Marcus Keßler*, Ilka Wittig, Jörg Ackermann and Ina Koch*

Prediction and analysis of redox-sensitive cysteines using machine learning and statistical methods

https://doi.org/10.1515/hsz-2020-0321
Received May 10, 2020; accepted December 7, 2020;
published online January 6, 2021

Abstract: Reactive oxygen species are produced by a number of stimuli and can lead both to irreversible intracellular damage and signaling through reversible post-translational modification. It is unclear which factors contribute to the sensitivity of cysteines to redox modification. Here, we used statistical and machine learning methods to investigate the influence of different structural and sequence features on the modifiability of cysteines. We found several strong structural predictors for redox modification. Sensitive cysteines tend to be characterized by higher exposure, a lack of secondary structure elements, and a high number of positively charged amino acids in their close environment. Our results indicate that modified cysteines tend to occur close to other post-translational modifications, such as phosphorylated serines. We used these features to create models and predict the presence of redox-modifiable cysteines in human mitochondrial complex I as well as make novel predictions regarding redox-sensitive cysteines in proteins.

Keywords: cysteine; human mitochondrial complex I; machine learning; post-translational modification; proteomics; redox.

Introduction

Reactive oxygen species (ROS) are metabolic by-products of cellular processes, such as oxidative phosphorylation, or are directly produced by enzymes like NADPH (nicotinamide adenine dinucleotide phosphate) oxidases (Groitl and Jakob 2014). Among the various ROS, superoxide (O_2^-) and hydrogen peroxide (H_2O_2) are physiologically most relevant. A number of stimuli can cause an imbalance of ROS production and consumption. Examples are hypoxia, ischemia/reperfusion injury, inflammation, environmental pollution, strenuous exercise, smoking, nutrition, and psychological stress (Poljsak et al. 2013; Pell et al. 2018). An elevation of the concentration of ROS may lead to oxidative stress, damage of macromolecules, and apoptosis.

Despite their well-known negative effects, ROS have vital biological functions as targeted secondary messengers. Many oxidative post-translational modifications are reversible and act as a binary switch. Functional consequences of ROS-signaling can be involved in changes in many different pathways, for instance, gene transcription, translation and protein folding, metabolism, signal transduction, apoptosis and others (Brandes et al. 2009). The majority of functional redox modifications occur with redox-reactive cysteines. Oxidation of the thiol forms reactive sulfenic acid and may establish disulfide bonds with nearby cysteines or undergo further oxidation to sulfinic or sulfonic acid, resulting in changes in structure and/or function of the protein (Ray et al. 2012). While the latter two are generally irreversible, sulfenic acids and disulfide bonds are reversible by reducing proteins such as thioredoxins and glutaredoxins (Åslund et al. 1997; Holmgren 1989). Reversibility is a necessary precondition for redox-modified cysteines to function in non-pathological pathways.

Because of the importance of redox modification to intercellular processes, numerous studies have been conducted to classify the many redox-sensitive proteins, their underlying stimuli as well as their effects. Most work has involved experimental research, both in the form of case studies for proteins as well as in the form of large-scale proteomics investigations. Examples include experimental research into the inhibition of mouse complex I through the S-nitrosation of Cys39 on the ND3 (NADH dehydrogenase 3) subunit. Reversible S-nitrosation slows the reactivation of complex I during reperfusion, protecting tissue from oxidative damage through ROS imbalance. The nitrosation of the cysteine in complex I has been marked using a

***Corresponding authors: Marcus Keßler and Ina Koch**, Molecular Bioinformatics Group, Institute of Computer Science, Goethe-University, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany, E-mail: Marcus.Kessler@bioinformatik.uni-frankfurt.de (M. Keßler); ina.koch@bioinformatik.uni-frankfurt.de (I. Koch).
https://orcid.org/0000-0002-5652-8424 (M. Keßler)
Ilka Wittig, Functional Proteomics Group, Medical School, Goethe-University, Theodor-Stern-Kai 7, 60590, Frankfurt am Main, Germany
Jörg Ackermann, Molecular Bioinformatics Group, Institute of Computer Science, Goethe-University, Robert-Mayer-Str. 11-15, 60325, Frankfurt am Main, Germany

mitochondria-selective S-nitrosating agent, MitoSNO (mitochondria-targeted S-nitrosothiol), and then identified using SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) (Chouchani et al. 2013). Redox-proteomics investigations include the studies of Murphy et al. into ROS-sensitive thiols and their implications for mitochondrial function and redox signaling, the research of Bleier et al. for generator-specific targets of ROS, or the experiments by Martìnez-Acedo et al. for their new method for the global analysis of the redox proteome (Bleier et al. 2015; Chouchani et al. 2010; Hurd et al. 2007; Martìnez-Acedo et al. 2015; Requejo et al. 2010).

To assist investigations on redox systems, we provide a novel application of a machine learning approach, which may be able to predict redox-modifiable cysteines and hence, to reduce the necessary load of time and effort for researchers. Similar approaches have been attempted for related problems and have reported differing levels of success. Marino & Gladyshev have developed an approach based on active-site similarity and cysteine reactivity to predict the presence of thiol oxidoreductases among proteomes. Testing their method for the proteome of *Saccharomyces cerevisiae*, they have been able to identify the majority of known yeast thiol oxidoreductases (Marino and Gladyshev 2009). In another study, the group has analyzed S-nitrosylation and has found that nitrosylation could be predicted by the presence of a distantly situated, exposed acid-base motif (Marino and Gladyshev 2010).

In 2012 they published a review of methods of the computational analysis of reactive cysteines (Marino and Gladyshev 2012), praising new insights for the understanding of catalytic redox cysteines, metal-binding cysteines, and disulfide bonds, despite the lack of progress in the case of regulatory cysteines, sites of stable post-translational modifications (PTMs) and catalytic non-redox cysteines due to the complexity of problem. A machine learning approach based on the support vector machine algorithm has been used to differentiate between cysteines involved in ligand binding and cysteines forming disulfide bridges, using multiple alignment profiles. Their approach has yielded better results than predictions based on PROSITE patterns (Passerini and Frasconi 2004). An approach to predict whether cysteine exists in a free state, metal bound, or participate in disulfide bridges has used support vector machine and neural networks algorithms, applying position–specific evolutionary profiles and features such as chain length and amino acid composition of the protein. The approach has achieved similar results as predictions based on PROSITE patterns (Passerini et al. 2006).

Despite initial successes, there is no reliable and fast way to predict the redox sensitivity of cysteines in proteins. Our approach applied statistical as well as machine learning methods to computationally find and highlight redox-modifiable cysteines. By additionally considering the structure of the proteins, we wanted to achieve better results than comparable approaches that use only amino acid sequences. We utilized features like physicochemical properties, amino acid accessibility, Half Sphere Exposure (HSE) and secondary structure elements to train models. We applied various supervised machine learning algorithms. We tested the models for the prediction of redox-sensitive cysteines. We compared properties of cysteines that had been experimentally shown to be modifiable (hereafter referred to as Cys+) with properties of cysteines that had been investigated but not found to be modifiable (hereafter referred to as Cys–). We considered only proteins for which at least one redox-sensitive cysteine had been experimentally reported.

Results

Statistics

We studied the differences between composition of residues in the 3D neighborhood of Cys+ and Cys–. We considered the 10 spatially nearest residues. Significance values are indicated as follows: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

Figure 1 shows the deviation of the composition of residues in the spatial neighborhood of Cys+ from the composition of residues in the spatial neighborhood of Cys– of protein set 1. The frequencies of lysine, glycine, proline, serine, and threonine are significantly elevated. It has been suggested (Tanner et al. 2011) that serine and threonine may be involved in the formation of sulfenyl esters together with sulfenic acids formed by the oxidation of cysteines, while lysines and histidines may be involved in the formation of sulfenamides, such as in the case of Protein Tyrosine Phosphatase 1B (PTP1B) (Salmeen et al. 2003; Sarma and Mugesh 2007), which may be a possible explanation for the relative abundance of these amino acids. The frequency of leucine is significantly reduced.

When grouped for their features, significant differences could be found between the two groups of cysteines for all features, see Supplementary Figure S2. We found the strongest difference for positive charge as well as the presence of aliphatic side chains. While there is much overlap between the data for modifiable and unmodifiable cysteines, these significant differences may be one

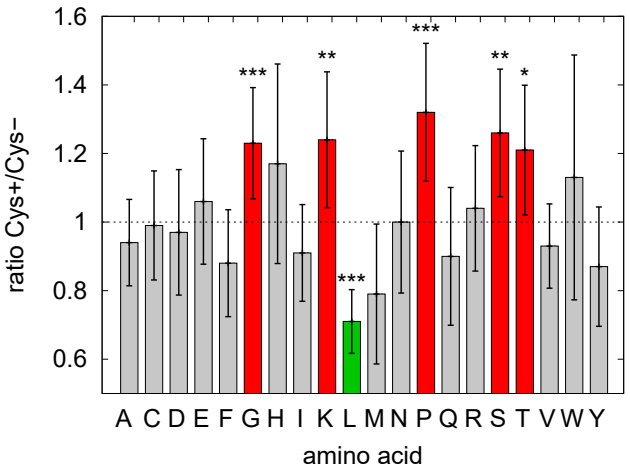

Figure 1: Relative frequencies of residues in the 3D neighborhood of Cys+. Red bars indicate significantly elevated frequencies and green bars significantly reduced frequencies. The error bars are $3 \times \sigma$ with σ being the Poisson standard deviation for the total number of counts. The frequencies of the amino acids glycine (G), lysine (K), proline (P), serine (S), and threonine (T) are significantly elevated. The relative frequency of leucine (L) is significantly reduced.

important piece of the puzzle when trying to predict modifications.

We created a sequence logo (Crooks et al. 2004) of the different frequencies of the closest amino acids in the sequence neighborhood of the cysteines in the two groups of protein set 2, as seen in Figure 2. The hydrophobic leucine (L) and cysteine (C) are reduced in the sequence logo around a Cys+, especially and significantly at position –1 and positions –3, 3 and 6, respectively.

Positively charged amino acids like lysine (K) and arginine (R) are unaffected in the direct sequence neighborhood of Cys+, but are enriched in the further sequence neighborhood, i.e., for more than four positions away from the cysteine, confirming the results of Chen et al. (2015), who also found an abundance of positively charged residues around *S*-nitrosylation sites and a reduced occurrence of C. Note that positively charged amino acids are dominant also in the 3D neighborhood, see Figure 1 and Supplementary Figure S2. Aromatic amino acids like phenylalanine (F) and tryptophan (W) were also reduced in the sequence neighborhood of Cys+, see Supplementary Figure S1.

We were unable to confirm the presence of an acid-base motif which was found in some previous studies on smaller datasets (Greco et al. 2006; Hess et al. 2005). Marino and Gladyshev (2010) proposed the presence of a modified acid-base motif, consisting of a positively charged residue in close proximity to the cysteine and a negatively charged amino acid up to eight Å away.

We detected a higher abundance of phosphorylated residues in the sequence neighborhood of Cys+ than Cys– in protein set two according to Uniprot (Consortium 2019) annotations. On average, each Cys+ had around 0.097 phosphorylated serines, threonines and tyrosines in its neighborhood of 10 residues upstream and downstream, while we observed only 0.055 phosphorylated residues around Cys–, making phosphorylation about 1.76 times more common around Cys+ (data not shown). We found similar results for other post-translational modifications, like acetylation. We discovered 0.063 acetylated residues in the neighborhood of the average Cys+, with only 0.045 modified lysines present around Cys–, a factor of 1.48 (data not shown). Ubiquitination also appeared more commonly around Cys+, where we located an average of 0.016 ubiquitination sites around modified cysteines, while we only detected an average of 0.002 ubiquitination sites around unmodified cysteines. These results were statistically significant even when corrected for the different abundances of serines and lysines around Cys+ in the case of phosphorylation and ubiquitination with p-value <0.001 according to the Mann-Whitney U test. Results for acetylation were not statistically significant.

We explored the accessible surface area as predicted by the algorithm DSSP (Define Secondary Structure of Proteins)

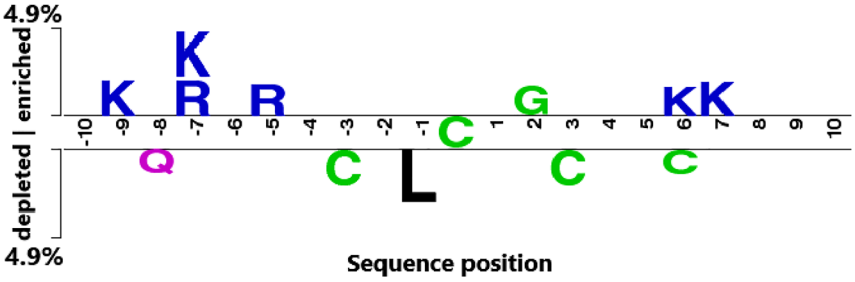

Figure 2: Sequence logo (Crooks et al. 2004) of differences between the residues in the neighborhood of Cys+ and Cys–. Sequence position in relation to cysteine are shown on the *x*-axis, percentage difference of Cys+ in relation to Cys– on the *y*-axis. Enriched residues around Cys+ are shown at the top, depleted residues at the bottom. Size of symbols is proportional to the difference between the two samples. Only differences with a p-value <0.05 according to the t-test are shown.

(Kabsch and Sander 1983) as well as HSE in the 3D neighborhood of cysteines in protein set 1. Cys+ showed a higher accessible surface area than Cys– with distributions that differed with high statistical significance (p-value $<4 \times 10^{-6}$), despite much overlap in the range of values, see Supplementary Figure S3. This appears reasonable, as an exposed cysteine should be assumed to be more easily accessible to ROS. HSE showed similar results, see Supplementary Figure S4 and Supplementary Figure S5. We also tested the accessibility and HSE of residues in the 3D neighborhood of Cys+. Again, both DSSP and HSE showed that residues around Cys+ were significantly more accessible.

We examined the relative frequencies of secondary structure elements (SSEs) in the neighborhood of Cys+ in protein set 1. In the 3D neighborhood of Cys+, we found the frequencies of β-strands as well as of bends and unstructured loop regions to be significantly elevated, see Supplementary Figure S6. Statistical significance could not be found for the other SSEs, often due to their low frequency of occurrence in general. This differs from the results of Marino and Gladyshev (2009), who found a marked preference for both α-helical and loop geometries around thiol oxidoreductases, testing a more limited dataset of 75 structures. We found that Cys+ themselves had a much higher chance than Cys– to be present in loop structures, and found a higher incidence of β-strands upstream and α-helices downstream from Cys+ than the reverse, while the ratios for Cys– were more balanced, confirming the findings of Fomenko et al. (2007).

We created a heatmap of the Pearson correlation coefficient between the features in protein set 1, see Supplementary Figure S7. The heatmap showed a high negative correlation between accessibility and HSE, as both features display related physical properties, whereby HSE shows a stronger correlation to cysteine redox-sensitivity. Features of residues closer to the investigated cysteine tended to have a stronger correlation to its redox-sensitivity than those of more distant residues. The most highly correlated features were SSE and HSE of the cysteine.

Machine learning

We applied all three machine learning algorithms on the datasets to be able to infer redox-sensitive cysteine sites in proteins. We used both a dataset consisting of proteins from mammals, plants and fungi as well as a dataset consisting of proteins strictly from mammals. We used 5-fold cross-validation to guard against overfitting. We also implemented feature selection to remove superfluous data.

The best Area Under the Curve (AUC) value for the Receiver Operating Characteristic (ROC) curve was 0.72 for both the imputed mammal and the full imputed datasets using the ET algorithm, see Figure 3, Table 1 and Table 2. The ROC curves depict average values over the different cross-validation folds.

To compare the performance of the algorithms with the statistical results, we ran training and testing with cross-validation using only a subset of our features in the same configuration, see Supplementary Table S1. In general, we achieved the highest AUC value when only using the residue identity in the neighborhood of cysteines. We also received favorable results by using only HSE of close residues, which was superior to using only accessibility. This is in agreement with the results of the feature heatmap, which indicated a higher correlation of cysteine redox-sensitivity with HSE. The AUC value using secondary structure was on a similar level as for accessibility.

We evaluated the variance of our features using the ANOVA F-test to compare the feature importances based on the not imputed sequence dataset for all species, see Supplementary Table S2. A higher score signifies that the feature is more dependent in the target variable, i.e. the redox activity of the cysteine. We find that in general, features concerning the cysteine itself, like its secondary structure or exposure, show a high impact on the modifiability of the cysteine. Residue identity also showed high importance, followed by HSE and accessibility.

We trained the ET algorithm, which showed the highest AUC value during model training, on the full dataset 2, using imputation to make novel predictions for cysteine sensitivity in proteins. These proteins were also collected from the RedoxDB and are suspected to contain redox-sensitive cysteines, but their sequence positions are not known. The list included 121 proteins with 575 cysteines. We ran the predictions 10 times for more robust results. Out of the 121 proteins, 44 included cysteines with a prediction score larger than 0.45, which we applied as a threshold for a cysteine likely being a Cys+. We chose this value, instead of the usual value of 0.50, since the algorithm tends to strongly underestimate the number of Cys+ due to the large number of Cys– in our dataset. 11% of the cysteines were predicted as Cys+. For the full list of positively predicted cysteines, see Table 3 and Table 4. We compared the predictions with known data of post-translational modifications from the UniProt (Consortium 2019) database. We found that two of our predictions, CYS37 and CYS40 in Protein Data Bank (PDB) file 1X5D (available at https://www.rcsb.org/structure/1x5d) of protein disulfide-isomerase A6, are known to form a redox-active disulfide bond according to UniProt. None of the other examined

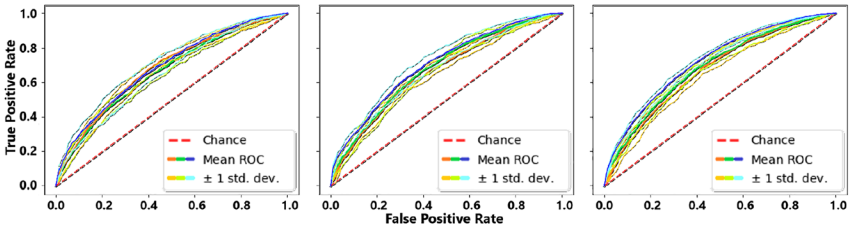

Figure 3: ROC curves of the ET (blue), RF (green) and SVM (orange) algorithms for the full structural (left), sequence (middle) and imputed sequence (right) datasets. Red line shows a completely random prediction. The ROC curve shows average values over the different cross-validation folds.

Table 1: AUC of the three different algorithms ET, RF and SVM on the mammal dataset using structural data, sequence data or imputed sequence data, as well as average results for the algorithms (rightmost column) and datasets (bottom row). The highest result is underlined. The AUC is an average value over the different cross-validation folds.

	Structure	Sequence	Imputation (Seq.)	$\bar{x}$
ET	0.67	0.70	<u>0.72</u>	0.70
RF	0.66	0.70	0.67	0.68
SVM	0.66	0.65	0.65	0.65
$\bar{x}$ dataset	0.66	0.68	0.68	

Table 2: AUC of the three different algorithms ET, RF and SVM on the full dataset using structural data, sequence data or imputed sequence data, as well as average results for the algorithms (rightmost column) and datasets (bottom row). The highest result is underlined. The AUC is an average value over the different cross-validation folds.

	Structure	Sequence	Imputation (Seq.)	$\bar{x}$
ET	0.69	0.70	<u>0.72</u>	0.70
RF	0.66	0.68	0.69	0.68
SVM	0.68	0.66	0.67	0.67
$\bar{x}$ dataset	0.68	0.68	0.69	

cysteines were known locations of redox activity according to UniProt. These findings reflect positively on the reliability of our methods. We predicted redox activity in disulfide bridges not known to be redox-active in two more proteins, CYS 120 and CYS 127 in PDB file 1B56 (Hohoff et al. 1999) for fatty acid-binding protein 5, where both cysteines were predicted as CYS+, as well as CYS105 and CYS137 in PDB files 3O0W (Kozlov et al. 2010) and 3POW (Chouquet et al. 2011) for calreticulin, where only CYS137 was predicted to be CYS+. These results do not seem unexpected and may even be a positive outcome, as disulfide bridges are one of the redox modifications we aim to predict, and not all disulfide bridges that are known to be redox-active according to redoxDB are marked as such in other databases like UniProt. Several other proteins contained disulfide bridges, which were not predicted as redox-active, like prothrombin (3HK3 (Gandhi et al. 2009)), the cytochrome b6-f complex iron-sulfur subunit (1RFS (Carrell et al. 1997)) and the non-specific lipid transfer protein (1BWO (Charvolin et al. 1999)).

Use case

After using different algorithms to create models for the prediction of redox-sensitive cysteines, we applied our models to generate predictions for the *NDUFS1*, *MT-ND3*, *NDUFA2* and *NDUFC2* subunits of mammalian respiratory complex I, based on the structural data from PDB entries 6G2J (Agip et al. 2018), 6G72 (Agip et al. 2018), 5LC5 (Zhu et al. 2016), 5LNK (Fiedorczuk et al. 2016) and 5XTD (Guo et al. 2017).

Mitochondrial complex I is with its mass of around 1 MDa the largest protein complex in the respiratory chain. It is responsible for the oxidization of NADH produced mainly by the Krebs cycle, transferring electrons to the ubiquinone pool in the process (Wirth et al. 2016). The respiratory chain continuously reduces O_2 into H_2O, whereby a small amount of ROS in the form of O_2^- is generated. It has also been proposed that ROS formed at complex III could have a direct feedback on complex I cysteines (Larosa and Remacle 2018).

Which cysteines are likely targets of such activity could be predicted by our models. We applied the models trained with the random forest (RF), support vector machine (SVM) and extra trees (ET) algorithms to the complex, using the parameters and training data detailed in the previous sections. Highest AUC was received for the mammalian structural and imputed datasets. For illustration, we used the PDB entry with the accession number 5XTD of mitochondrial complex I (Guo et al. 2017), as it is the most complete structure of human origin.

We compared the predictions of all algorithms based on the mammalian structural and imputed datasets to the experimental results from Chouchani et al. (2013). 25% of

Table 3: PDB ID, chain ID, amino acid ID and prediction score of a list of cysteines from proteins suspected of being redox-sensitive collected from RedoxDB. Only cysteines with scores greater than 0.45 are shown. Higher scoring cysteines are more likely to be redox-sensitive according to our methods.

Protein name	PDB/Chain ID	AA ID	Prediction	Molecular function
Fructose-bisphosphate aldolase B	1QO5/A	201	0.452	Lyase
Fructose-bisphosphate aldolase B	1QO5/A	239	0.464	Lyase
Nuclear factor NF-kappa-B p100 subunit	3JV5/A	250	0.466	Activator, DNA-binding, repressor
All-trans-retinol dehydrogenase [NAD(+)] ADH4	3COS/A	246	0.465	Oxidoreductase
Amine oxidase [flavin-containing] A	1O5W/A	3321	0.477	Oxidoreductase
Amine oxidase [flavin-containing] A	1O5W/A	3323	0.472	Oxidoreductase
Serine/threonine-protein phosphatase 2A	1B3U/A	389	0.480	Chromosome partition
Inosine-5′-monophosphate dehydrogenase 2	1NF7/A	26	0.467	DNA-binding, oxidoreductase, RNA-binding
Inosine-5′-monophosphate dehydrogenase 2	1NF7/A	173	0.477	DNA-binding, oxidoreductase, RNA-binding
Drebrin-like protein	1X67/A	134	0.464	Actin-binding
Adenylosuccinate lyase	2J91/A	98	0.465	Lyase
Adenylosuccinate lyase	2J91/A	399	0.456	Lyase
Glutathione peroxidase 2	2HE3/A	55	0.476	Oxidoreductase, peroxidase
Aflatoxin B1 aldehyde reductase member 3	1GVE/A	66	0.470	Oxidoreductase
Aflatoxin B1 aldehyde reductase member 3	1GVE/A	322	0.461	Oxidoreductase
Adenylosuccinate synthetase, chloroplastic	1DJ3/A	111	0.455	Ligase
Deoxycytidine kinase	1P5Z/B	45	0.451	Kinase, transferase
Deoxycytidine kinase	1P5Z/B	59	0.486	Kinase, transferase
Fructose-bisphosphate aldolase A	1ALD/A	201	0.457	Lyase
Uroporphyrinogen decarboxylase, chloroplastic	1J93/A	45	0.450	Decarboxylase, lyase
Maltose binding protein fusion with RACK1	3DM0/A	1184	0.459	Sugar transport, transducer
Translationally-controlled tumor protein	1YZ1/A	28	0.452	
Peptidyl-prolyl cis-trans isomerase A	4DGD/A	52	0.511	Isomerase, rotamase
Peptidyl-prolyl cis-trans isomerase A	4DGD/A	70	0.484	Isomerase, rotamase
Peptidyl-prolyl cis-trans isomerase A	4DGD/A	161	0.455	Isomerase, rotamase
Eukaryotic translation initiation factor 5A-1	3CPF/A	38	0.469	Elongation factor, RNA-binding
Eukaryotic translation initiation factor 5A-1	3CPF/A	73	0.467	Elongation factor, RNA-binding
Nucleoside diphosphate kinase B	3BBB/A	145	0.467	Activator, DNA-binding, kinase, transferase
Protease do-like 1, chloroplastic	3QO6/A	409	0.450	Hydrolase, protease, serine protease
Fatty acid-binding protein 5	1B56/A	47	0.469	Transport
Fatty acid-binding protein 5	1B56/A	67	0.456	Transport
Fatty acid-binding protein 5	1B56/A	120	0.451	Transport

the cysteines of the relevant subunits were predicted correctly by all algorithms in both datasets, see Supplementary Table S5. Additionally, 30% were predicted correctly by the majority of algorithms. Overall, 68% of the individual predictions of the algorithms for the investigated cysteines agreed with experimental data. For the 3D structure of complex I with cysteines experimentally validated to be modified, see Figure 4. For a closer look at the structural features of CYS367, CYS554 and CYS564 in subunit *NDUFS1*, see Supplementary Figure S8. Supplementary Figure S9 shows all investigated cysteines as well as their structural features, with features that are more commonly observed in Cys+ in green, in red features observed in Cys−, and in black those observed in both groups at about the same frequency. The machine learning models tend to show positive predictions for cysteines with a high number of features that are statistically associated with redox modification.

Discussion

We applied machine learning approaches to the prediction of redox-sensitive cysteines, successfully inferring the modifiability of 72% of the cysteines in both the mammal and the full imputed dataset. We showed that several structural features of the close neighborhood of cysteines were significant for functional analysis and prediction. We received the highest AUC with an imputed dataset for training.

The statistical analyses revealed which features of the residue environment of cysteines possess a strong

Table 4: PDB ID, chain ID, amino acid ID and prediction score of a list of cysteines from proteins suspected of being redox-sensitive collected from RedoxDB. Only cysteines with scores greater than 0.45 are shown. Higher scoring cysteines are more likely to be redox-sensitive according to our methods.

Protein name	PDB/Chain ID	AA ID	Prediction	Molecular function
Fatty acid-binding protein 5	1B56/A	127	0.454	Transport
Bifunctional epoxide hydrolase 2	3I28/A	81	0.498	Hydrolase, multifunctional enzyme
Formate dehydrogenase	3JTM/A	81	0.492	Oxidoreductase
Ubiquitin-protein ligase E3A	1C4Z/D	17	0.479	Transferase
Non-specific lipid-transfer protein	1BWO/A	87	0.481	Transport
Catechol O-methyltransferase	3U81/A	69	0.476	Methyltransferase, transferase
Catechol O-methyltransferase	3U81/A	95	0.468	Methyltransferase, transferase
Catechol O-methyltransferase	3U81/A	157	0.498	Methyltransferase, transferase
Catechol O-methyltransferase	3U81/A	191	0.476	Methyltransferase, transferase
Cytochrome b6-f complex iron-sulfur subunit	1RFS/A	112	0.456	Translocase
Ubiquitin thioesterase otubain-like	4DHI/D	87	0.453	Hydrolase, protease, thiol protease
Ubiquitin carboxyl-terminal hydrolase 16	2I50/A	31	0.465	Activator, chromatin regulator, hydrolase, protease
Calreticulin	3O0W/A	137	0.462	Chaperone
Glycine–tRNA ligase	2ZT5/A	126	0.471	Aminoacyl-tRNA synthetase, hydrolase, ligase
Glycine–tRNA ligase	2ZT5/A	177	0.499	Aminoacyl-tRNA synthetase, hydrolase, ligase
Thiamine thiazole synthase, chloroplastic	1RP0/A	172	0.505	Transferase
Eukaryotic translation initiation factor 5A-2	3HKS/A	39	0.457	Initiation factor
Prothrombin	3HK3/B	168	0.453	Hydrolase, protease, serine protease
Glutathione S-transferase Z1	1E6B/B	154	0.470	Isomerase, oxidoreductase, peroxidase, transferase
DNA-directed RNA polymerase II subunit RPB1	3PO3/S	271	0.465	DNA-binding, nucleotidyltransferase, transferase
DNA-directed RNA polymerase II subunit RPB1	3PO3/S	302	0.481	DNA-binding, nucleotidyltransferase, transferase
Phosphoglycerate mutase 1	1YFK/A	153	0.454	Hydrolase, isomerase
Protein disulfide-isomerase A6	1X5D/A	37	0.510	Chaperone, isomerase
Protein disulfide-isomerase A6	1X5D/A	40	0.517	Chaperone, isomerase
Crk-like protein	2BZY/A	13	0.459	
Phosphoglycerate kinase 1	2WZB/A	98	0.476	Kinase, transferase
Calreticulin	3POW/A	137	0.466	Chaperone
Calreticulin	3POW/A	163	0.458	Chaperone
Isocitrate dehydrogenase [NAD] subunit 1	3BLX/A	150	0.512	Allosteric enzyme, oxidoreductase, RNA-binding
Isocitrate dehydrogenase [NADP]	2QFY/A	383	0.464	Oxidoreductase
Serotransferrin	1RYO/A	137	0.451	Ion transport, iron transport, transport

influence on redox sensitivity. Concerning the amino acids, we could show that Cys+ had a high abundance of PTMs, such as phosphorylated serines, as well as positively charged amino acids in their neighborhood, verifying previous research (Greco et al. 2006; Hess et al. 2005; Marino and Gladyshev 2010). Cys+, as well as their amino acid environment, tended to be characterized by a higher value in accessibility. They were surrounded by a lower number of aliphatic amino acids, since redox-sensitive cysteines are found closer to the surface, while aliphatic amino acids tend to be in the interior (Sandberg and Terwilliger 1991) of proteins. Cys+ could also be shown to be situated near bends, turns or loop regions, for example, see Supplementary Figure S8. Specific sequence motifs were not found in our investigations.

Finally, our results suggest that a machine learning approach may be a valuable tool for the prediction and analysis of redox-sensitive cysteines, but that more research will have to be necessary to increase the robustness of predictive models. We propose that in the future, a number of additional attributes will need to be taken into account. Protein datasets, if large enough, will need to be separated by organism type, e.g. mammal, plant, fungus. Another important factor may be the location of a protein inside the cell, as well as the type and strength of the stimulus responsible for its redox modification. The type of modification the cysteine is able to undergo will further improve computational analyses. One potential source of complications may be the fact that it is possible that some undiscovered redox-modifiable cysteines could still be found within our dataset of Cys–. We tried to minimize this possibility by only utilizing proteins that had been investigated through experimental means before, but further research may still improve future datasets in that respect.

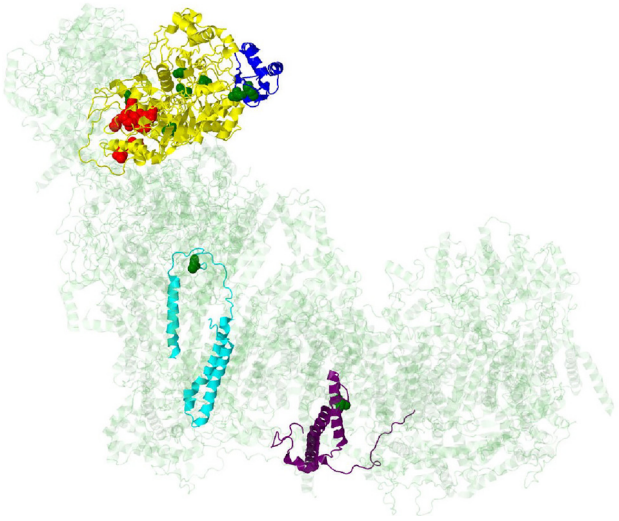

Figure 4: 3D structure of human complex I using PDB entry 5XTD (Guo et al. 2017). The subunits *NDUFS1*, *MT-ND3*, *NDUFA2* and *NDUFC2* are colored yellow, cyan, blue and purple, respectively. Green dots indicate the positions of experimentally verified (Chouchani et al. 2013) Cys+. Red dots indicate the positions of Cys−.

Materials and methods

Data sets

We generated two data sets, the first one contains structurally resolved proteins, whereas the second one was extended by proteins with unknown structure. For the protein set 1, we collected 439 redox-active proteins with 644 Cys+ and 1692 Cys− from the latest update of RedoxDB (Sun et al. 2012) up to May 2020. We selected only proteins with known structure stored in the PDB (Berman et al. 2000) and at least one Cys+. To assign SSEs and accessibility to amino acids, we applied the algorithm DSSP (Kabsch and Sander 1983). Protein set one included 369 mammal proteins, 25 plant proteins and 45 fungal proteins.

For the second data set, we considered 1097 additional redox-active proteins listed in the RedoxDB for which no structures were available. For these proteins, we imputed structural properties of amino acid residuess in proteins, such as, e.g., secondary structure and accessibility, using the values from protein set 1. We applied the function IterativeImputer with Bayesian Ridge estimator (Pedregosa et al. 2011). Combined with model training, we performed imputation separately within each cross-validation fold. By imputation, we could extend the data set by 834 mammal proteins, 127 plant proteins, and 136 fungal proteins, such that it consisted of 1536 redox-active proteins in total. The proteins contained a total number of 2250 Cys+ and 13,373 Cys−. During model training, we removed 55% of randomly chosen Cys− from the dataset to reduce bias towards Cys− in our model. We additionally removed 80% of Cys− from the incomplete, imputed dataset for the same reason.

Statistical methods

We performed statistical analyses of the environment of both Cys+ as well as Cys− to show that the chosen features should theoretically enable our models to make useful predictions. We performed all statistical tests on the not imputed dataset before pre-processing. The neighborhood of residues is one of the main predictors of post-translational modifications and catalytic activity of residues (Blom et al. 2004). Good results have been obtained for the prediction of disulfide bridges (Passerini and Frasconi 2004). No specific amino acid sequence motifs around Cys+ have been found (Greco et al. 2006; Marino and Gladyshev 2010). We applied the Mann–Whitney *U* test (Mann and Whitney 1947) to test for significance and corrected for multiple testing, using the Benjamini-Hochberg Procedure (Benjamini and Hochberg 1995) as Bonferroni correction.

Feature extraction and pre-processing

We pre-processed protein set one by two methods. The first method processed amino acid residues in the sequence, while the second method utilized the 3D neighborhood in the structure.

The first method extracted 414 features for each cysteine. We considered the 20 closest amino acids in the sequence, according to a threshold adapted from Passerini and Frasconi (2004). The second method extracted 70 features for each cysteine. The features were intended to describe physical properties and were computed for each cysteine and each of the 13 amino acid residues in its closest 3D neighborhood (Euclidean distance). Preliminary statistical tests showed that statistical significance decreased with larger neighborhood (data not shown).

Only the first method was used for the imputed dataset, since the 3D neighborhood of cysteines was unknown for the proteins with unknown structure.

For each amino acid, we computed 24 features:

- 17 values of physicochemical properties of the side chain (abbreviations in parentheses): molecular mass (Mass), volume (Vol), surface area (Area), three values for the likelihood of either large, regular or small solvent exposed area (SEA1, SEA2, SEA3), three values of propensities for the SSEs, α-helix, β-strand, or turn (Alpha, Beta, Turn), and eight binary values for the chemical classification, which are polar, non-polar, charged, positive, tiny, small, aromatic and aliphatic (Polar, Non-P, Charge, Positive, Tiny, Small, Aromatic, Aliphatic). We adopted standard values for the physicochemical properties for the amino acids from Lin et al. (2011).
- Half Sphere Exposure values *HSE*1 and *HSE*2: We applied the function HSExposure of BioPython in version 1.74. The HSE is defined here as the number of amino acid neighbors within two half-spheres with a radius of 12 Å. The sphere is divided into two halves by a plane perpendicular to the C_β-C_α vector.
- Relative accessibility: To determine the accessibility from the 3D structure, we applied DSSP (Kabsch and Sander 1983). The accessible surface area is defined as the residue water-exposed surface in Å^2. Relative accessibility is defined here as the accessible surface area divided by the maximum accessible surface area as defined by Tien et al. (2013).
- Four values to assign an SSE to the residue according to DSSP:
  - α-helix (H), 3-helix (G), 5-helix (I), which we counted as "helix"
  - residue in isolated β-bridge (B), extended strand (E), hydrogen-bonded turn (T), which we counted as "strand"
  - bend (S),

– loop/irregular structure (–)

The secondary structure was encoded in a four-dimensional binary vector of length one.

For the sequence based method, the residues were treated separately and in sequence order. The 17 physicochemical properties plus the HSE1, HSE2 and relative accessibility for each of the 20 closest residues sum up to 400 values. The vectors for the SSE assignments were added together into four values. The number of known PTMs (phosphorylation, acetylation or ubiquination) of residues among the 20 closest residues of the investigated cysteine were also used as three separate features. We completed the set of values by the seven features of the cysteine itself, not including the redundant 17 physicochemical properties of the cysteine. For method 2, we added all values for any of the physicochemical properties together into one value for each of the 17 properties to avoid the introduction of a sequential arrangement into the data, as amino acids in a spatial neighborhood are not ordered consecutively. The vectors for the SSE assignments were again also added together into four values. We treated HSE and relative accessibility separately for each residue, resulting in 39 additional features. Together with the values for PTMs and the secondary structure, HSE and relative accessibility for the cysteine itself, this sums up to 70 features. We collected 3D positions of residues from the PDB files of highest resolution. For the imputed dataset, all data, except for the 17 physicochemical properties of the 20 neighboring residues and the PTMs, were imputed. All values were scaled using the StandardScaler function of scikit-learn (Pedregosa et al. 2011). We applied feature selection using the SelectKBest function of scikit-learn (Pedregosa et al. 2011) with the f_classif score function to compute the ANOVA *F*-value of the features and reduce their number to 30. ANOVA is a statistical test for determining whether the means of two or more samples of data came from the same distribution or not. An *F*-test is used here to calculate the ratio between the explained and unexplained variance by a statistical test like ANOVA. Feature selection was performed independently within each cross-validation fold.

Machine learning methods and tools

For modeling and prediction of Cys+, we utilized the following machine learning techniques: SVM (Cortes and Vapnik 1995), RF (Ho 1995) and ET (Geurts et al. 2006). We applied the package scikit-learn 0.22 (Pedregosa et al. 2011) in Python version 3.7. We used the support vector classification (SVC) function with an RBF (radial basis function) kernel for SVM prediction. For all algorithms, 5-fold cross-validation was performed to avoid overfitting.

To evaluate the algorithms, we used the AUC value for the ROC curve, which displays the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds for a positive prediction. TPR is the ratio between the number of Cys+, which were accurately predicted divided by the full number of Cys+ in the dataset. FPR is the ratio between the number of Cys– which were falsely predicted as Cys+ divided by the full number of Cys– in the dataset. The AUC is the probability that the algorithm will rank a randomly chosen Cys+ higher than a randomly chosen Cys–. A value of one would be a perfect score, while a value of 0.5 signifies a completely random classification. For detailed information on the machine learning methods, see the section *Methods* in the supplemental material.

Acknowledgment: This study was supported by the Deutsche Forschungsgemeinschaft: SFB 815/Z1 (I. W.).
Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This study was supported by the Deutsche Forschungsgemeinschaft: SFB 815/Z1 (I. W.).
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Agip, A., Blaza, J., Bridges, H., Viscomi, C., Rawson, S., Muench, S., and Hirst, J. (2018). Cryo-EM structures of complex I from mouse heart mitochondria in two biochemically defined states. Nat. Struct. Mol. Biol. 25: 548–556.

Åslund, F., Berndt, K., and Holmgren, A. (1997). Redox potentials of glutaredoxins and other thiol-disulfide oxidoreductases of the thioredoxin superfamily determined by direct protein-protein redox equilibria. J. Biol. Chem. 272: 30780–30786.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57: 289–300.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. Nucleic Acids Res. 28: 235–242.

Bleier, L., Wittig, I., Heide, H., Steger, M., Brandt, U., and Dröse, S. (2015). Generator-specific targets of mitochondrial reactive oxygen species. Free Radic. Biol. Med. 78: 1–10.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 4: 1633–1649.

Brandes, N., Schmitt, S., and Jakob, U. (2009). Thiol-based redox switches in eukaryotic proteins. Antioxidants Redox Signal. 11: 997–1014.

Carrell, C., Zhang, H., Cramer, W., and Smith, J. (1997). Biological identity and diversity in photosynthesis and respiration: structure of the lumen-side domain of the chloroplast Rieske protein. Structure 5: 1613–1625.

Charvolin, D., Douliez, J., Marion, D., Cohen-Addad, C., and Pebay-Peyroula, E. (1999). The crystal structure of a wheat nonspecific lipid transfer protein (ns-LTP1) complexed with two molecules of phospholipid at 2.1 a resolution. Eur. J. Biochem. 264: 562–568.

Chen, Y., Lu, C., Su, M., Huang, K., Ching, W., Yang, H., Liao, Y., Chen, Y., and Lee, T. (2015). dbSNO 2.0: a resource for exploring structural environment, functional and disease association and regulatory network of protein S-nitrosylation. Nucleic Acids Res. 43: 503–511.

Chouchani, E., Hurd, T., Nadtochiy, S., Brookes, P., Fearnley, I.M., Lilley, K., Smith, R., and Murphy, M. (2010). Identification of S-nitrosated mitochondrial proteins by S-nitrosothiol difference in gel electrophoresis (SNO-DIGE): implications for the regulation of mitochondrial function by reversible S-nitrosation. Biochem. J. 430: 49–59.

Chouchani, E., Methner, C., Nadtochiy, S., Logan, A., Pell, V., Ding, S., James, A., Cochemé, H., Reinhold, J., Lilley, K., et al. (2013). Cardioprotection by S-nitrosation of a cysteine switch on mitochondrial complex I. Nat. Med. 19: 753–759.

Chouquet, A., Paidassi, H., Ling, W., Frachet, P., Houen, G., Arlaud, G., and Gaboriaud, C. (2011). X-ray structure of the human calreticulin globular domain reveals a Peptide-binding area and suggests a multi-molecular mechanism. PloS One 6: e17886.

Consortium, T. U. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47: D506–D515.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Mach. Learn. 20: 273–297.

Crooks, G., Hon, G., Chandonia, J., Steven, E., and Brenner, S. (2004). WebLogo: a sequence logo generator. Genome Res. 14: 1188–1190.

Fiedorczuk, K., Letts, J., Degliesposti, G., Kaszuba, K., Skehel, M., and Sazanov, L. (2016). Atomic structure of the entire mammalian mitochondrial complex I. Nature 538: 406–410.

Fomenko, D., Xing, W., Adair, B., Thomas, D., and Gladyshev, V. (2007). High-throughput identification of catalytic redox-active cysteine residues. Science 315: 387–389.

Gandhi, P., Page, M., Chen, Z., Bush-Pelc, L., and Di Cera, E. (2009). Mechanism of the anticoagulant activity of thrombin mutant W215A/E217A. J. Biol. Chem. 284: 24098–24105.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. 63: 3–42.

Greco, T., Hodara, R., Parastatidis, I., Heijnen, H., Dennehy, M., Liebler, D., and Ischiropoulos, H. (2006). Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. Proc. Natl. Acad. Sci. U. S. A. 103: 7420–7425.

Groitl, B. and Jakob, U. (2014). Thiol-based redox switches. Biochim. Biophys. Acta 1844: 1335–1343.

Guo, R., Zong, S., Wu, M., Gu, J., and Yang, M. (2017). Architecture of human mitochondrial respiratory megacomplex $I_2III_2IV_2$. Cell 170: 1247–1257.

Hess, D., Matsumoto, A., Kim, S., Marshall, H., and Stamler, J. (2005). Protein S-nitrosylation: purview and parameters. Nat. Rev. Mol. Cell Biol. 6: 150–166.

Ho, T. (1995). Random decision forests. In: *Proceedings of the 3rd international conference on document analysis and recognition*, IEEE, Montreal, Quebec, Canada.

Hohoff, C., Borchers, T., Rustow, B., Spener, F., and van Tilbeurgh, H. (1999). Expression, purification, and crystal structure determination of recombinant human epidermal-type fatty acid binding protein. Biochemistry 38: 12229–12239.

Holmgren, A. (1989). Thioredoxin and glutaredoxin systems. J. Biol. Chem. 264: 13963–13966.

Hurd, T., Prime, T., Harbour, M., Lilley, K., and Murphy, M. (2007). Detection of reactive oxygen species-sensitive thiol proteins by redox difference gel electrophoresis: implications for mitochondrial redox signaling. J. Biol. Chem. 282: 22040–22051.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.

Kozlov, G., Pocanschi, C., Rosenauer, A., Bastos-Aristizabal, S., Gorelik, A., Williams, D., and Gehring, K. (2010). Structural basis of carbohydrate recognition by calreticulin. J. Biol. Chem. 285: 38612–38620.

Larosa, V. and Remacle, C. (2018). Insights into the respiratory chain and oxidative stress. Biosci. Rep. 38: 1–14.

Lin, C., Lin, K., Yang, C., Chung, I., Huang, C., and Yang, Y. (2011). Protein metal binding residue prediction based on neural networks. Int. J. Neural Syst. 15: 71–84.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18: 50–60.

Marino, S. and Gladyshev, V. (2009). A structure-based approach for detection of thiol oxidoreductases and their catalytic redox-active cysteine residues analyzing amino acid and secondary structure composition of the active site and its similarity to known active sites containing redox Cys and calculating accessibility, active site location, and reactivity of Cys. PLoS Comput. Biol. 5: 1–13.

Marino, S. and Gladyshev, V. (2010). Structural analysis of cysteine S-nitrosylation: a modified acid-based motif and the emerging role of trans-nitrosylation. J. Mol. Biol. 395: 844–859.

Marino, S. and Gladyshev, V. (2012). Analysis and functional prediction of reactive cysteine residues. J. Biol. Chem. 287: 4419.

Martìnez-Acedo, P., Núñez, E., Gómez, F., Moreno, M., Ramos, E., Izquierdo-Álvarez, A., Miró-Casas, E., Mesa, R., Rodriguez, P., Martínez-Ruiz, A., et al. (2015). A novel strategy for global analysis of the dynamic thiol redox proteome. Mol. Cell. Proteomics 11: 800–813.

Passerini, A. and Frasconi, P. (2004). Learning to discriminate between ligand-bound and disulfide-bound cysteines. Protein Eng. Des. Sel. 17: 367–373.

Passerini, A., Punta, M., Ceroni, A., Rost, B., and Frasconi, P. (2006). Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins 65: 305–316.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12: 2825–2830.

Pell, V., Spiroski, A., Mulvey, J., Burger, N., Costa, A., Logan, A., Gruszczyk, A., Rosa, T., James, A., Frezza, C., et al. (2018). Ischemic preconditioning protects against cardiac ischemia reperfusion injury without affecting succinate accumulation or oxidation. J. Mol. Cell. Cardiol. 123: 88–91.

Poljsak, B., Šuput, D., and Milisav, I. (2013). Achieving the balance between ROS and antioxidants: when to use the synthetic antioxidants. Oxid. Med. Cell Longev 2013: 1–11.

Ray, P., Huang, B., and Tsuji, Y. (2012). Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cell. Signal. 24: 981–990.

Requejo, R., Chouchani, E., James, A., Prime, T., Lilley, K., Fearnley, I., and Murphy, M. (2010). Quantification and identification of mitochondrial proteins containing vicinal dithiols. Arch. Biochem. Biophys. 504: 228–235.

Salmeen, A., Andersen, J., Myers, M., Meng, T., Hinks, J., Tonks, N., and Barford, D. (2003). Redox regulation of protein tyrosine phosphatase 1B involves a sulphenyl-amide intermediate. Nature 423: 769–773.

Sandberg, W. and Terwilliger, T. (1991). Repacking protein interiors. Trends Biotechnol. 9: 59–63.

Sarma, B. and Mugesh, G. (2007). Redox regulation of protein tyrosine phosphatase 1B (PTP1B): a biomimetic study on the unexpected formation of a sulfenyl amide intermediate. J. Am. Chem. Soc. 129: 8872–8881.

Sun, M., Wang, Y., Cheng, H., Zhang, Q., Ge, W., and Guo, D. (2012). RedoxDB - a curated database of experimentally verified protein redox modification. Bioinformatics 28: 2551–2552.

Tanner, J., Parsons, Z., Cummings, A., Zhou, H., and Gates, K. (2011). Redox regulation of protein tyrosine phosphatases: structural and chemical aspects. Antioxidants Redox Signal. 15: 77–97.

Tien, M., Meyer, A., Sydykova, D., Spielman, S., and Wilke, C. (2013). Maximum allowed solvent accessibilites of residues in proteins. PloS One 8: 1–8.

Wirth, C., Brandt, U., Hunte, C., and Zickermann, V. (2016). Structure and function of mitochondrial complex I. Biochim. Biophys. Acta 1857: 902–914.

Zhu, J., Vinothkumar, K., and Hirst, J. (2016). Structure of mammalian respiratory complex I. Nature 536: 354–358.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/hsz-2020-0321).