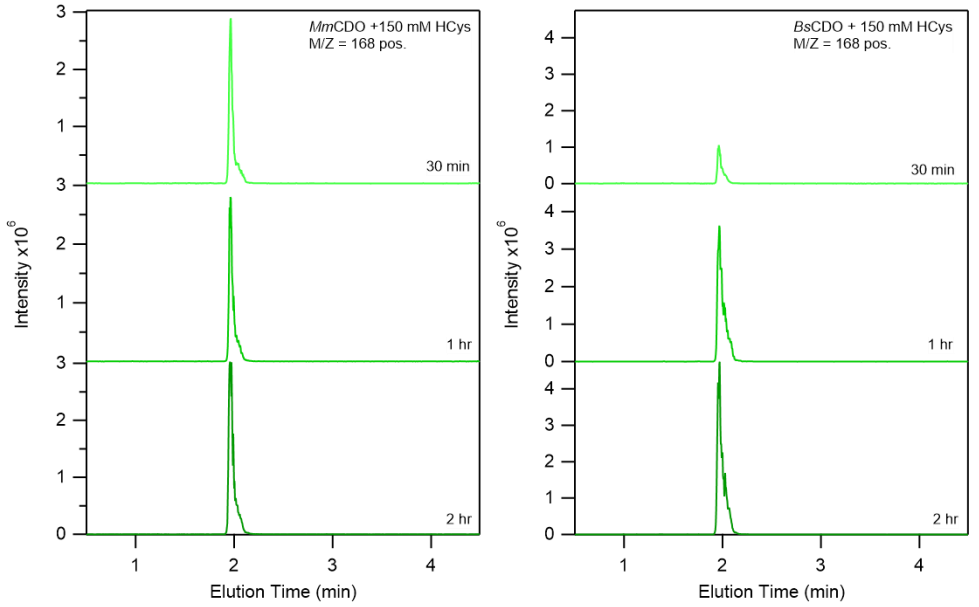


Figure A.3.4 Mass spectra at the M/Z corresponding to HCSA of quenched reaction mixtures of *Mm*CDO and *Bs*CDO with HCys at three time points. A 2 hour no-enzyme HCys control has been background subtracted from each spectrum.

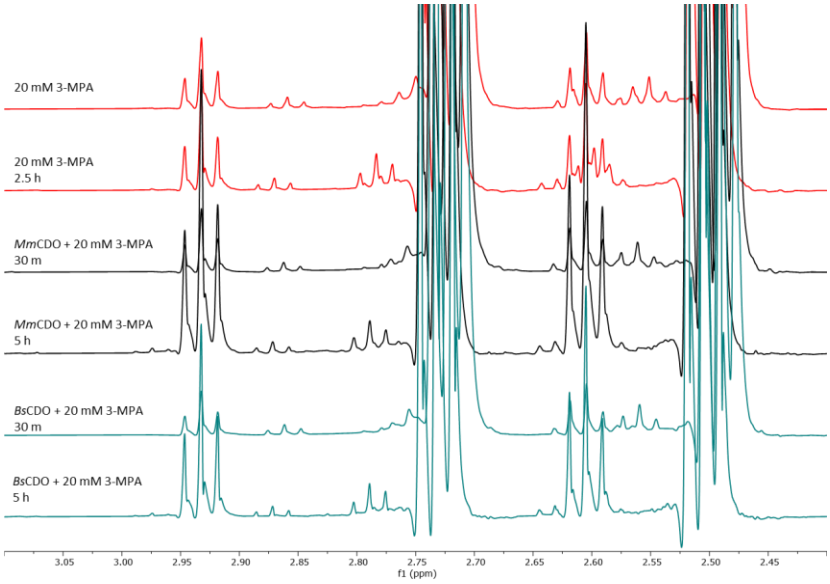


Figure A.3.5 Close up view of the ^{1}H-NMR spectra collected at two time points after incubating MmCDO (black) and BsCDO (teal) with substrate analogue 2-AET. Spectra of standard solutions of reactant 2-AETand product HT are shown at the top.

Matlab Code A.4.1 FindY_Indices

```
clear all;
AAs = cell2mat(importdata('AllCDOs Aligned Pos Y157.txt'));
Y_indices = find(AAs=='Y');


CDOs = importdata('AllCDOsAlignedText.txt');
counter = 1;
for k = 1:size(CDOs,1)
    if CDOs{k}(1) == '>'
        CDO_index{counter} = CDOs{k};
        counter = counter + 1;
    end
end


for a = 1:length(Y_indices)
    IdentifiersWithYs{a} = CDO_index{Y_indices(a)};
end
```

```
for b = 1:length(Y_indices)
    Slashes = find(IdentifiersWithYs{b}=='|',2);
    CharactersYouNeed{b} = IdentifiersWithYs{b}((Slashes(1)+1):(Slashes(2)-1));
end
 CDOSwithY157=char(CharactersYouNeed);
```

Matlab Code A.4.2 FindC_Indices

```
clear all;
AAs = cell2mat(importdata('AllCDOs Aligned Pos C93.txt'));
C_indices = find(AAs=='C');

CDOs = importdata('AllCDOsAlignedText.txt');
counter = 1;
for k = 1:size(CDOs,1)
    if CDOs{k}(1) == '>'
        CDO_index{counter} = CDOs{k};
        counter = counter + 1;
    end
end

for a = 1:length(C_indices)
    IdentifiersWithCs{a} = CDO_index{C_indices(a)};
end

for b = 1:length(C_indices)
    Slashes = find(IdentifiersWithCs{b}=='|',2);
    CharactersYouNeed{b} = IdentifiersWithCs{b}((Slashes(1)+1):(Slashes(2)-1));
End
```

Matlab Code A.4.3 CreateClusterAlignment

```
clear all;

AllCDOs=importdata('AllCDOsAlignedText.txt');
String=strjoin(AllCDOs);
PreSplit=regexprep(String,'>','\n>');
Split=strsplit(PreSplit,'\n');

SubsetIDs=string(importdata('noncrosslinked eukaryotes IDs.txt'));
FindInSeqs=contains(Split,SubsetIDs);
Subset_indices = find(FindInSeqs==1);

Answer=Split(Subset_indices);
JoinAnswer=strjoin(Answer);
AAs=[" ","-","G","P","A","V","L","I","M","C","F","Y","W","H","K","R","Q","N","E","D","S","T"];



for i=1:22
    first=AAs(i);
    for j=1:22
        last=AAs(j);
        original(i,j)=strcat(first,{' '},last);
        replacement(i,j)=strcat(first,'{b}',last);
    end
end

replacement=regexprep(replacement,'{b}','\n');
FormatAnswer=regexprep(JoinAnswer,original(:,:),replacement(:,:));
```

```
clipboard('copy',FormatAnswer);
```

Matlab Code A.4.4 FindY157and93Pairs

```
%% get frequencies and indices from positions 157


clear all;
AAs_157 = cell2mat(importdata('AllCDOs Aligned Pos Y157.txt'));
uniqueAAs_157 = unique(AAs_157);              %list of all AAs at position 157


%CDOs = importdata('AllCDOsAlignedText.txt');
% counter = 1;
% for k = 1:size(CDOs,1)
%     if CDOs{k}(1) == '>'
%         CDO_index{counter} = CDOs{k};
%         counter = counter + 1;
%     end
% end


for k = 1:length(uniqueAAs_157)
   IndicesForSpecificAAs{k} = find(AAs_157 == uniqueAAs_157(k))        %%indices of specific AAs, e.g. 4,5, 24 for Alanine, or whatever
end


for a = 1:length(uniqueAAs_157)
   AA157_frequencies(a) = sum(IndicesForSpecificAAs{a});              %%frequencies of the 157 residues
end


% %for b = 1:length(Y_indices)
```

```
%     Slashes = find(IdentifiersWithYs{b}=='|',2);


%     CharactersYouNeed{b} = IdentifiersWithYs{b}((Slashes(1)+1):(Slashes(2)-1));
% %end
%  %CDOSwithY157=char(CharactersYouNeed);


AAs_93 = cell2mat(importdata('AllCDOs Aligned Pos C93.txt'));
uniqueAAs_93 = unique(AAs_93)'              %list of all AAs at position 157


for k = 1:length(uniqueAAs_157)
   AA_157_93_pairs{k} = AAs_93(IndicesForSpecificAAs{k});              % for each 157 residues, there are these 93 residues
end


Frequency_of_this_93_157_pair = zeros(length(uniqueAAs_157),length(uniqueAAs_93));


for k = 1:length(uniqueAAs_157)
    Unique_93residues_forthis157residue{k} = unique(AA_157_93_pairs{k});
    for m = 1:length(uniqueAAs_93)
        Frequency_of_this_93_157_pair(k,m) = sum(AA_157_93_pairs{k}(:)==uniqueAAs_93(m));
    end
end


%PrettyTable                                                                                                    = zeros(size(Frequency_of_this_93_157_pair,1)+1,size(Frequency_of_this_93_157_pair,2)+1);
%PrettyTable(1,1:end) = [0,(uniqueAAs_93)];
%PrettyTable(1:end,1) = [0,(uniqueAAs_157)'];
% PrettyTable(2:end,2:end) = Frequency_of_this_93_157_pair;
```

Matlab Code A.4.5 FormatTaxIDs

```
clear all;
filename = 'Bottom Left Xlinked Bacteria TaxLineage.txt';
Data = readtable(filename);

for k = 1:size(Data)
    tempdata = cell2mat(table2array(Data(k,1)));
    [~,ind] = find(tempdata =='|',3);
    if length(ind)>2
        newData{k} = tempdata(ind(3)+2:end);
    else
        newData{k} = tempdata;
    end
end
String=strjoin(newData);
clipboard('copy',String);
```

Matlab Code A.4.6 SortTaxLineage

```
clear all;
CharacteristicNums = importdata('Low E Top Right TaxLineage.txt');

if isstruct(CharacteristicNums)==1
    tempdata = CharacteristicNums.textdata;
    UniqueCharNums = unique(tempdata);
    UniqueCharNums{end}(1) = [];

    RelFreq = zeros(1,length(UniqueCharNums));
    for k = 1:length(UniqueCharNums)
        k.*100./length(UniqueCharNums)
```

```
        for j = 1:length(tempdata)
            if string(UniqueCharNums{k})==string(tempdata{j})
                RelFreq(k) = RelFreq(k)+1;
            end
        end
    end
else
    tempdata = CharacteristicNums;
    UniqueCharNums = unique(tempdata);

    RelFreq = zeros(1,length(UniqueCharNums));
    for k = 1:length(UniqueCharNums)
        k.*100./length(UniqueCharNums)
        for j = 1:length(tempdata)
            if string(UniqueCharNums(k))==string(tempdata(j))
                RelFreq(k) = RelFreq(k)+1;
            end
        end
    end
end

[RelFreq,RelFreqOrder] = sort(RelFreq);
UniqueCharNums = UniqueCharNums(RelFreqOrder);


ListOfRelativeAbundances = [string(char(UniqueCharNums)),RelFreq];

% for a = 1:length(tempdata)
%
```

```
% for k = 1:length(UniqueCharNums)
%     tempData = char(CharacteristicNums.textdata);
%
%     RelativeInt(k) = sum((char(CharacteristicNums.textdata)) == (char(UniqueCharNums(k))));
% end
```

Matlab Code A.4.7 FindOrganismInfo

```
clear all;
Phyla = string(char(importdata('class high e bottom left.txt')));
Unique_Phyla = string(char(unique(Phyla)));

for k = 1:length(Unique_Phyla)
   Phyla_frequency(k) = sum(string(char(Phyla)) == string(char(Unique_Phyla(k))));
end

[Phyla_frequency,sort_order] = sort(Phyla_frequency,'descend');
Unique_Phyla = Unique_Phyla(sort_order);

ListOfRelativeAbundances = [Unique_Phyla,Phyla_frequency'];
```

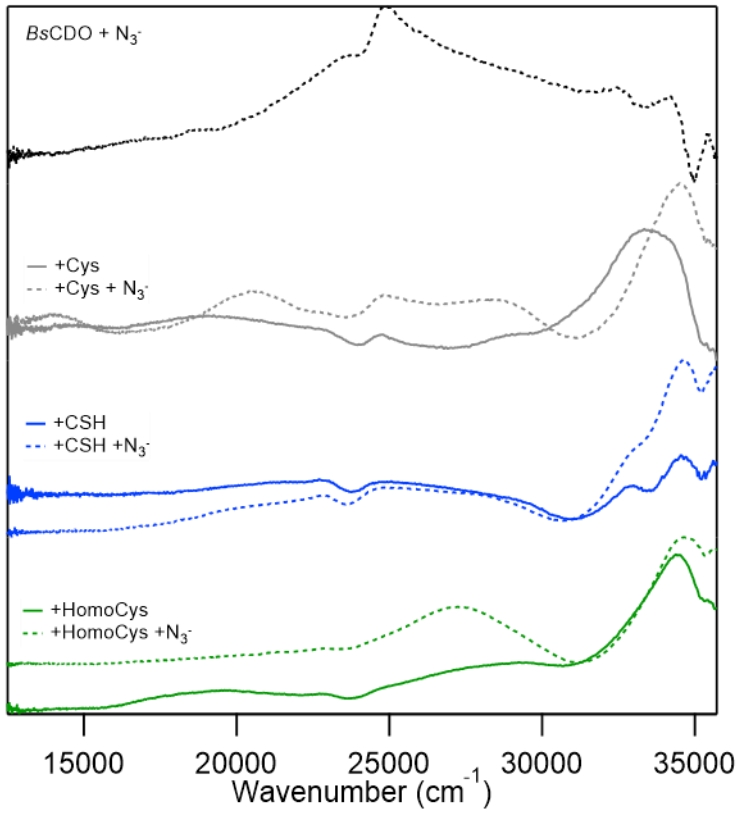


Figure A.5.1 MCD spectra of substrate (analogue)-bound *Bs*CDO with and without addition of 100-fold excess azide.

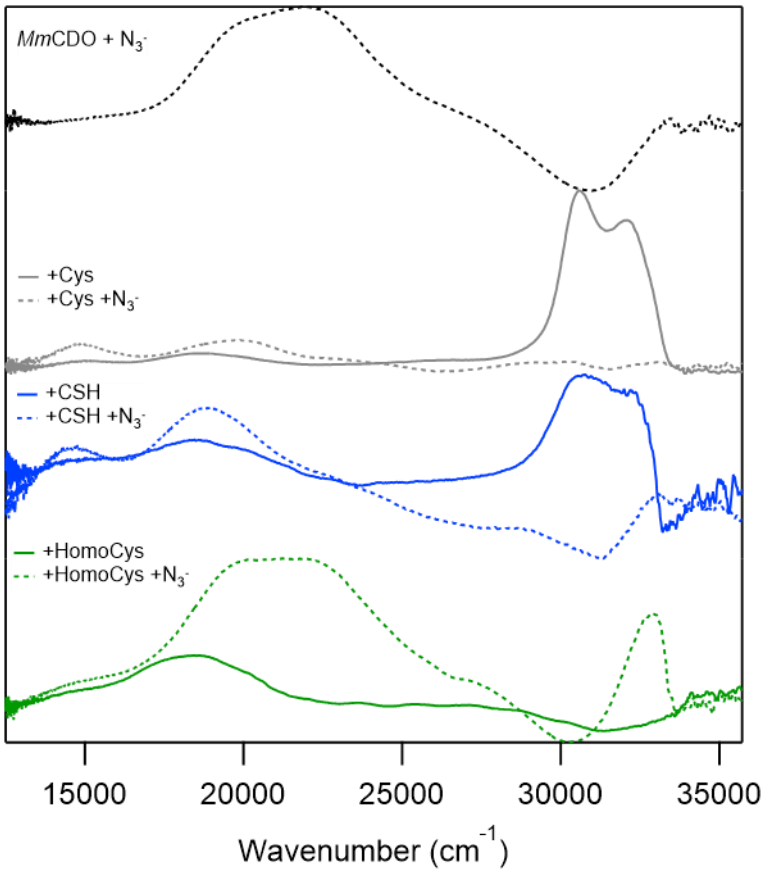


Figure A.5.2 MCD spectra of substrate (analogue)-bound *Mm*CDO with and without addition of 100-fold excess azide.