Machine Learning Prediction and Statistical Analysis of Redox Modifications in Proteins

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von

Marcus D. Keßler
aus Heidelberg

Frankfurt 2021

vom Fachbereich 12 der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:

Gutachter:

Datum der Disputation:

Contents

Preliminary remarks

(1)

A study was published during the course of this work, which is not part of this dissertation. Parts of the sections "Feature Extraction and Preprocessing" and "Redox Modifications" have been adapted from this study: Keßler, M., Wittig, I., Ackermann, J. and Koch, I., 2021. Prediction and analysis of redox-sensitive cysteines using machine learning and statistical methods. Biological Chemistry, 1 (ahead-of-print)

(2)

Any work concerning Hidden Markov Models, structural invariants near cysteines and shapemers of proteins were carried out as part of a project under my supervision. They were part of a Masters thesis by:
Nover, N., 2021. Prediction of Redox Modifications in Proteins using Machine Learning Methods. Masters Thesis, Department of Informatics and Mathematics, Goethe University Frankfurt.

(3)

The source code used in this thesis can be found using the following link: https://github.com/mal099/ExtraRedox

List of Abbreviations

-	Loop/irregular structure or unknown AA
AA	Amino acid
Aliphatic	Aliphatic AA
Alpha	α-helix SSE
ALT	Alternative
Area	Surface area
Aromatic	Aromatic AA
ASA	Accessible surface area
AUC	Area Under the Curve
B	β-bridge
Beta	β-strand SSE
Charge	Charged AA
CTD	Composition, Transition, Distribution
Cys+	Cysteine experimentally shown to be modifiable
Cys−	Cysteine not shown to be modifiable
DIGE	Difference gel electrophoresis
DSSP	Define Secondary Structure of Proteins
E	Extended strand
ET	Extra Trees
Fox0	Forkhead box class O
FPR	False Positive Rate
G	3-helix
GB	Gradient Boosting
H	α-helix
H_2O_2	Hydrogen peroxide
HMM	Hidden Markov Model
HSE	Half Sphere Exposure
I	5-helix
ISF	Isoelectric focusing
K_a	Acid dissociation constant
LC	Liquid-chromatography
Mass	Molecular mass
MitoSNO	Mitochondria-targeted S-nitrosothiol
MS	Mass spectrometry
NADPH	Nicotinamide adenine dinucleotide phosphate
ND3	NADH dehydrogenase 3

List of Abbreviations

NEM	N-ethyl maleimide
NMR	Nuclear magnetic resonance
Non-P	Non-polar AA
Nox4	NADPH oxidase 4
O_2^-	Superoxide
PAGE	Polyacrylamide gel electrophoresis
PDB	Protein Data Bank
pK_a	Ionization constant
Polar	Polar AA
Positive	Positive AA
PSIBLAST	Position Specific Iterated - BLAST
PSIPRED	PSIBLAST-based secondary structure PREDiction
PTM	Post-translational modification
RBF	Radial basis function
RF	Random Forest
RMSD	Root-mean-square deviation
ROC	Receiver Operating Characteristic
ROS	Reactive oxygen species
RSA	Relative surface area
S	Bend
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SEA1	Large solvent exposed area
SEA2	Regular solvent exposed area
SEA3	Small solvent exposed area
Small	Small AA
SNP	Single nucleotide polymorphism
SSE	Secondary structure element
SVM	Support Vector Machine
T	Hydrogen-bonded turn
Tiny	Tiny AA
TPR	True Positive Rate
Turn	Turn SSE
Vol	Volume
WT	Wild type

Zusammenfassung: Reaktive Sauerstoffspezies sind eine Klasse natürlich vorkommender, hochreaktiver Moleküle, die die Struktur und Funktion von Makromolekülen verändern. Dies kann oft zu irreversiblen intrazellulären Schäden führen. Gleichzeitig können sie auch reversible Veränderungen durch posttranslationale Modifikation von Proteinen bewirken, die in der Zelle zur Signalübertragung genutzt werden. Die meisten dieser Modifikationen treten an spezifischen Cysteinen auf. Welche strukturellen und physikalisch-chemischen Eigenschaften zur Sensitivität von Cysteinen gegenüber Redoxmodifikationen beitragen, ist derzeit unklar. Hier habe ich den Einfluss von Proteinstruktur- und Sequenzmerkmalen auf die Modifizierbarkeit von Proteinen und den darin enthaltenen spezifischen Cysteinen mit statistischen und maschinellen Lernmethoden untersucht. Ich fand mehrere starke strukturelle Prädiktoren für Redoxmodifikationen, wie zum Beispiel eine höhere Zugänglichkeit zum Cytosol und eine hohe Anzahl von positiv geladenen Aminosäuren in unmittelbarer Nähe. Ich stellte eine hohe Häufigkeit anderer posttranslationaler Modifikationen wie Phosphorylierung und Ubiquitinierung in der Nähe von modifizierten Cysteinen fest. Die Verteilung von Sekundärstrukturelementen scheint eine wichtige Rolle bei der Modifizierbarkeit von Proteinen zu spielen. Unter Nutzung dieser Eigenschaften erstellte ich Modelle zur Vorhersage des Vorhandenseins von redoxmodifizierbaren Cysteinen in Proteinen, einschließlich des menschlichen mitochondrialen Komplexes I, der natürlichen NKG2E-Killerzellrezeptoren und der proximalen Tubuluszellproteine, und verglich einige dieser Vorhersagen mit früheren experimentellen Ergebnissen.

Abstract: Reactive oxygen species are a class of naturally occurring, highly reactive molecules that change the structure and function of macromolecules. This can often lead to irreversible intracellular damage. Conversely, they can also cause reversible changes through post-translational modification of proteins which are utilized in the cell for signaling. Most of these modifications occur on specific cysteines. Which structural and physicochemical features contribute to the sensitivity of cysteines to redox modification is currently unclear. Here, I investigated the influence of protein structural and sequence features on the modifiability of proteins and specific cysteines therein using statistical and machine learning methods. I found several strong structural predictors for redox modification, such as a higher accessibility to the cytosol and a high number of positively charged amino acids in the close vicinity. I detected a high frequency

of other post-translational modifications, such as phosphorylation and ubiquitination, near modified cysteines. Distribution of secondary structure elements appears to play a major role in the modifiability of proteins. Utilizing these features, I created models to predict the presence of redox modifiable cysteines in proteins, including human mitochondrial complex I, NKG2E natural killer cell receptors and proximal tubule cell proteins, and compared some of these predictions to earlier experimental results.

1 Introduction

Reactive oxygen species (ROS) are metabolic by-products of cellular processes, such as oxidative phosphorylation in mitochondria, or are directly and deliberately produced by enzymes such as NADPH (nicotinamide adenine dinucleotide phosphate) oxidases [48] in response to specific physiological stimuli [69], such as in phagocytic cells, where NADPH oxidases can participate in intracellular and intercellular redox signaling. They can also produce large quantities of superoxide during respiratory bursts, which contribute to the elimination of ingested organisms [78]. Among the various ROS, the comparatively stable hydrogen peroxide (H_2O_2) and its more unstable precursor molecule superoxide (O_2^-) are physiologically most relevant [46]. The superoxide anion is often produced through the leakage of electrons from redox centers, partially reducing oxygen molecules [104]. An imbalance of ROS production and consumption may lead to irreversible modifications in proteins and other macromolecules, including DNA or lipids. This oxidative stress may give rise to negative physiological consequences such as critical cell damage and play a role in cardiovascular, metabolic and neurodegenerative diseases[57].

The cessation of blood flow to the heart or brain often results in hypoxia in portions of the affected organs, resulting in ischemia/reperfusion injury. This is the result of the generation of ROS and RNS (reactive nitrogen species), which have been shown to irreversibly damage the protein complexes of the respiratory chain as well as enzymes of the Krebs cycle [70].

In diabetic patients, hyperglycemia has been shown to lead to a high glucose-dependent production of electron transfer donors, increasing the electron flux through the mitochondrial electron transport chain and the ATP/ADP ratio. This results in a hyperpolarization of the mitochondrial membrane, hindering the transport of electrons at complex III and accu-

mulating electrons at coenzyme Q. This leads to the rapid generation of O_2^-, which is believed to be the main cause of mitochondrial dysfunction in diabetic patients [70].

The major neurodegenerative diseases, such as Alzheimer's, Parkinson's and Huntington's, are all characterized by either misfolded, misprocessed or mutated proteins forming insoluble aggregates and leading to mitochondrial dysfunction. ROS and RNS generated by such malfunctioning mitochondria have been implicated in the loss of neurons and glutathionylation of proteins typical for these diseases, although the exact role of mitochondrial thiols in their development is not yet fully understood and still subject of ongoing research[70].

In recent years, a positive role for redox modifications has increasingly been recognized. They have often been found to have vital biological functions as targeted secondary messengers, acting as a binary or multistate switch [110]. 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 [12]. The majority of functional redox modifications occur with redox-reactive cysteines [105], often leading to structural and/or functional changes to the protein. While many types of redox modifications are reversible by reducing proteins such as thioredoxins and glutaredoxins [5, 54], enabling the formation of non-pathological ROS pathways, others may lead to permanent changes. This may leave the protein damaged or non-functional [88]. More detailed information on the different types of redox modifications can be found in chapter 2.1.

One early example of a protein gaining function through redox modification was OxyR, a peroxide sensitive transcription factor in *E. coli*. OxyR induces antioxidant gene expression after its cysteine is oxidized by H_2O_2 [48]. A role for oxygen sensing has also been found in mitochondria, especially during hypoxia. Under these conditions, production of O_2^-, most likely by complex III, is elevated, which is then quickly converted to H_2O_2 and diffused into the cytosol. Here, it stabilises hypoxia-inducible factor-1α (HIF-1α), leading to the transcription of genes enabling the cellular response against the harmful effects of hypoxia [24]. While many early examples of redox regulation involved stress responses, there is also evidence that redox signaling plays an important role in normal cellular metabolism. Endogenously generated ROS acts as a second messenger to receptor agonists, such as growth factors and hormones, enabling their associated metabolic changes.

They may also modulate the activation of transcription factors, membrane channels and metabolic enzymes as well as control calcium-dependent and phosphorylation signaling pathways, having an important influence on the majority of aspects of cell physiology [108].

Due to the important influence of redox modifications to both benign and harmful intercellular processes, a large variety of case studies and large-scale proteomics investigations have been conducted to identify and classify redox-sensitive proteins, their underlying stimuli and their effects. Examples include the research by Chouchani *et al.* on the inhibition of mouse complex I through reversible S-nitrosylation of Cys39 on the ND3 (NADH dehydrogenase 3) subunit, which has been shown to be exposed under hypoxic conditions. This slows the reactivation of mitochondria during reperfusion of ischemic tissue and stops ROS production in complex I, protecting tissue from oxidative damage and tissue necrosis caused by ROS imbalance. The redox-active cysteine was marked through the use of a mitochondria-selective S-nitrosylating agent, MitoSNO (mitochondria-targeted S-nitrosothiol), and then identified using SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) [21].

Large-scale proteomics studies include the work of Murphy *et al.* into S-nitrosated mitochondrial proteins, vicinal dithiols, ROS-sensitive thiols in general and their implications for mitochondrial function and redox signaling [55, 20, 89]. They identified redox-sensitive cysteines by first targeting them with a fluorescent tag and then detecting the difference in fluorescence between the control and treated groups through redox difference gel electrophoresis.

Bleier *et al.* analyzed generator-specific targets of ROS in rat heart mitochondria through the application of redox fluorescence difference gel electrophoresis analysis, finding that distinctly different subsets of proteins were modified by ROS which had been produced by the main mitochondrial ROS generators complex I and complex III. [9] Martìnez-Acedo *et al.* developed their new GELSILOX (GEL-based Stable Isotope Labeling of OXidized Cys) method, which combines a proteomics protocol with a computational approach to analyze variance at the peptide level, which they used to demonstrate a significant increase in the status of oxidized thiols induced by hypoxia. They were also able to detect thiols that had been redox modified by ischemia-reperfusion and showed that these reactions were no longer present in ischemia-preconditioned animals [76]. More detailed information on the principles of redox proteomics can be found in chapter 2.2.

Several machine learning approaches have been developed to assist researchers in the identification of redox-active cysteines and similar problems, reducing the necessary load of time and effort for experimentation, reporting differing levels of success. These methods are often hindered by several factors. For one, redox cysteines and their environment do not appear to conform closely to any specific recognizable pattern or motif. Additionally, there are many different types of possible redox modifications which may result from a variety of stimuli, further complicating the problem.

One approach developed by Marino & Gladyshev used amino acid (AA) and secondary structure composition, accessibility, active site location and cysteine reactivity. Applying these features, they attempted to predict the presence of thiol oxidoreductases, an extensively studied group of enzymes containing catalytic redox-active cysteines, among proteomes. Testing their method for the proteome of *Saccharomyces cerevisiae*, they were able to identify the majority of known yeast thiol oxidoreductases [73]. In another study, the group analyzed an extensive dataset of S-nitrosylated proteins, testing for the influence of a variety of features which had been shown to facilitate nitosylation in previous studies, such as pK_a, sulfur atom exposure, cysteine conservation or hydrophobicity. The acid dissociation constant pK_a denotes the strength of an acid. This is often thought to play an important role in the oxidization of thiols. The pK_a values used in this study were predicted using the machine learning-based PROPKA [101] tool. Hydrophobic protein surfaces may concentrate lipophilic NO and molecular oxygen, enabling the formation of nitrosylating species close to the cysteine. The conservation of cysteines among different proteins in their sequence may suggest an important regulatory role of the residue. High thiol exposure would facilitate the approach of nitrosylating agents. The researchers found no evidence for the hypothesis that S-nitrosylation sites were defined by cysteine pK_a as predicted by PROPKA, exposure, hydrophobicity or cysteine conservation. They instead discovered that nitrosylation could be predicted by the presence of a distantly situated, exposed acid-base motif [74].

In 2012, Marino & Gladyshev published a review detailing the current understanding of the properties and functions of reactive cysteine residues and the computational methods to analyze them [75]. Focusing first on pK_a and residue exposure, they gave a brief overview of current tools aiding in their prediction. They found much progress and many new insights provided by bioinformatics tools in the analysis of catalytic redox cysteines, metal-binding cysteines, and disulfide bonds, while methods for the investi-

gation of regulatory cysteines, sites of stable post-translational modifications (PTMs) and catalytic non-redox cysteines still appeared lacking. They appeared hopeful about the further advances that could be accomplished due to new experimental methodologies and data, despite continued difficulties due to the complex nature of the problem of the identification, categorization and comprehension of redox thiols and the necessary conditions leading to redox modification.

Passerini & Frasconi have developed an approach applying the support vector machine algorithm on windows of multiple alignment profiles to differentiate between cysteines involved in ligand binding and cysteines forming disulfide bridges. They compared their method to predictions based on PROSITE pattern hits. PROSITE is a database for the functional characterization and annotation of proteins, consisting of entries that show detailed descriptions of protein families, domains and functional sites, as well as amino acid patterns and profiles in them. Using this approach, they were able to find the majority of relevant PROSITE patterns, but were also able to detect signal in the profile sequence PROSITE was not sensitive to [83]. In a later study, they conceived an approach to predict whether cysteine exists in a free state, metal bound, or participate in disulfide bridges. The method uses a two-stage approach consisting of a support vector machine in the first stage and a bidirectional recurrent neural network in the second stage. They utilized only sequence information in the form of 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 other state-of-the-art methods [84].

1.1 Motivation

Traditionally, redox modifiable cysteines have been identified using biochemical characterization of proteins. In recent years, much progress has been accomplished using redox-DIGE (difference gel electrophoresis) and ICAT (isotope coded affinity tag) methods to streamline experimentation and identify large quantities of reactive thiols. Despite these advances, the experimental verification of the redox modifiability of cysteines remains a costly and time-intensive process to this day. Biochemical methods tend to also be insensitive to proteins with low abundance, such as transcription factors, leading to an experimental bias towards more abundant target proteins, such as enzymes

or ribosomal factors [100]. To aid researchers in the process of identifying redox active cysteines and reduce the required experimental workload, I developed a novel application of a machine learning approach, which may be able to predict redox modifiable cysteines. I also hope to expand our understanding of the necessary conditions and structural properties that facilitate redox modification of cysteines.

In the last years, multiple computational methods for the prediction of cysteine disulfides have been successfully developed. Methods for the prediction of redox cysteines, on the other hand, have been less effective. In their own study on the characterization of thiol oxidoreductases, enzymes containing catalytic redox-active Cys residues, Fomenko *et al.* [40] wrote that the "[i]dentification and characterization of thiol oxidoreductases is challenging because of high divergence of protein families that represent these enzymes and a variety of folds that were adapted by this protein group". Since then, many attempts at the statistical and computational characterization of redox cysteines have been undertaken, yet most approaches have been limited in scope.

In the aforementioned study by Fomenko *et al.* [40], the researchers attempted to identify active sites of thiol oxidoreductases by searching for sporadic Cys/selenocysteine (Sec) pairs in homologous sequences. This approach, while highly accurate, can only identify a small subset of redox modifiable cysteines, and will especially show weaknesses for less well-known proteins where fewer sequences are available.

Marino *et al.*[73] analyzed structural features of redox modifiable cysteines in thiol oxidoreductases, such as amino acid and secondary structure composition, calculating accessibility, active site location, and reactivity of the cysteine, developing an algorithm for their identification based on these characteristics. Sanchez *et al.*[94] developed a classifier based on the distance of the active cysteine to the nearest cysteine sulfur atom, its solvent accessibility, and its pK_a. Both approaches seemed successful, but were only based on data from a small number of redox active cysteines (75 in the former and 161 in the latter case), limiting their scope and robustness.

Recently, Sun *et al.*[100] used the large RedoxDB dataset to develop a classifier based on the support vector machine algorithm to predict redox sensitive cysteines, utilizing only sequence data in order to offer the broadest possible application, even when structural data was not available. They focused on the features of sequential distance of the active cysteine to nearby cysteines, a Position-Specific Scoring Matrix (PSSM) profile, solvent accessi-

bility, physicochemical properties and predicted secondary structure of flanking residues.

My approach applied statistical as well as machine learning methods to computationally find and highlight redox modifiable cysteines and proteins, combining structural and sequence data from the RedoxDB dataset. By utilizing a large dataset and additionally considering the structure of the proteins, I wanted to achieve better results than comparable approaches that use only amino acid sequences. I utilized features like physicochemical properties, amino acid accessibility, Half Sphere Exposure (HSE) and secondary structure elements (SSEs), combining them with the new approach of so-called "shapemers" based on rotation-invariant structural characteristics. When structural data was not available, I utilized prediction methods and imputation to fill in the gaps for the features. I applied various supervised machine learning algorithms, comparing the quality of their models, both for the prediction of redox-sensitive cysteines and proteins. I compared properties of cysteines and proteins that had been experimentally shown to be modifiable (hereafter referred to as Cys+) with those suspected to not be modifiable (hereafter referred to as Cys−) to further researchers' understanding of the structural and physicochemical properties influencing the redox modifiability of cysteines as part of proteins and polypeptides.

I applied my approach to several use cases. I predicted the redox modifiability of the cysteines of human mitochondrial complex I, the largest protein complex in the respiratory chain responsible for the oxidization of NADH produced mainly by the Krebs cycle. I compared my predictions to prior experimental results from the literature to see how well the different algorithms would perform for an unknown real-life example. I produced new predictions for the cysteines of a proximal tubule cell dataset which is currently being used in experimental research by another research group. I used the models to compare different single nucleotide polymorphism variants of the NKG2E Natural Killer Cell Receptor to better understand the impact of small mutations on redox modifiability. The method was also used to make novel predictions of the sequence position of redox modifiable cysteines in proteins experimentally verified to contain at least one such cysteine.

2 Biological Foundations

2.1 Redox Modifications

Aerobic metabolism leads inevitably to the production of ROS [48], including hydrogen peroxide (H_2O_2) and its precursor molecule superoxide (O_2^-), which can lead to thiol oxidation, see Figure 1.

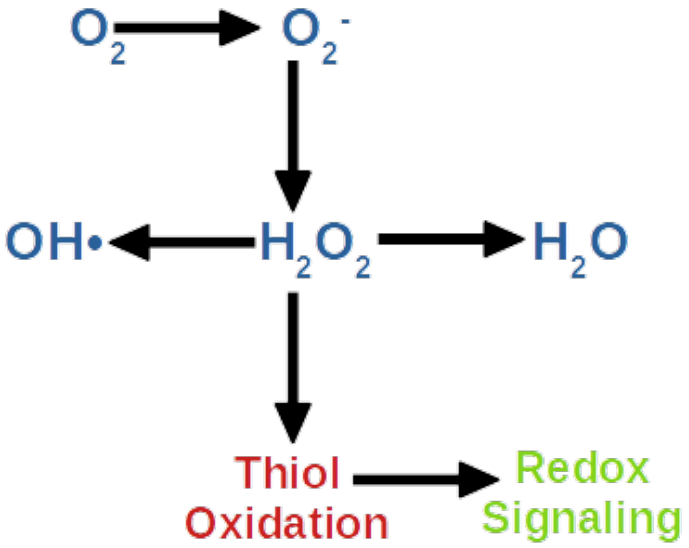

Figure 1: The initial ROS formed within cells is often superoxide. It is however not the primary redox signal, as it is quickly converted to hydrogen peroxide, which will either lead to thiol oxidation or be regulated by enzymes such as peroxiredoxin, which forms a sulfenic acid intermediate while reducing hydrogen peroxide to water. The active site cysteine is then reduced and regenerated in a manner specific to the type of peroxiredoxin [48]. Peroxiredoxin may also serve as an oxidant receptor or sensor, which has been observed in yeast for the activation of the transcription factor Yap1 by the peroxiredoxin-like protein GPx3. The reaction between GPx3 and ROS forms a sulfenic acid, which may then form an intermolecular disulfide bridge with Yap1. Interaction with a second cysteine produces an intramolecular disulfide bridge, recycling GPx3 [108]. Figure adapted from Sullivan *et al.* [98].

A number of stimuli can cause an imbalance of ROS production and con-

sumption. Examples are hypoxia, ischemia/reperfusion injury, inflammation, environmental pollution, strenuous exercise, smoking, nutrition, and psychological stress [87, 86]. An elevation of the concentration of ROS may lead to oxidative stress, damage of virtually all types of macromolecules, such as DNA, proteins and lipids, and apoptosis, see Figure 2.

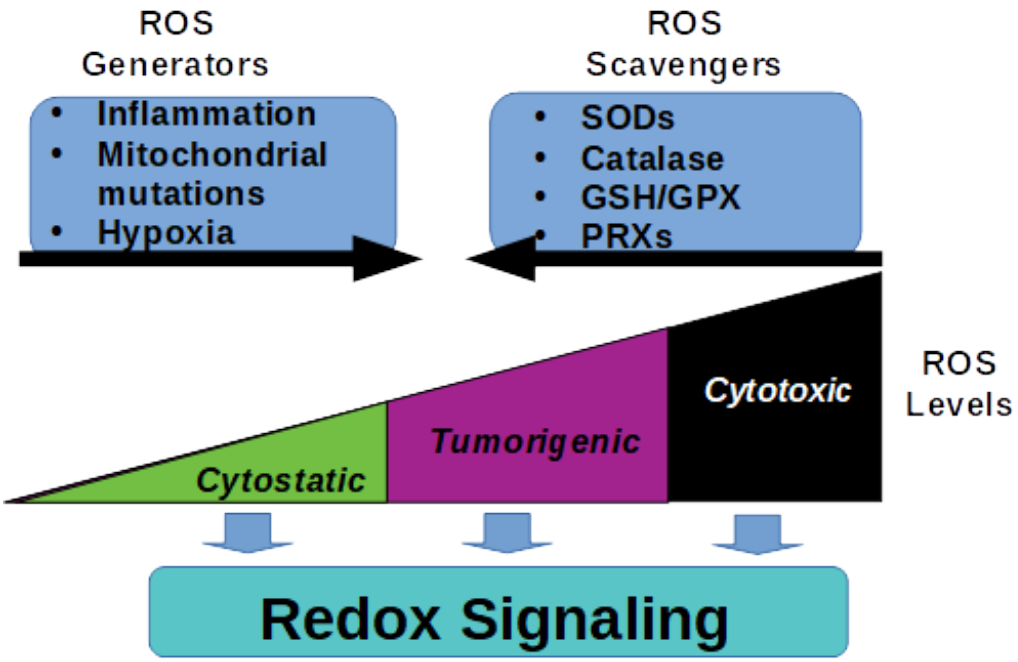

Figure 2: ROS generators increase the amount of ROS in the cell, while ROS scavengers serve to downregulate it. High levels of ROS may lead to tumorigenic or even cytotoxic conditions in the cell. ROS signaling may happen both at lower and higher levels of ROS concentration. Figure adapted from Sullivan *et al.* [98].

Despite their well-known negative effects, ROS have vital biological functions as targeted secondary messengers. In order to keep the amount of ROS in the cytosol at a productive level permissible of regular cellular function, a number of conserved enzymatic and non-enzymatic systems are employed in both prokaryotic and eukaryotic organisms that detoxify excess ROS as well as prevent and repair oxidative damage. Examples of such systems include transcriptional changes mediated by oxidative modification of transcription factors, such as the expression of catalase, peroxiredoxin, thioredoxin and

glutaredoxin by the bacterial peroxide sensor OxyR. Other examples include the activation of stress-specific chaperones and the ROS-mediated change in metabolic pathways, from energy production towards NADPH generation [48].

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 [12]. 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 irreversible oxidation to sulfinic or sulfonic acid, resulting in changes in structure and/or function of the protein [88], see Figure 3. Oxidation to sulfenic acids and the formation of disulfide bonds are reversible by reducing proteins such as thioredoxins and glutaredoxins [5, 54]. Reversibility is a necessary precondition for redox modified cysteines to function in non-pathological pathways. Free thiols may also reversibly form nitrosothiols (SNO) or sulfhydrated thiols through S-nitrosylation with nitric oxide and S-sulfhydration with hydrogen sulfide, respectively. Sulfenic acids may form glutathionylated thiols or sulfenamides.

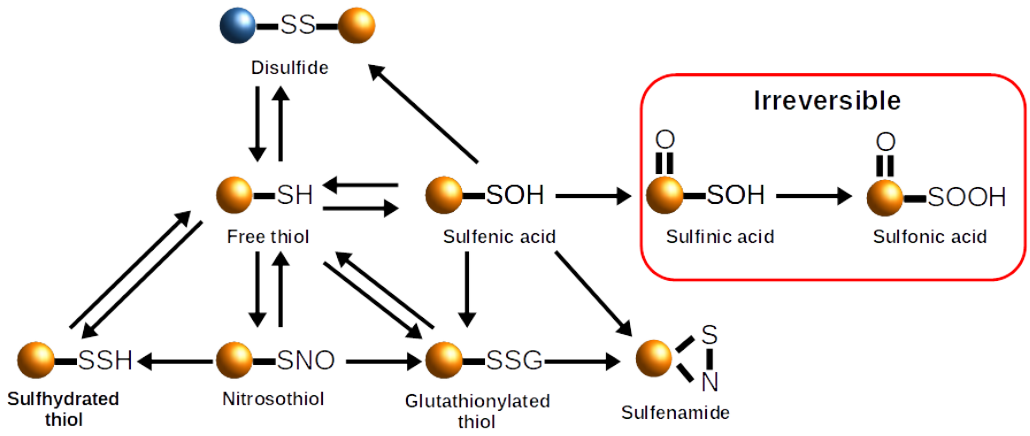

Figure 3: Different types of cysteine modifications by ROS. After reacting with hydrogen peroxide, redox-sensitive thiol groups (SH) will form sulfenic acid (SOH). The sulfenic acid may react with nearby cysteine thiols, forming disulfide bonds (SS). Alternatively, they may form glutathionylated thiols or sulfenamides. Free thiols may also undergo protein S-nitrosylation by reacting with nitric oxide into nitrosothiols (SNO), or S-sulfhydration by reacting with hydrogen sulfide into sulfhydrated thiols. All of the aforementioned modifications are reversible. Further oxidation of sulfenic acids to sulfinic acids or sulfonic acids is generally irreversible. Figure adapted from Chung *et al.* [23].

Recent studies suggest a role of redox signaling in regulating the function of mitochondria through the redox modification of specific cysteines of respiratory chain complexes [24, 32]. Mitochondrial redox balance may also be controlled by the generation of ROS by respiratory chain complexes, especially complex I and complex III, through reversible and irreversible redox modification of specific target proteins involved in both redox signaling and pathophysiological processes [32]. The importance of mitochondrial redox signaling has been shown in cases such as the severely myopathic skeletal muscle-specific COX15 knockout (KO) mice which were crossed with AOX-transgenic mice [31]. In these mice, complex III and complex IV of the respiratory chain are bypassed by alternative oxidases (AOXs) by transferring electrons from coenzyme Q directly to O_2. This lead to decreased ROS production, which impaired AMPK/PGC-1α signaling and PAX7/MYOD-dependent muscle regeneration, decreasing the lifespan of the mice. Treat-

ment with the antioxidant N-acetylcysteine had a similar effect on KO mice.

Significant overproduction of ROS by isolated mitochondria results either when mitochondria are not producing ATP in sufficient quantities and thus possess a high protonmotive force and a reduced coenzyme Q pool, or when there is a high NADH/NAD+ ratio in the mitochondrial matrix [80]. It may be possible to prevent oxidative damage through redox regulation of respiratory chain activity and S-nitrosylation of complex I [32].

2.2 Redox Proteomics

Redox proteomics is a field that aims to identify redox modified proteins and investigate the extent and location of oxidative modifications in them. Since their direct identification through mass spectrometry (MS) is complicated by their instability and low abundance, a large array of tools have been developed for their detection, enrichment and quantification.

The first step for the detection of oxidative modifications is to stabilize the redox status of the sample to avoid the creation of artifacts during sample preparation, either through rapid protonation and precipitation of proteins with trichloracetic acid or by blocking free thiols with reagents such as N-ethyl maleimide (NEM) or iodoacetamide (IAM). Reversible redox modifications are then reduced, e.g. by dithiothreitol or tris(2-carboxyethyl)phosphine (TCEP). A secondary thiol label is applied to mark reversible redox modifications of cysteines, while irreversible modifications remain unlabeled, see Figure 4. Now, the sample can be analyzed using either gel-based methods or pure liquid-chromatography (LC)-MS.

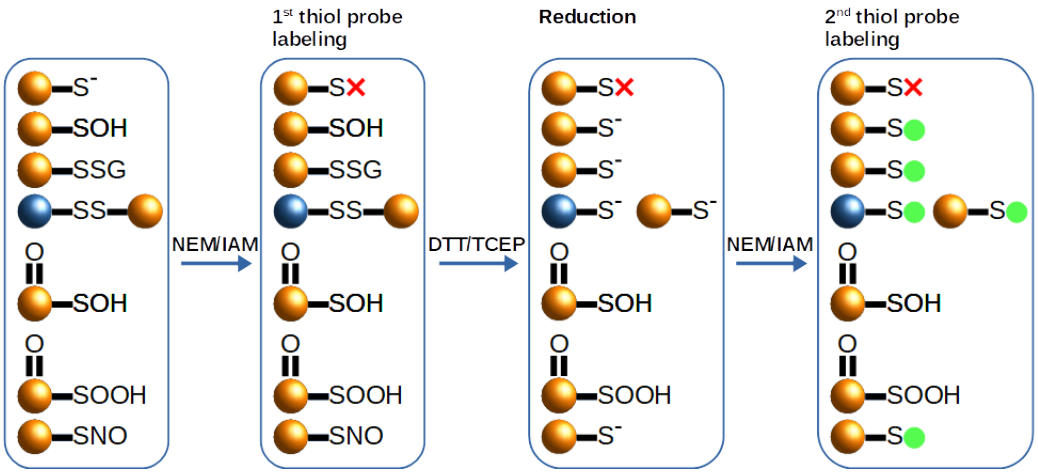

Figure 4: Redox modified cysteine labeling. After proteins are extracted from tissue or cells, the redox state is stabilized by labeling the free thiols with N-ethyl maleimide (NEM) or iodoacetamide (IAM) (red x). Reversible redox modifications are reduced through the application of dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP). Reversible redox modifications are marked using a second label (green dot). Irreversible redox modifications remain unlabeled. Figure adapted from Dröse *et al.* [32].

In gel-based methods, redox modifications are first labeled using fluorescent dyes. Proteins are then separated by polyacrylamide gel electrophoresis (PAGE). Traditionally, 2D isoelectric focusing or sodium dodecyl sulphate (IEF/SDS) gels are used to separate proteins according to their isoelectric point in the first dimension, then according to molecular mass in the second dimension, providing a detailed mapping of redox modifications in the proteome. Recently, newer methods, such as redox difference gel electrophoresis (redox-DIGE), have been developed, which are used to differentially label proteins to identify ROS specific targets.

Pure LC-MS based redox proteomics distinguishes between reduced and oxidized thiols qualitatively after alkylation with reagents such as NEM. Identification of generator-specific thiols requires quantitative approaches. Cysteines are labeled using an isotopically light probe in one sample, and with an isotopically heavy version in the other. After analysis with LC-MS, the ratios of signal intensities of differentially tagged thiol containing peptide pairs can be quantified to determine the redox state or the differential oxidation state under two conditions in the two samples [32].

OxICAT, another method for the detection of redox modifiable cysteines, uses the isotope-coded affinity tag (ICAT) technology. The ICAT reagent consists of the thiol-trapping IAM-moiety (iodoacetamide), a cleavable biotin affinity tag, and a 9-carbon linker, which exists in an isotopically light ^{12}C-form and an isotopically heavy ^{13}C-form. Free thiols are first irreversably labeled with ICAT after denaturation. Reversible redox modifications are then reduced using TCEP, so the thiols can be tagged with heavy ICAT. Using LC and MS, it is now possible to not only detect the redox status, but also to quantify the isotope ratio between modified and unmodified thiols of any particular peptide [32, 67, 6].

2.3 Amino Acid Sequence and Distance

Conservation of AA sequence has been found to often be of structural or functional importance in proteins [43, 90]. Residues in the immediate environment of active sites as well as charge-charge interactions play a prominent role in the value of the acid dissociation constant pK_a, which has been linked to redox modifications of cysteines [91]. There has also been research showing a predisposition towards basic and acidic residues regulating S-nitrosylation and denitrosylation by altering thiol nucleophilicity, and it has been established that deprotonation of thiol to nucleophilic thiolate can be suppressed and enhanced, respectively, by neighboring acidic and basic groups. Thiolates are known to play a role in the formation of many types of redox modifications of cysteines, such as sulfenic acids, disulfide bridges [50] and S-nitrosylation [51]. Hydrophobic pockets have additionally been suggested to present favorable conditions for S-nitrosylation. Such motifs may appear both in the sequence as well as spatial neighborhood [47, 51].

One of the largest studies on the subject has systematically investigated 4165 S-nitrosylation sites within 2277 proteins, mapping them to PDB (Protein Data Bank) structures resulting in information of spatial amino acid composition, solvent-accessible surface area, spatially neighboring AAs, and side chain orientation for 298 substrate cysteine residues. While the researchers found no significant motif surrounding the sites of S-nitrosylation, the abundance of positively charged and hydrophilic AAs were enriched, while the hydrophobic AA cysteine was depleted [19].

Similar investigations were undertaken for the functional sites of sulfenic acids, where it was shown that, in functional site signatures, the frequency of charged residues around cysteines modifiable to sulfenic acid were signifi-

cantly smaller than around other cysteines [93].

There is currently still much disagreement on the importance of such enriched or depleted areas of specific residues. In a different study, researchers found that proximal acid-base motif or cysteine pK_a could not define the specificity of S-nitrosylation. They instead proposed a revised acid-base motif located up to 8 Å from the cysteine with exposed charged groups [74]. The pK_a values used in this study were predicted by PROPKA [101], and it may be possible that experimentally verified pK_a values under physiologically relevant conditions may show a higher correlation between cysteine pK_a and redox modifiability.

2.4 Secondary Structure

The secondary structure of a protein describes the three-dimensional form of small repeating sub-segments of the protein. Several common SSEs have been identified, the most common being α-helices and β-sheets. Secondary structure has been most commonly defined by the hydrogen bonds between the amino hydrogen and the carboxyl oxygen in the peptide backbone. The DSSP (Define Secondary Structure of Proteins) [60] algorithm is often used to assign secondary structure classification to proteins.

Secondary structure can significantly affect the general structure and function of both active sites and the protein in general. It has been shown that the structures of α-helices tend to be more robust to changes in the amino acid sequence than β-strands, due to the relatively higher number of inter-residue contacts they possess. Thus, they may accumulate sequence changes more rapidly than strands within the same domain. Both helices and strands tend to be more robust than less ordered coil regions [1].

Regarding redox modifications, researchers have found a preference for both loop and α-helix regions within 6 Å of the active cysteine of thiol oxidoreductases. β-strands were found to be less common [73]. Another study investigated the role of SSEs both upstream and downstream of the redox-active cysteine. A heightened number of β-strands was found upstream of the cysteines, while α-helices were most commonly found downstream. The active cysteine itself was often present in loop regions [41]. A larger study of more than 1500 redox-active cysteines found an over-representation of coil regions in their vicinity, while helices appeared depleted [100].

2.5 Acid Dissociation Constant

The acid dissociation constant(pK_a) is a method often used to indicate the strength of an acid. It denotes the negative base-10 logarithm of the acid dissociation constant (K_a) of a solution:

$$HA \rightarrow A^- + H^+$$

$$K_a = \frac{[A^-][H^+]}{[HA]}$$

$$pK_a = -log_{10}(K_a)$$

where quantities in square brackets represent the concentrations of the species at equilibrium.

The pK_a for the unperturbed cysteine thiol in aqueous solution is commonly accepted to be between 8.3 and 8.8, depending on the measurement [79, 63, 91]. However, the pK_a values of cysteines in proteins may vary significantly depending on local environmental factors, such as the presence of threonine and other polar or charged amino acids, hydrogen bonds or secondary structure. It has been shown that pK_a of some modifiable cysteines are influenced only by backbone features [93]. Low pK_a thiols in proteins are often called reactive cysteines, and low pK_a is thought to be one key factor in oxidization susceptibility, as the accepted mechanism for the generation of sulfenic acids through H_2O_2-mediated oxidation involves initial cysteine deprotonation, suggesting a lowered thiol pK_a. A study calculating cysteine pK_a using the MEAD multiflex package found a shift in pK_a from 8.14 for a control set to 6.9 for a set of modifiable cysteines [93]. Yet, there is a 10^6-fold difference in reaction rate constants between peroxiredoxin-2 (pK_a=5-6) and PTP1B (pK_a=5.4) [97], despite similar dissociation constants. It appears that a low pK_a alone is not sufficient to explain the redox activity of thiols. Other factors affecting nucleophilicity, such as solvation, steric hindrance, hydrogen bonding and the formation of cyclic transition states may play an important role [50]. Some redox modifiable cysteines displayed unusually high pK_a values. One explanation may be a possible modulation of pK_a values through conformational changes and local flexible loops, which is not captured by the static strucural models underlying pK_a calculation [93]. There have also been studies reporting little discernible link between calculated pK_a using PROPKA and propensity for S-nitrosylation [74], suggesting other structural signifiers may serve as better predictors than pK_a, at least

when calculated using static structures and ML methods.

2.6 Residue Accessibility

Proteins are known to interact with their environment mainly through their solvent exposed surface. Due to this fact, the exposure of a given residue can often act as a predictor to its reactivity [36]. This quality is most often described by the accessible surface area (ASA), although other descriptors, like half sphere exposure, exist. ASA is typically calculated using the "rolling ball" method, where the accessible surface is created by tracing the center of a small probe sphere, often with a radius of 1.4 Å, as it rolls along the van der Waals surface of the biomolecule. This value was converted to the relative surface area (RSA). RSA is defined here as the accessible surface area divided by the maximum accessible surface area as defined by Tien *et al.* [103].

The HSE is defined as the number of C_α-atoms of amino acid neighbors within two half spheres of a chosen radius, generally around 10 Å, around the C_α-atoms of the central amino acid [96], see Figure 5.

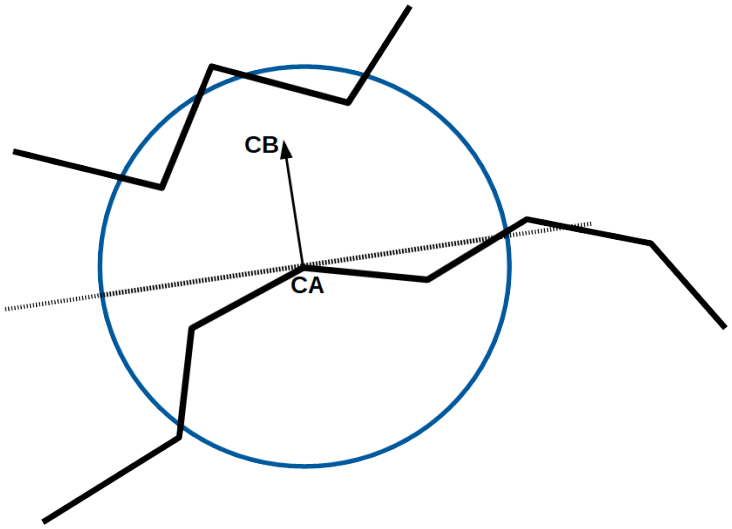

Figure 5: Half sphere exposure can be calculated by counting the number of C_α-atoms in two half spheres of a certain radius, shown here as a blue circle, around a central C_α atom, marked here as CA. The thick black line shows amino acid sequences, with edges denoting C_α-atoms. Dashed line shows the border between the two half spheres, perpendicular to the vector between the C_α- and C_β-atoms.

From a physicochemical point of view, cysteine residues can have varying properties depending on their levels of solvent exposure. While a buried cysteine may possess hydrophobic features due to the hydrophobic effects of amino acid packing inside the protein, interactions with H-bond partners of highly exposed cysteines may lead to considerable polarization, decreasing its pK_a and increasing its reactivity.

Exposed cysteines have been estimated to have a pK_a value very close to the physiological pH when compared to other titratable amino acids. This may lead exposed cysteines to easily switch their ability to function as nucleophiles and experience sudden charge shifts and significant electrostatic changes. High responsiveness to changes in physiological states, leading to a higher or lower ability to interact with the protein environment or other charged molecules, offers one explanation why cysteine residues tend to be comparatively rare, especially on the protein surface, unless employed for a specific function [75].

Cysteine exposure has been investigated for its role regulating S-nitrosylation. One study found that approximately 48% of active cysteines had an ASA value of more than 1.0 Å^2 when a 1.4 Å probe was employed. This value rose to 65% using a probe with a radius of 1.2 Å to account for the small size of the NO molecule. This shows that, while redox-active cysteines tend to be more exposed to the surrounding solvent, high exposure does not appear to be a clear prerequisite for redox activity. The same study found that the exposure of other residues in the vicinity of the active cysteine may also be of importance, as an acid-base motif was discovered often pointing outward with respect to the cysteine sulfur atom. The motif was generally found in conserved regions of solvent accessible surface areas [74].

2.7 Post-translational Modifications

Protein signaling is often controlled by PTMs acting as a binary switch, influencing catalytic activity, turnover and local targeting. Many types of PTMs exist, such as ubiquitination, acetylation and, most prominently, phosphorylation. It has been hypothesized that these types of PTMs may work in conjunction with redox modifications to further regulate protein activity. Researchers have studied the cooperation between redox regulation and phosphorylation/dephosphorylation events both in unicellular and higher organisms with a particular focus on peroxiredoxins. They found evidence for a high prevalence of redox control of metabolism and signaling as well as a high amount of crosstalk between redox-controlled signaling and phosphorylation cascades. Such regulatory cysteines were found both at and more distant to the active site of the enzyme [62].

Further examples of such crosstalk include studies of cyanobacterial protein phosphorylation, where it was found that the Serine/Threonine kinase SpkB could be inhibited by oxidation and reactivated by thioredoxin-catalyzed reduction. Mutants lacking the SpkB kinase were unable to phosphorylate the glycyl-tRNA synthetase β-subunit (GlyS), while purified SpkB could phosphorylate purified GlyS, showing a link between redox regulation through modulation of the cysteine redox state and cyanobacterial phosphorylation. A mutant lacking SpkB kinase showed hypersensitivity to oxidative stress, displaying severe growth retardation or death in response to several types of ROS stimuli [77].

The catalytic activity of of the serine/threonine kinase Aurora A could

be shown to be inhibited by the oxidation of a conserved cysteine residue lying adjacent to a critical phosphorylation site in the activation segment. The redox cysteine was found to be highly conserved among serine/threonine kinases, pointing to an important regulatory mechanism [15].

Activity of signal transducer and activator of transcription 3 (STAT3), a latent transcription factor promoting cell survival and proliferation often found active in cancers, has been reported to be modulated by S-glutathionylation, preventing Janus kinase 2 (JAK2) mediated phosphorylation of Tyr705. While the molecular mechanism of this event is as of yet poorly understood, it has been suggested that glutathionylation may interfere with tyrosine accessibility, hampering its recognition by JAK2.

Similar results were found for other PTMs, such as in the case of Forkhead box class O (FoxO) transcription factors, where it was found that ROS can induce the formation of disulfide complexes with p300/CBP acetyltransferase, and that modulation of FoxO activity by p300/CBP-mediated acetylation depends on the formation of this redox-dependent complex [30].

Evidence for the role of oxidative stress also exists during the degradation of rat myosin heavy chain associated with the activation of a ubiquitin-dependent pathway. Levels of markers for oxidative stress in mice with suspended tails were measured and found to be increased over a period of several days together with an enhancement of protein ubiquitination. Supplementation of antioxidative nutrients decreased not only the levels of antioxidative markers, but also suppressed ubiquitination and fragmentation of the rat myosin heavy chain protein, suggesting an important role of oxidative stress in some forms of ubiquitination [56].

2.8 Mitochondrial Complex I

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 [109]. 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 [66].

Which cysteines are likely targets of such activity could be predicted by my models.

I used data for the *NDUFS1*, *MT-ND3* and *NDUFA2* subunits of mammalian respiratory complex I, based on the structural data from PDB entries 6G2J [2], 6G72 [2], 5LC5 [114], 5LNK [38] and 5XTD [49], as the basis for the prediction of redox modifiable cysteines.

2.9 Proximal Tubule Cells

The kidney is an organ responsible for the removal of waste products produced by metabolism into the urine. It also regulates the acid-base balance, electrolyte concentrations, extracellular fluid volume, and blood pressure. It secretes hormones regulating blood composition and pressure.

As the kindey is a highly complex organ, many of its functions, as well as pathologies concerning those functions, are not yet completely understood. To answer central questions in kidney biology and disease pathogenesis, single-cell transcriptional profiling has been performed, allowing researchers to monitor global gene regulation in thousands of individual cells [82]. As a central question, the role and physiological function of the NADPH oxidase 4 (Nox4) in the kidney has emerged.

Nox4 produces only hydrogen peroxide and is highly expressed in proximal tubule cells, and may be responsible for maintaining normal kidney function through redox signaling. It has been considered a potential pharmacological target for a long time.

To aid in the identification of potential targets, I used machine learning models to perform predictions of redox active cysteines coded by genes which are all expressed by proximal tubule cells identified by aforementioned single-cell profiling, providing a basis for later research on promising candidates for further experimentation.

2.10 NKG2E Natural Killer Cell Receptor

Natural killer cells respond to abnormal cell activity by monitoring levels of MH1 protein expression, which is often disrupted during viral infection, inflammation or neoplastic transformation. Their activity is controlled in part by the binding of NKG2x receptors to CD94. NKG2A forms an inhibitory receptor, while NKG2C and NKG2E serve an activating role [37, 61].

In humans, NKG2E exists in several variants due to single nucleotide polymorphisms (SNPs). I received data for four variants of this proteins for

further analysis to better understand the impact of small mutations on redox modifiability.

3 Methods

3.1 Data

3.1.1 PDB

The Protein Data Bank (PDB, www.rcsb.org) [8] is a database containing 3D structural data of more than 170.000 macromolecular structures, including proteins, DNA and RNA. Knowledge and understanding of structural data of biological macromolecules is necessary in order to comprehend and interpret their role in both disease and health as well as its function in physiological processes.

PDB data is obtained through crystallography, NMR (nuclear magnetic resonance) spectroscopy or cryo-electron microscopy by independent scientists and researchers from all around the world and updated weekly. Protein structures deposited in the PDB are used as the basis for numerous other databases and tools, such as the secondary structure assignment tool DSSP, the protein structural domain classification database SCOP, the protein structure classification database CATH and PDBsum, a database providing at-a-glance overviews of macromolecular structures. PDB files are mainly made up of 3D coordinates of atoms, along with additional data, such as molecular bonds or methodological data. In my research, structural data as well as the resolution of the structure were mainly used.

Each structure contained within the PDB can be uniquely identified via a four-character PDB identifier. Several identifiers may correspond to the same biomolecule, as different structural files may comprise data for different environments, ligands, methods or conformations.

3.1.2 RedoxDB

The RedoxDB (http://biocomputer.bio.cuhk.edu.hk/RedoxDB/index.php) is a manually curated database containing experimentally verified data of proteins with at least one redox modified cysteine. It consists of two datasets. Dataset A consists of 1998 redox proteins with verified positions and modification types of modified cysteines. Dataset B comprises 865

additional redox proteins, but lacks data on the positions of modified cysteines. Many, although not all, of the database entries contain a PDB identifier for structural data of the protein. The database incorporates additional information, including amino acid sequence, gene name, protein function, literature references, links to UniProt and taxonomic data. RedoxDB represents one of the largest repositories for redox modified proteins in existence, making it a useful tool for computational analysis of redox proteomics [99].

3.1.3 UniProt

UniProt is a worldwide database of protein knowledge, incorporating data for over 120 million proteins, including sequence and functional information. Over half a million database entries contain detailed annotation extracted from scientific literature and curated by experts. These annotations are further supplemented through automated rule-based systems.

UniProt contains data in a variety of categories, including protein function, subcellular location, taxonomy, pathology, PTMs, domains and many others. For this study, the presence of PTMs near redox cysteines was used. Other categories, like subcellular location or taxonomy, may prove useful in the future [25].

3.1.4 Proximal Tubule Cell Data

I received a set of proximal tubule cell data for analysis and prediction using machine learning methods. This dataset consists of a list of 122 proteins with 1394 cysteines. PDB entries for 14 of those proteins were found, containing 99 cysteines.

3.1.5 Protein Sets

I used four protein sets to be utilized with different machine learning approaches. The first one contains structurally resolved proteins, whereas the second one was extended by proteins with unknown structure. For protein set 1, 439 redox-active proteins with 644 Cys+ and 1692 Cys− from the latest update of RedoxDB [99] up to May 2020 were collected. I selected only proteins with known structure stored in the PDB [8] and at least one Cys+. To assign SSEs and accessibility to amino acids, I applied the algo-

rithm DSSP [60]. Protein set 1 included 369 mammalian proteins, 25 plant proteins and 45 fungal proteins.

For protein set 2, I considered 1097 additional redox-active proteins listed in the RedoxDB for which no structures were available. I used imputation, the process of replacing missing data with substituted values, to be able to use the full set of features. The imputed structural properties of amino acid residues in proteins, such as, e.g., secondary structure and accessibility, were calculated using the values from protein set 1. I applied the function IterativeImputer with Bayesian Ridge estimator [85]. Combined with model training, I performed imputation separately within each cross-validation fold. By imputation, the dataset was extended by 834 mammalian 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, 55% of randomly chosen Cys− were removed from the dataset . I removed 80% of Cys− from the incomplete, imputed protein set to reduce bias towards Cys− in my model. I balanced the protein sets to consist of 50% of Cys+ for the same reason.

Entries in the UniProt database often possess annotations concerning redox modifications for specific cysteines. I used these annotations to search through the entire UniProt database to supplement the RedoxDB dataset with additional data. I categorized all proteins with the feature type "DISULFID" when coupled with the note "redox-active" as well as the feature type "MOD_RES" when coupled with any of the notes "Cysteine sulfenic acid (-SOH)", "Cysteine sulfinic acid (-SO2H)", "Cysteine sulfonic acid (-SO3H)", "S-nitrosocysteine" or "S-glutathionyl cysteine". Only eukaryotic data in the manually annotated and reviewed Swiss-Prot dataset was considered. Out of 563,972 entries, I found 1149 eukaryotic proteins with known redox modifications. Out of these, 1026 entries had not been categorized in the RedoxDB before, including 2148 redox-active cysteines, see Figure 6. The UniProt data was parsed with a Python script and transformed into the same format used by RedoxDB so that the two datasets could be combined easily. I added proteins from Uniprot to both protein sets while removing duplicates with 90% sequence identity applying the cluster_fast tool from the USEARCH v11 sequence analysis tool [35]. This brought the numbers up to 2678 cysteines and 840 proteins for protein set 1 and 26325 cysteines and 2478 proteins for protein set 2.

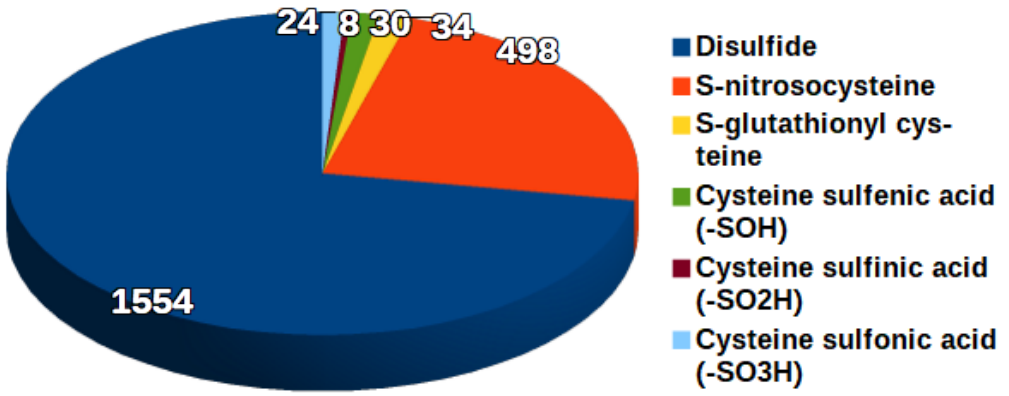

Figure 6: Number and type of redox-active cysteines found in the UniProt database.

Protein set 3 consists of 400 random eukaryotic redox-active proteins from the RedoxDB and 400 random eukaryotic proteins from Uniprot.

Protein set 4 adds 3957 proteins from a list of proteins generated with PISCES [106] to the redox-active proteins of protein set 1, to generate a set containing a mix of redox-active and non redox-active proteins with resolved structures. The PISCES proteins had up to 20% sequence identity, 1.6 Å resolution and a maximum R-factor of 0.25.

3.2 Tools

3.2.1 Define Secondary Structure of Proteins (DSSP)

DSSP [60] is a program mainly designed to standardize secondary structure assignment. It uses PDB files as the basis for its assignments and also contains a database of secondary structure assignments for every PDB entry. The program calculates the most likely secondary structure assignment based on the 3D coordinates of the atoms of the protein, as well as a calculation of the H-bond energy between nitrogen atoms and oxygen atoms of the backbone. All hydrogen atoms present in the PDB file are discarded

before hydrogen atoms are added by optimally placing them 1.0 Å from the backbone N in the opposite direction of the backbone C=O bond. The secondary structure is then assigned by using the two best H-bonds for each atom. DSSP contains eight different classes of secondary structures:

- α-helix (H)
- 3-helix (G)
- 5-helix (I)
- residue in isolated β-bridge (B)
- extended strand (E)
- hydrogen-bonded turn (T)
- bend (S)
- loop/irregular structure (-)

DSSP files also contain values for the ASA of each amino acid. The ASA is defined as the residue water-exposed surface in $Å^2$.

3.2.2 PSIPRED

PSIBLAST-based secondary structure PREDiction (PSIPRED) [59, 14] is a method for predicting SSEs in proteins. The predictions use two feed-forward neural networks performing an analysis on output from PSIBLAST (Position Specific Iterated - BLAST).

PSIPRED is able to predict the secondary structure of an unknown amino acid sequence based on known information regarding evolutionarily related proteins. These proteins are found using PSIBLAST. They are used to construct a position-specific scoring matrix, which is then further processed by the neural networks.

The first neural network is fed with a window of fifteen amino acids, adding additional information indicating the position of the window at the C or N terminus. The final input layer contains 315 input features, divided into fifteen groups of 21 features. The network contains a hidden layer of 75 features and three output nodes for the SSEs helix, sheet and coil.

These secondary structure predictions are filtered by the second network. The input layer again contains data for the same window of fifteen amino acids, including features for terminus position and SSE, resulting in 60 features in fifteen groups of four. The network contains another hidden layer of 60 units and again has three output nodes for each SSE. Out of these three nodes, the highest score is chosen as the prediction for the SSE of the central amino acid.

PSIPRED has achieved a Q_3-score of 81.6% for its predictions using stringent evaluation methods [65]. I evaluated PSIPRED together with DSSP and received similar results, with 80.0% of PSIPRED predictions agreeing with DSSP secondary structure assignments (data not shown).

3.2.3 ASAquick

ASAquick [36] is a fast approach for predicting the ASA of the AAs in a single sequence. The program utilizes a neural network to analyze a sequential window of AAs, as well as global features, such as single residue and directional two residue composition as well as sequence length of the protein chain. Based on empirical optimization, this sequential window includes the data of 21 residues. The neural network contains hidden layers of 31 nodes.

Unlike similar predictors, ASAquick forgoes the use of residue mutation profiles generated by multiple sequence alignments to known protein structures. This results in a speed-up of predictions by several orders of magnitude, while retaining a similar accuracy to other predictors. This method is less dependent on sequence similarity and may thus be useful for predicting the ASA of novel proteins.

ASAquick has been trained using ASA data provided with DSSP, making it a good substitute when predicting ASA values for proteins with unknown structure in my investigations.

3.2.4 PROPKA

The protein pK_a prediction tool PROPKA [101] was developed by the Jensen Research Group of the University of Copenhagen and is one of the most commonly used empirical pK_a predictors. Unlike many other tools, its predictions are not based on molecular dynamics or the Poisson-Boltzmann equation, but instead entirely on empirical rules relating to protein structure based on PDB entries. This allows for both accurate and rapid prediction of pK_a values in

proteins.

The most current version of the program is PROPKA 3.4. This new version of PROPKA has a number of advantages over its predecessor, PROPKA 2.0. Specifically, it enables modeling of interactions between ligand groups and implements methods for modeling of both covalently and noncovalently coupled titrational events. In their own testing, the Jensen Group found significant improvements in the root-mean-square deviation (RMSD) values for Asp, Glu, Tyr, Lys and His residues [81]. Despite these advances, studies have suggested that PROPKA 2.0 may enable superior predictions in the case of cysteine residues [91]. Calculations could be further improved by energy minimizing the raw X-ray structures using CHARMM [13]. In this study, I compared the value of the predictions of PROPKA 2.0 to PROPKA 3.4 for the purposes of the prediction of redox-active cysteines. I did not prepare the structures using CHARMM due to time constraints and the large size of the dataset.

3.2.5 ConCavity

I characterized functional sites and surface cavities applying the ConCavity algorithm [16]. ConCavity uses sequence conservation estimates and structure-based methods to predict small molecule binding sites. By combining these two approaches, ConCavity is able to outperform similar tools and has been shown to possess high accuracy at the identification of drug targets as well as ligand binding sites. It can be suspected that redox-active cysteines may be located within similar surface cavities in proteins, since there is evidence that they tend to be situated in more highly accessible areas [75, 74]. I used the ConCavity algorithm on the dataset of redox-active proteins to compare the prediction values between Cys+ and Cys−.

3.2.6 UCLUST

UCLUST is an algorithm that divides sequences into clusters by identity. It is part of the USEARCH [35] software package. Clusters are defined by one representative sequence, also called the centroid. Each sequence that is part of the cluster must have sequence identity above a certain threshold with the centroid. UCLUST is a greedy algorithm which assigns any sequence as a new centroid that can not be assigned to an existing cluster. Consequentially, the order of sequences in an input file matters. I arranged input sequences

by sequence length in descending order. I applied the cluster_fast variant of UCLUST.

3.2.7 Geometricus

Geometricus [33] is an approach that translates the 3D structure of protein fragments into a set of four 3D moment invariants to then discretize these fragments into shapemers, easily comparable and countable vectors. These vectors can then be used for a number of different tasks, such as structure similarity search, unsupervised clustering, supervised machine learning and structure classification.

The fragments used in this approach can either be sequential k-mers, i.e. a sequence fragment of fixed length, or radius-based around a central residue. The k-mer method is generally preferable for describing structures that are sequential in nature, such as α-helices and loops, while the radius-based method tends to excel when longer ranged structural contacts are of importance, such as in the case of β-sheets, PTMs and enzyme active sites.

The 3D moment invariants are computed using the formula for the central moment μ as defined below:

$$\mu_{pqr} = \sum_{i=1}^{c} (x_i - \bar{x})^p (y_i - \bar{y})^q (z_i - \bar{z})^r$$

where $(\bar{x}, \bar{y}, \bar{z})$ is the centroid and (x_i,y_i,z_i) is the coordinates corresponding to the C_α of the i-th residue of the radius or k-mer-based structural fragment. The three second-order rotation invariants O_3, O_4 and O_5 as described by Mamistvalov [71] and the third-order rotation invariant F as described by Flusser [39] are defined below:

$$O_3 = \mu_{200} + \mu_{020} + \mu_{002}$$

$$O_4 = \mu_{200} \cdot \mu_{020} + \mu_{200} \cdot \mu_{002} + \mu_{020} \cdot \mu_{002} - \mu_{110}^2 - \mu_{101}^2 - \mu_{011}^2$$

$$O_5 = \mu_{200} \cdot \mu_{020} \cdot \mu_{002} + 2\mu_{110} \cdot \mu_{101} \cdot \mu_{011} - \mu_{110}^2 \cdot \mu_{002} - \mu_{101}^2 \cdot \mu_{020} - \mu_{011}^2 \cdot \mu_{200}$$

$$\begin{aligned} F = {} & 15\mu_{111}^2 + \mu_{003}^2 + \mu_{030}^2 + \mu_{300}^2 - 3\mu_{102} \cdot \mu_{120} - 3\mu_{021} \cdot \mu_{201} - 3\mu_{030} \cdot \mu_{210} \\ & -3\mu_{102} \cdot \mu_{300} - 3\mu_{120} \cdot \mu_{300} - 3\mu_{012} \cdot (\mu_{030} + \mu_{210}) - 3\mu_{003} \cdot (\mu_{021} + \mu_{201}) \\ & +6\mu_{012}^2 + 6\mu_{120}^2 + 6\mu_{201}^2 + 6\mu_{210}^2 + 6\mu_{102}^2 + 6\mu_{021}^2 \end{aligned}$$

Any structural fragment thus corresponds to a vector of four values, and similar structural features will correspond to similar vectors. These values can be used as a feature for machine learning, or discretized into shapemers,

enabling researchers to easily find important structural features by simply finding the most abundant shapemers.

The moment invariants O_3, O_4, O_5 and F calculated through Geometricus are discretized into shapemers by multiplying them with a resolution parameter m. The resulting value is then rounded down, as seen in the following formula:

$$(O'_3, O'_4, O'_5, F') = (\lfloor m \times ln(O_3) \rfloor, \lfloor m \times ln(O_4) \rfloor, \lfloor m \times ln(O_5) \rfloor, \lfloor m \times ln(F) \rfloor)$$

The resolution parameter m defines the coarseness of the discretization, with higher values leading to a finer separation of the shapemers. We chose a relatively low value of 0.5. We calculated the shapemers for protein set 4.

3.2.8 Machine Learning Tools

I applied the SVC, GradientBoostingClassifier, ExtraTreesClassifier and RandomForestClassifier functions of the package scikit-learn 0.22 [85] in Python version 3.7 for the Support Vector Machine (SVM), Gradient Boosting (GB), Extra Trees (ET) and Random Forest (RF) algorithms. For more details, see Section 3.4.1: Machine Learning Approaches.

All values were scaled using the StandardScaler function of scikit-learn [85], which standardizes features by subtracting the mean and dividing by variance of the feature using the following formula:

$z = (x - u)/s$

where z is the standardized value, x is the original value, u is the mean of the training samples and s is their standard deviation.

I applied feature selection using the SelectFromModel function of scikit-learn [85]. This method utilizes the classification algorithm itself to eliminate all features displaying a lower feature importance than the mean.

3.3 Statistical Methods

I performed statistical analyses of the environment of both Cys+ as well as Cys− to show that the chosen features should theoretically enable the models to make useful predictions. I performed all statistical tests on protein set 1 before preprocessing. The neighborhood of residues is one of the main predictors of post-translational modifications and catalytic activity of residues [10]. Good results have been obtained for the prediction of disulfide

bridges [83]. No specific amino acid sequence motifs around Cys+ have been found [74, 47]. For all features, the set of values for Cys+ in either the sequence or Euclidean neighborhood of Cys+ was compared to the set of values in the neighborhood of Cys− by dividing their mean values.

I applied the **Mann-Whitney *U* test** [72], which examines the null hypothesis that randomly selected values X and Y from two populations have equal chance for X > Y as for Y > X, to test for significance. When testing multiple hypotheses, it becomes more likely to observe a seemingly significant result by chance. I corrected for multiple testing using **Bonferroni correction**, which divides the cutoff value for significance by the number of tests to avoid spurious positives.

I used the **Poisson standard deviation** to estimate a lower bound for the error rate, applying the following equation:

$$\sigma = \frac{\sqrt{\overline{A}_{Cys+}}}{\overline{A}_{Cys-}}$$

where σ is the Poisson standard deviation and $\overline{A}_{Cys+}$ and $\overline{A}_{Cys-}$ are the means over the values of one feature for Cys+ and Cys− respectively.

Gini coefficient was applied to assess the quality of data splits performed by tree based machine learning methods. It is calculated according to the formula:

$Gini = 1 - \sum_{i=1}^{C} p_i^2$

where $Gini$ is the Gini coefficient, p_i is the proportion of the i-th class (here Cys+ and Cys−) after the split and C is the total number of classes. An even distribution of two classes would have a value of 0.5, while a completely perfect classification produces a value of 0.

Feature importances of protein set 3 were compared by computing the **ANOVA F-value**. ANOVA is a statistical test for determining whether the means of two or more samples of data came from the same distribution or not. ANOVA was applied as it can be used when the feature data is numeric, while the prediction is categorical. We seek to confirm or reject the null hypothesis that, for the tested feature, the mean value for Cys+ or Cys− is significantly different from the overall mean value. This is accomplished by calculating the sum of squares within samples and between

samples. We then calculate the F-value from the sums. The sum of squares is calculated according to the following formulae:

$SSB = \sum_{j=1}^{m} n \cdot (\overline{g}_j - \overline{X})^2$

and

$SSE = \sum_{i=1}^{n} (x_i - \overline{g})^2$

where SSB is the sum of squares between groups and SSE within groups. m denotes the number of groups, n the number of data points, $\overline{X}$ the average across all data points, $\overline{g}$ the average within a specific group, x_i the value of one observation.
The F-value is calculated according to the formula:

$F = \frac{SSB/(m-1)}{SSE/(n-1)}$

The F-value can then be compared to a minimum value depending on the confidence value to either accept or reject the null hypothesis.

I used a **two sample sequence logo** for a clear graphical representation of the differences between the sequence neighborhoods of Cys+ and Cys−. Two sample t-test was used to check for significance. Bonferroni correction was used again.

3.4 Machine Learning

The aim of machine learning is to utilize algorithms to analyze, categorize and classify data that is too complex for humans alone to comprehend. Using machine learning methods and tools, complex relationships between hundreds of features can be used mathematically to gain new knowledge. A large amount of clean, noiseless data is a necessary precondition for this endeavor. I tried out several algorithms resting on different theoretical foundations to find the one best suited for this problem. All of the methods used are supervised machine learning algorithms for classification and regression, using labeled data, i.e. we know whether the training data comes either from a Cys+ or a Cys−. The algorithm attempts to learn different features corresponding to one of the two groups, and will be able to classify

new data from the test dataset or an unknown dataset. These methods can be contrasted to unsupervised machine learning algorithms, where training data is unlabeled and the algorithm attempts to find categories within the data on its own. I used four different machine learning methods which I will explain in the current chapter.

3.4.1 Machine Learning Approaches

For modeling and prediction of Cys+, I utilized the following machine learning techniques: Hidden Markov Models (HMM) [7], SVM [27], GB [42], RF [52] and ET [45]. I chose these algorithms because they cover a variety of different approaches to the problem of classification while being state of the art methods for dealing with noisy, incomplete or small datasets. It is generally not possible to predict in advance how well any machine learning algorithm will perform on any specific problem due to the complexity of the learning process.

Hidden Markov Models are statistical Markov models, where the underlying system is assumed to be a stochastic process with hidden states and emissions specific to each state. Transitions from one state to another conform to the Markov property, meaning the probability of a transition is only dependent on the most recent state of the model. We used the data from the sequential and 3D environment of cysteines to produce a Profile HMM [34] for Cys+. Here, the hidden states are Cys+ and Cys−. These states depend on observations, or emissions, of the 20 different types of residues found in most living organisms. Instead of observing the transitions between hidden states, we compare the two hidden states to each other. Using the datasets of Cys+ and Cys−, we can calculate an observation probability matrix given the state of the cysteine, i.e. the probability of an observed residue given the sequence position. This means that the HMM is not homogenous, i.e. the emission probabilities change in each step. From these probabilities, we calculate a score for each individual position according to the formula:

$$Score = ln([AA]^i_{Cys+}/[Cys+]) - ln([AA]^i_{Cys-}/[Cys-])$$

where $[AA]^i_{Cys+}$ is the total count of a specific residue at position i in the

neighborhood of Cys+, $[Cys+]$ is the total number of Cys+ in the dataset, equivalently for Cys−. The final score for each cysteine equals the sum of all individual AA scores. A high score is more likely to correspond to Cys+.

Support Vector Machine [27] is a classifier method that attempts to construct a hyperplane to optimally separate data points belonging to different classes in a high-dimensional space. The separating hyperplane is constructed in such a way that it has the highest possible distance, generally referred to as the "margin", to the data points of the training dataset, see Figure 7. New data points can then be mapped to the high-dimensional space and predicted to belong to a class depending on which side of the hyperplane they fall [27].

Originally, the maximum-margin hyperplane algorithm constructed a linear classifier. Later, it was proposed to create nonlinear classifiers using the kernel trick [3]. Here, the data points are mapped using a nonlinear function into the transformed high-dimensional space. While the classifier is still a hyperplane in the transformed space, it may be nonlinear in the original feature space, see Figure 8. Common kernel functions include polynomial and radial basis function (RBF) kernels. I decided to use the RBF kernel for my models, which includes two important parameters: the cost parameter C and the reach parameter γ. C represents a tradeoff between misclassification of training data and the smoothness of the decision surface. A low value may not grasp the complexity of the classification problem, while a high value is likely to lead to overfitting. γ can be seen as the inverse of the radius of influence of a single training example. If this radius is too small, each training sample will only influence itself, leading to overfitting. If the radius is too large, the model will be too constrained, behaving similarly to a linear model.

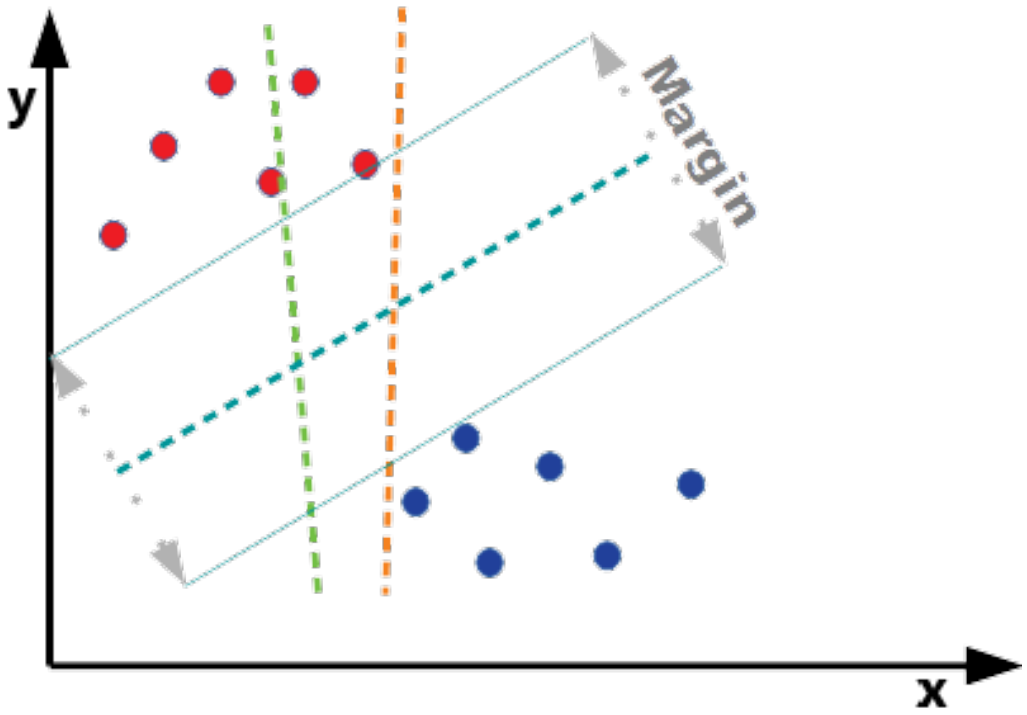

Figure 7: Simplified illustration of three different hyperplanes separating training data with the widest possible margin. Red and blue dots correspond to data points belonging to two different classes, x- and y-axes correspond to arbitrary features. The green hyperplane does not classify all training data points correctly. The orange hyperplane classifies correctly, but only has a small margin to some data points, potentially leading to new data being wrongly classified. The teal hyperplane would be chosen by the SVM, as it separates the training datasets correctly and by a large margin, leading to a high likelihood of correct classification on new data. Adapted from Coqueret & Guida [26].

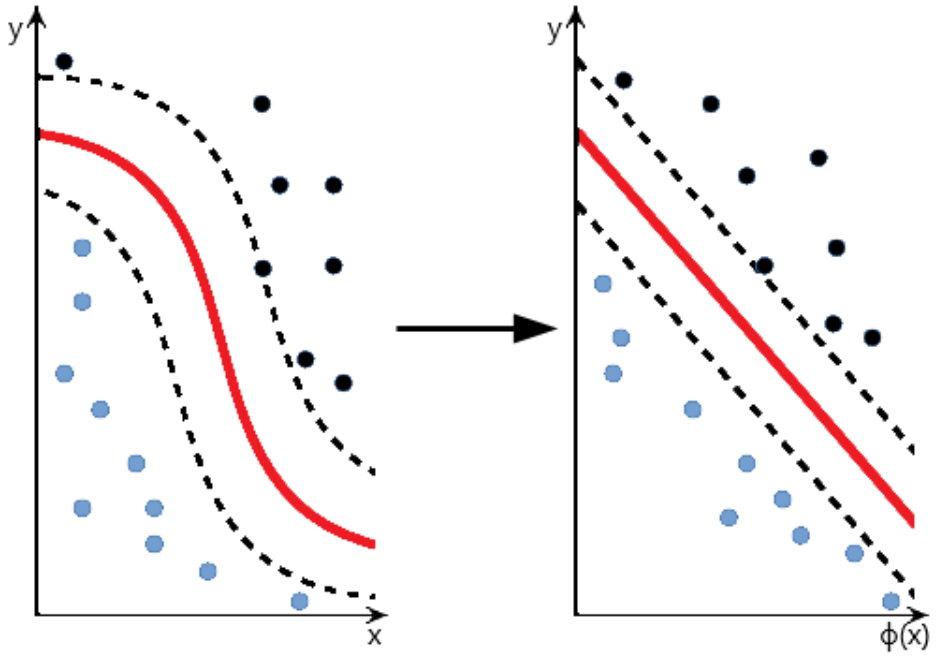

Figure 8: Simplified illustration of the kernel trick. Feature space is transformed in such a way that the two classes of training data points can be linearly separated with a wide margin by a flat hyperplane, shown here as a straight red line. This leads to faster computation. Adapted from Bock *et al.* [11].

Random Forest (RF) is a classification method based on the construction of a multitude of randomly varying decision trees. Classification of a data point is based on the class chosen by the largest subset of trees [52], see Figure 9. Trees are trained using *bagging*, i.e. in a parallel way.

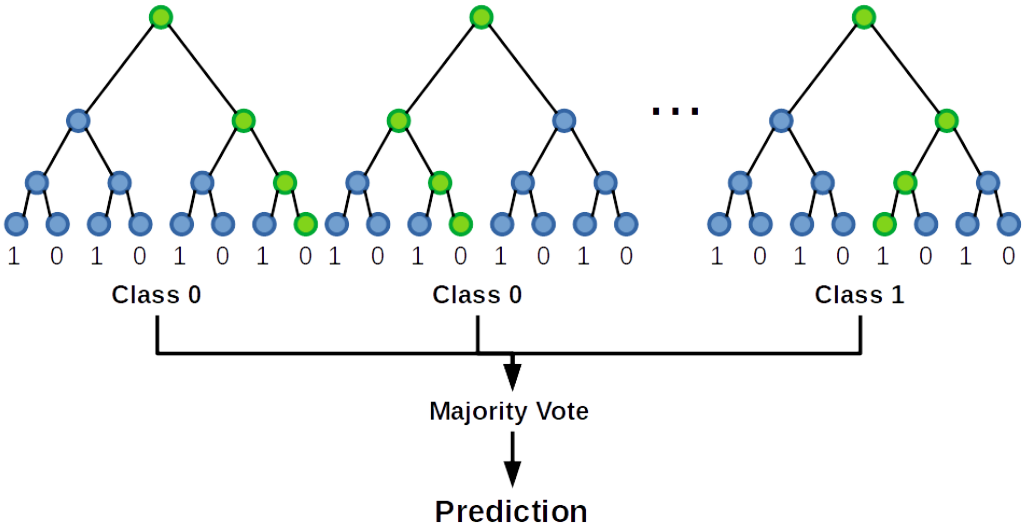

Figure 9: Illustration of an example for a machine learning method based on a multitude of randomly varying decision trees, such as RF and ET, using a bagging method, i.e. only using a subset of the training set for the construction of each tree. The algorithm constructs many decision trees with varying splits, represented here as nodes, using the training data. Each split shows a decision point in the model, which can be compared to a feature for a prediction. Green nodes are the path chosen for the specific features of a data point, blue nodes the decisions points not reached for the data point, numbers show a prediction at the end of each possible path. Each tree reaches a prediction on a new data point. The result chosen by the highest number of trees is returned as the final prediction.

Classifiers using only a single decision tree will split a dataset at nodes to arrive at a classification at the leaves of the tree. The main problem with the use of a single tree is that it tends to overfit the training data, causing it to perform badly on new datasets. This problem can be solved using multiple randomly varying trees with low correlation. This is accomplished by the method of bagging, i.e. only using a subset of the training set for the construction of each tree. Random forest algorithms use an additional method called feature bagging, i.e. the random selection of only a subset of the features at each split in the construction of decision trees. This decreases the correlation further, as it prevents features that are very strong predictors for the classification to be chosen for too many trees.

The optimal amount of trees, as well as the optimal numbers of features used at each split, differs for each problem and must be tuned using cross-validation or other methods.

Extremely Randomized Trees, or extra trees (ET), is a bagging classifier and works very similarly to random forest. Both algorithms are composed of a large multitude of trees, choosing the classification of data points based on the largest subset of trees making the same prediction [45]. Their methods for constructing decision trees are also very similar, with a few differences.

While random forests use only a subset of the original dataset for the construction of each tree, extra trees uses the entire dataset, reducing bias. Additionally, the cutoff point where nodes split is chosen randomly for extra trees, instead of being calculated optimally. Once the split points are selected, the best one between the subsets of features is accepted. This step reduces the variance. Since extra trees chooses its split points randomly, it is computationally faster than random forests. See Figure 10 for a comparison between RF and ET.

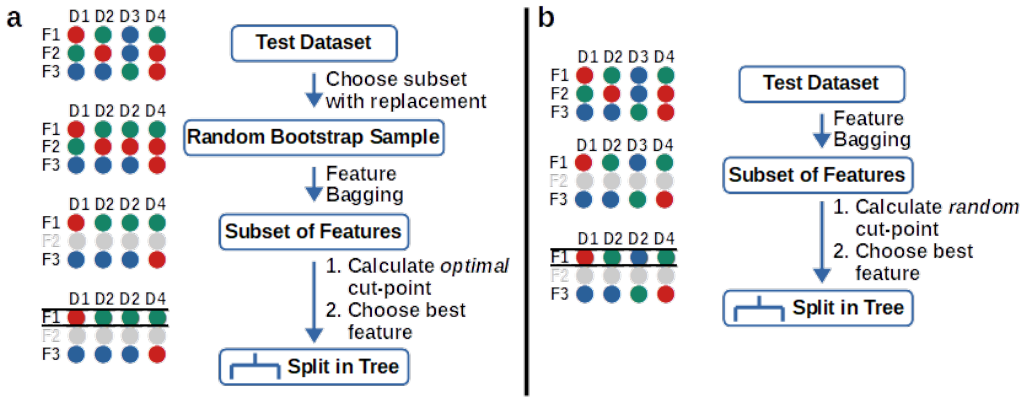

Figure 10: Comparison between RF and ET methods for splitting a node. D1 to D4 are data points, F1 to F3 are features, colored dots are meant to represent different values of features for different data points.

Gradient Boosting (GB) is another tree-based classifier, but instead of bagging, it utilizes *boosting*. This means that trees are not trained in a parallel manner, but instead consecutively, learning from the errors of their predecessors. GB begins by training a single decision tree on a dataset, then

uses that tree to make predictions. It will then calculate the residual error of this decision tree and use this error as the new target value for prediction. This process is repeated until a predetermined number of trees has been constructed, after which the final prediction is made, see Figure 11.

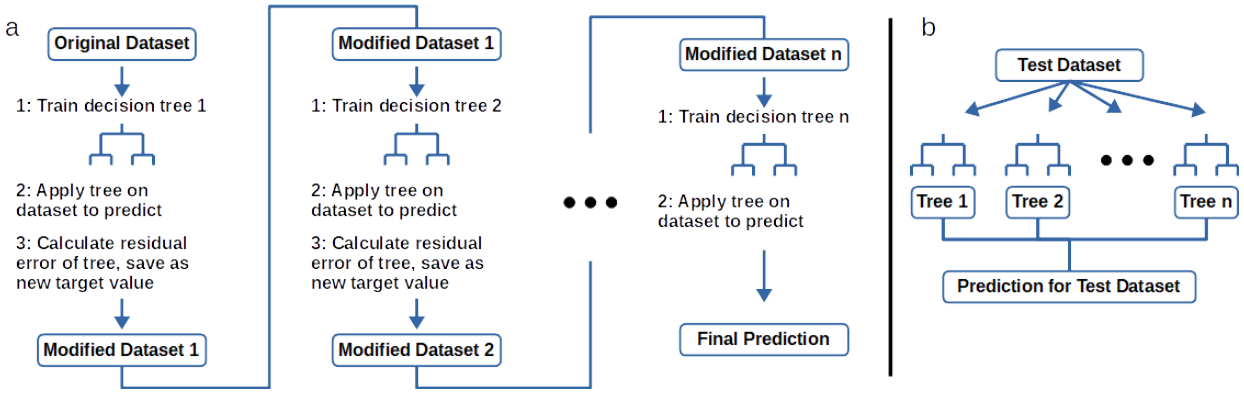

Figure 11: Illustration of the GB method. **a:** A decision tree is trained, then used for predictions. The residual error is used as the new target value. This process is repeated until the final prediction is reached. **b:** GB makes a prediction on the test dataset by adding up the predictions of all trees from the training step. Adapted from *towards data science* [18].

3.4.2 Feature Extraction and Preprocessing

Cysteine Features

Protein set 1 was preprocessed 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 413 features for each cysteine. I considered the twenty closest amino acids in the sequence, according to a threshold adapted from Passerini *et al.* [83].

The second method extracted 69 features for each cysteine. The features were intended to describe physical properties and were computed for each cysteine and each of the thirteen 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 protein set 2, since the

3D neighborhood of cysteines was unknown for the proteins with unknown structure.

For each amino acid, I computed 22 features:

- Seventeen 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). I adopted standard values for the physicochemical properties for the amino acids from Lin *et al.* [68].

- Half Sphere Exposure values $HSE1$ and $HSE2$: I 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, I applied DSSP [60]. 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.* [103].

- Two binary values to assign an SSE to the residue according to DSSP:
  - α-helix (H), 3-helix (G), 5-helix (I), which were counted as "helix"
  - residue in isolated β-bridge (B), extended strand (E), hydrogen-bonded turn (T), which were counted as "strand"
  - bend (S)
  - loop/irregular structure (-)

For the sequence-based method, the residues were treated separately and in sequence order. The seventeen physicochemical properties plus the HSE1,

HSE2 and relative accessibility for each of the twenty closest residues sum up to 400 values.

The vectors for the SSE assignments were added together into two values. I used the residue score for ligand binding sites calculated by ConCavity as an additional feature. PROPKA 2.0 and PROPKA 3.4 were used to calculate two features for the predicted pK_a value of the cysteine. The number of known PTMs (phosphorylation, acetylation or ubiquination) of residues among the twenty closest residues of the investigated cysteine were also used as three separate features. The set of values was completed by the five features of the cysteine itself (HSE1, HSE2, accessibility, two values for SSEs), not including the redundant seventeen physicochemical properties of the cysteine.

For method 2, I added all values for any of the physicochemical properties together into one value for each of the seventeen 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 two values. HSE and relative accessibility were treated separately for each residue, resulting in 39 additional features. Together with the values for ligand binding sites, PTMs and the secondary structure, HSE and relative accessibility for the cysteine itself, this sums up to 69 features.

I collected 3D positions of residues from the PDB files of highest resolution. For the imputed dataset, all data, except for the seventeen physicochemical properties of the 20 neighboring residues and the PTMs, were imputed. All values were scaled using the StandardScaler function of scikit-learn [85].

When building machine learning models, it is often unclear which features from the available data will prove to be useful for predictions. Too many features may not only slow computation down significantly and unnecessarily, but may even reduce the accuracy of the classifier. The model may more easily be able to fit the training data perfectly, but due to overfitting, it may not generalize to unfamiliar samples and perform less favorably on new data. A simple model is also preferable when compared to an unnecessarily complex one to explain a problem, as the best explanation is the one which makes the fewest assumption. I decreased the number of required features by applying feature selection. I utilized the SelectFromModel function of scikit-learn [85]. This method was chosen because it showed better performance than competing methods, like SelectKBest with F-score or Boruta, across

all machine learning algorithms used in this study. Feature selection was performed independently within each cross-validation fold. *Mean* feature importance was used as a threshold for the selection.

Protein Features We used Shapemers [33], CTD values (Composition, Transition, Distribution) of residues/SSEs and autocovariance [112] values of amino acids' physicochemical properties/accessibility as features to predict redox-active proteins.

Table 1: Grouping of amino acids for CTD values

Group	0	1	2	3	4	5	6
AA	A, G, V	C	D, E	F, I, L, P	H, N, Q, W	K, R	M, S, T, Y

NOTE: Grouping for composition, transition and distribution according to You *et al.* [112].

I divided the 20 amino acids into seven groups based on the dipoles and volumes of the side chains according to Table 1 [112] and the SSEs into coil, helix and strand to calculate CTD values. Composition represents the proportion of AA or SSE groups of the protein sequence, resulting in seven AA and three SSE features. Transition denotes the frequency with which one member of a group is followed by a member of a different or the same group, resulting in 49 AA and nine SSE features. Distribution shows the distribution patterns along the protein by indicating the location before which we find the first member as well as 25%, 50%, 75% and 100% of the members of a group. This results in 35 AA features and 15 SSE features, bringing the total to 118 features.

The protein is divided into seven regions for the calculation of these features. The regions represent first and second half, all quarters and the middle 50% of the sequence, resulting in an 826-dimensional feature vector.

I used autocovariance to transform the physicochemical properties of residues and their ASA into uniform matrices. Autocovariance can be used to show interactions between residues that are a specific number of positions apart, taking neighboring effects into account. Any protein P of length L can be represented by values calculated using the following equation:

$$AC(lag, j) = \sum_{i=1}^{L-lag} (P_{i,j} - \frac{1}{L}\sum_{i=1}^{L} P_{i,j}) \times (P_{(i+lag),j} - \frac{1}{L}\sum_{i=1}^{L} P_{i,j})/(L - lag)$$

Here, *lag* is the distance between residues, j is the jth physicochemical property, i is the position in the sequence. This results in a feature vector of $lag_{max} \times q$ dimensions, where q is the number of physicochemical properties and lag_{max} is the maximum lag distance. I chose a maximum lag of four, resulting in a 52-dimensional vector. CTD and autocovariance features were calculated for protein set 3.

3.4.3 Parameter Optimization

Values for the RBF parameters γ and C were chosen empirically using a grid search algorithm for all SVM predictions.

I tested values of γ in the range of $\gamma = 2^i,\ i = -15, -14, \ldots, 0$ and values of the cost C in the range of $C = 2^i,\ i = -5, -4 \ldots, 5$, using the GridSearchCV function of scikit-learn. For each of the 15×10 possible combinations, I performed test runs for the final dataset of features. I obtained preferable prediction results, i.e., high values of positive predictive value and sensitivity, for the combination $\gamma = 2^{-7}$ and $C = 2^0$.

For RF and ET, grid search was applied to find the most promising values for the parameters of the number of estimators or trees in the forest, the maximum number of features to consider for each split and the minimum number of samples required to split a node. Increasing the number of trees will usually allow the model to better learn the data. However, a number of trees that is too high may considerably slow down the process and even lead to suboptimal outcomes. A high maximum number of features considered for each split in the tree may lead to overfitting in some cases. Conversely, a high minimum number of samples required to split an internal node may lead to underfitting, as each tree is constrained by having to consider more samples at each node.

I tested values for the number of estimators in the range of $n(estimators) = 500 \times i,\ i = 1, 2, 3$, values of the maximum number of features in the range of $feature_{max} = 5 \times i,\ i = 2, 3, 4$ and values for the minimum sample split in the range of $split_{min} = 2 \times i,\ i = 1, 2, 3$. For each of the $3 \times 3 \times 3$ possible combinations, I performed test runs for the final set of features. I obtained preferable prediction results, i.e., high values of positive predictive value and sensitivity, for the combination $n(estimators) = 1000$, $feature_{max} = 20$ and $split_{min} = 4$.

For all algorithms, I applied 5-fold cross-validation to guard against over-

fitting: the dataset was split randomly into five subsets, each containing 20% of the dataset. I used four of those subsets, comprising 80% of the dataset, as training data, which were used to train the model containing predictive features for both Cys+ and Cys−. I then used the remaining 20% of the dataset to test the performance of the model.

3.4.4 Evaluation

To evaluate the algorithms, I used the Area Under the Curve (AUC) value for the Receiver Operating Characteristic (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 accurately predicted Cys+ 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 1 would be a perfect score, while a value of 0.5 signifies a completely random classification.

4 Results

4.1 Statistics

I studied the differences between composition of residues in the 3D and sequence neighborhood of Cys+ and Cys− to better understand and evaluate the usefulness of different features used in the machine learning models. I considered the twenty nearest residues in Euclidean space or sequence. Significance values are indicated as followed: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

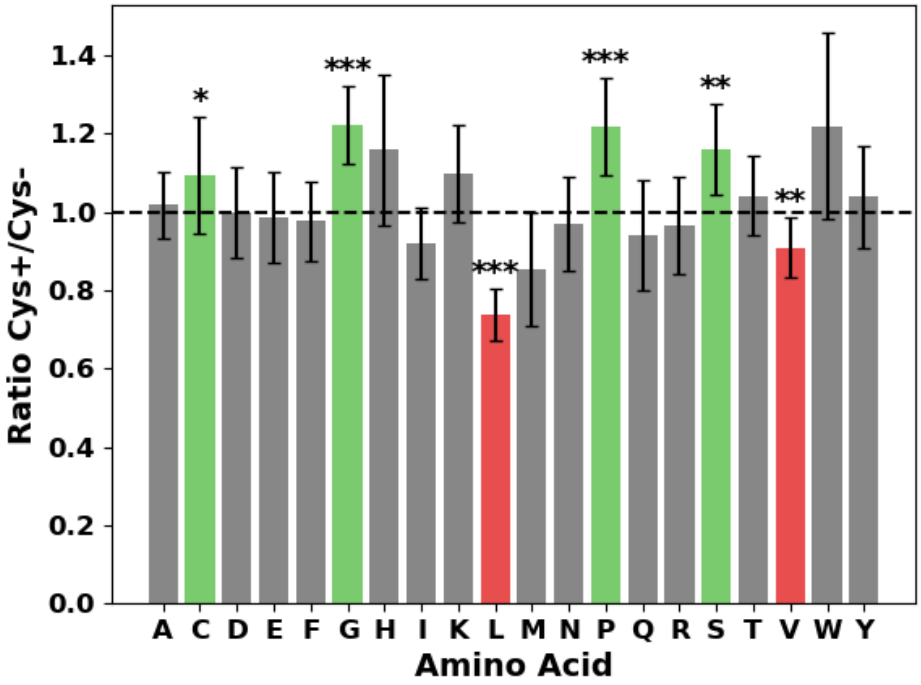

Figure 12: Relative frequencies of residues in the 3D neighborhood of Cys+. Green bars indicate significantly elevated frequencies and red bars significantly reduced frequencies. The error bars are $3 \times \sigma$ with σ being the Poisson standard deviation for the total number of counts. The frequencies of the amino acids cysteine (C), glycine (G), proline (P) and serine (S) are significantly elevated. The relative frequencies of leucine (L) and valine (V) significantly reduced.

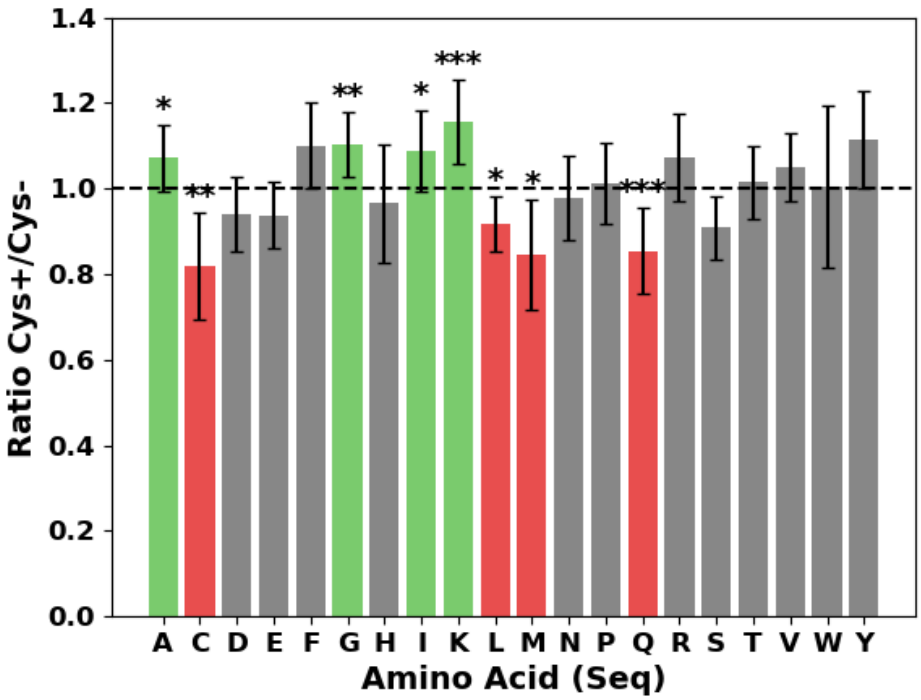

Figure 13: Relative frequencies of residues in the sequence neighborhood of Cys+. Green bars indicate significantly elevated frequencies and red bars significantly reduced frequencies. The error bars are $3 \times \sigma$ with σ being the Poisson standard deviation for the total number of counts. The frequencies of the amino acids alanine (A), glycine (G), isoleucine (I), and lysine (K) are significantly elevated. The relative frequencies of cysteine (C), leucine (L), methionine (M) and glutamine (Q) significantly reduced.

Figure 12 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, which only contains proteins with resolved 3D structures. The frequencies of cysteine, glycine, proline and serine are significantly elevated, while the frequencies of leucine and valine are significantly reduced. In protein set 2, which contains additional proteins with unknown structure, elevated frequencies of alanine, glycine, isoleucine and lysine in the sequence neighborhood of Cys+ were found, while the frequencies of cysteine, leucine, methionine and glutamine were reduced, see

Figure 13. It has been suggested [102] 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) [92, 95], which may be a possible explanation for the relative abundance of these amino acids. The frequency of leucine and valine are significantly reduced.

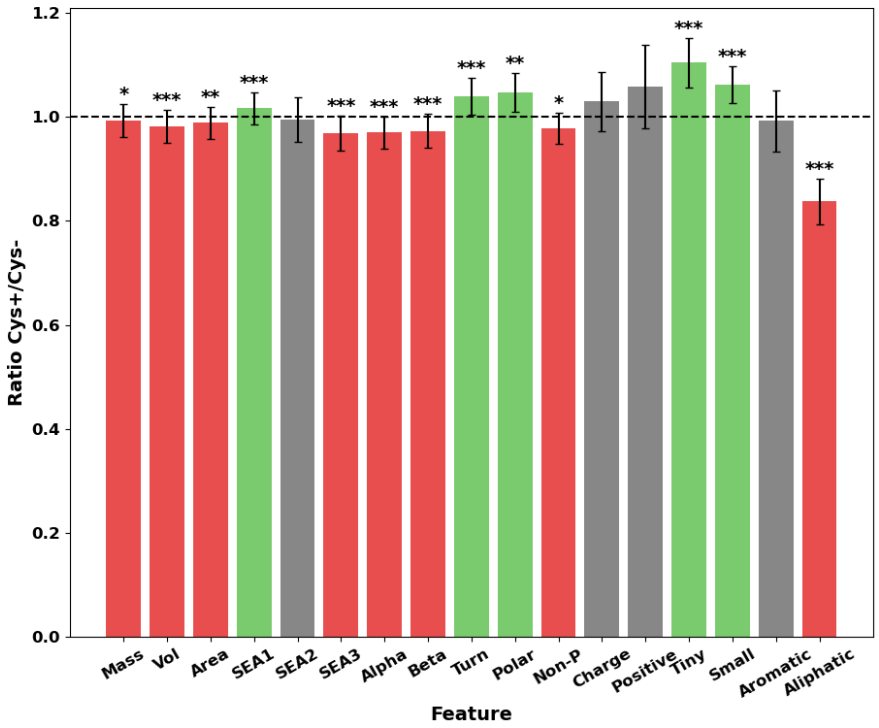

Figure 14: Physicochemical properties of residues in the 3D neighborhood of a Cys+. The values are averaged over the twenty closest residues and are given relative to the corresponding mean value of residues close to a Cys−. For example, the elevated first value of solvent exposed area (SEA1) indicates free space around the cysteine that may make it vulnerable to a modification by ROS, see text.

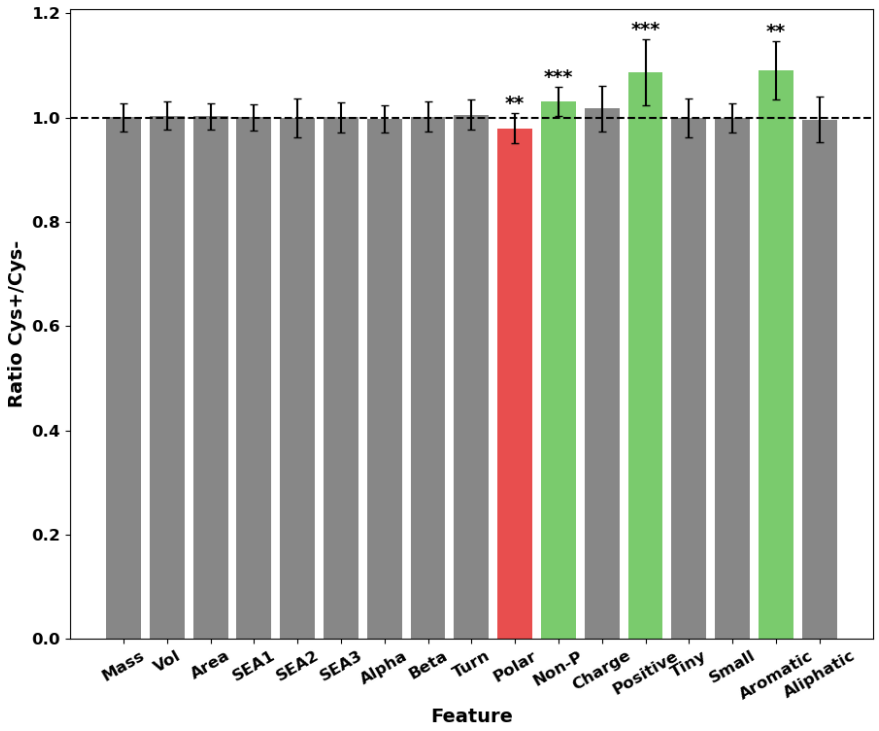

Figure 15: Physicochemical properties of residues in the sequence neighborhood of a Cys+. The values are averaged over the twenty closest residues and are given relative to the corresponding mean value of residues close to a Cys−.

When grouped for their features, significant differences could be found in the Euclidean neighborhood of cysteines between Cys+ and Cys− for most features, see Figure 14. I found the most significant differences for the likelihood of expressing specific SSEs as well as the presence of aliphatic side chains, according to the Mann-Whitney U test. The sequence neighborhood showed few significant differences, most significantly for non-polar and positively charged sidechains. While there is much overlap between the data for modifiable and unmodifiable cysteines, these significant differences may be one important piece of the puzzle when trying to predict modifications.

I created sequence logos [28] of the different frequencies of the closest amino acids in the Euclidean and sequence neighborhood of the cysteines in

the two groups of protein set 1 and 2, respectively, as seen in Figure 16 to 19. Sequence logos are a useful way to clearly visualize enriched and depleted amino acids at specific locations in sequences of the same length and help in the identification of sequence motifs. Logos are shown both with and without Bonferroni correction, which accounts for multiple hypotheses being tested. In a sequence logo, this type of correction may often hide general trends that hold true over the entire sequence. For example, the hydrophobic leucine (L) is reduced in nearly the entire Euclidean neighborhood and upstream in the sequence neighborhood of Cys+, especially and significantly at position -1. This is not visible in the Bonferroni corrected logo.

Positively charged amino acids like lysine (K) and arginine (R) are largely 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.* [19], who also found an abundance of positively charged residues around *S*-nitrosylation sites and a reduced occurrence of C. The polar residues glycine (G) and serine (S) are enriched in the close Euclidean neighborhood of Cys+. Glycine is likewise enriched in the close sequence neighborhood, while serine is depleted in a distance of more than five positions away from the central cysteine, see Figure 19.

I was unable to confirm the presence of an acid-base motif which was found in some previous studies on smaller datasets [47, 51]. Marino *et al.* [74] 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.

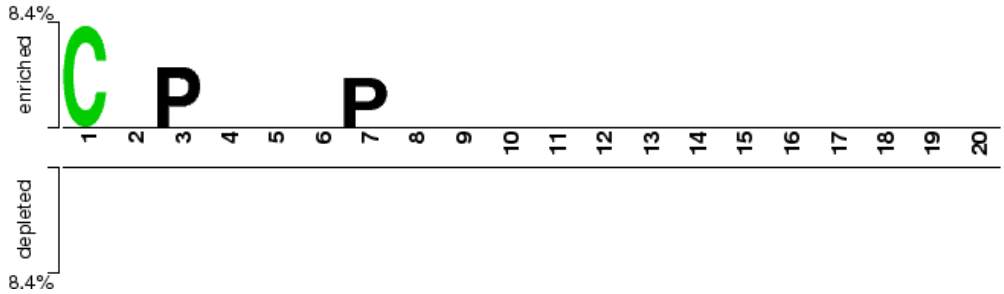

Figure 16: Sequence logo [28] of differences between the residues in the Euclidean neighborhood of Cys+ and Cys− in protein set 1. Only differences with a p-value < 0.05 according to the t-test after Bonferroni correction are shown. Distance order in relation to cysteine can be found on the x-axis, percentage difference of Cys+ in relation to Cys− on the y-axis. Enriched residues around Cys+ are illustrated at the top, depleted residues at the bottom. Size of symbols is proportional to the difference between the two samples.

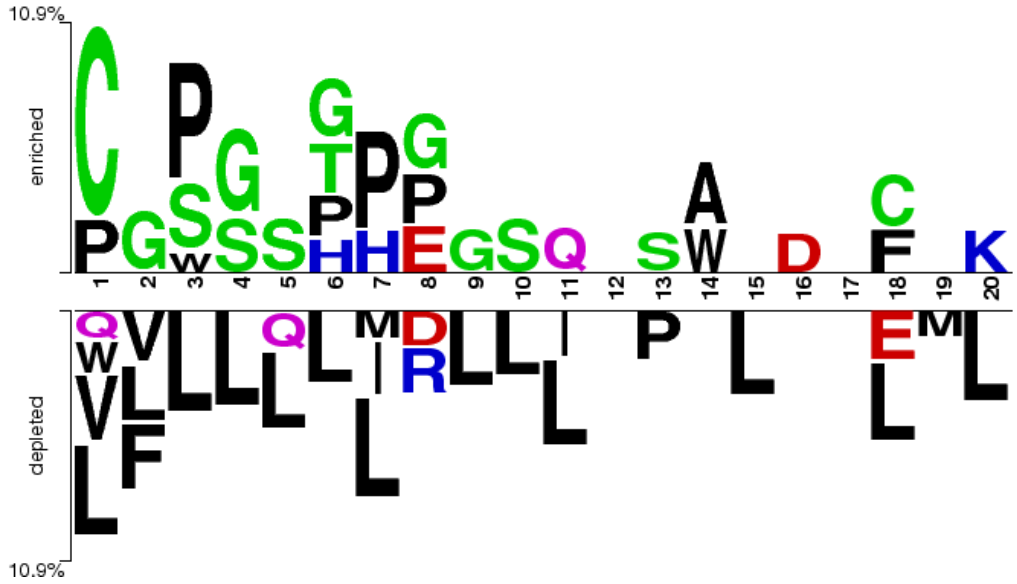

Figure 17: Sequence logo [28] of differences between the residues in the Euclidean neighborhood of Cys+ and Cys− in protein set 1. Only differences with a p-value < 0.05 according to the t-test are shown. Bonferroni correction was not applied. Distance order in relation to cysteine are shown on the x-axis, 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.

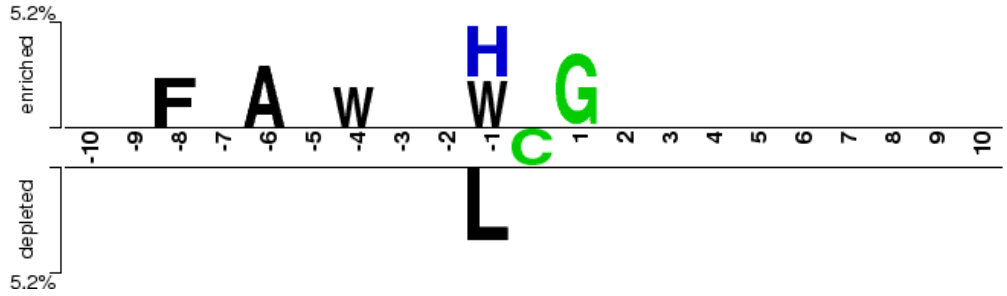

Figure 18: Sequence logo [28] of differences between the residues in the sequence neighborhood of Cys+ and Cys− in protein set 2. Only differences with a p-value < 0.05 according to the t-test after Bonferroni correction are shown. Sequence position in relation to cysteine are shown on the x-axis, percentage 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.

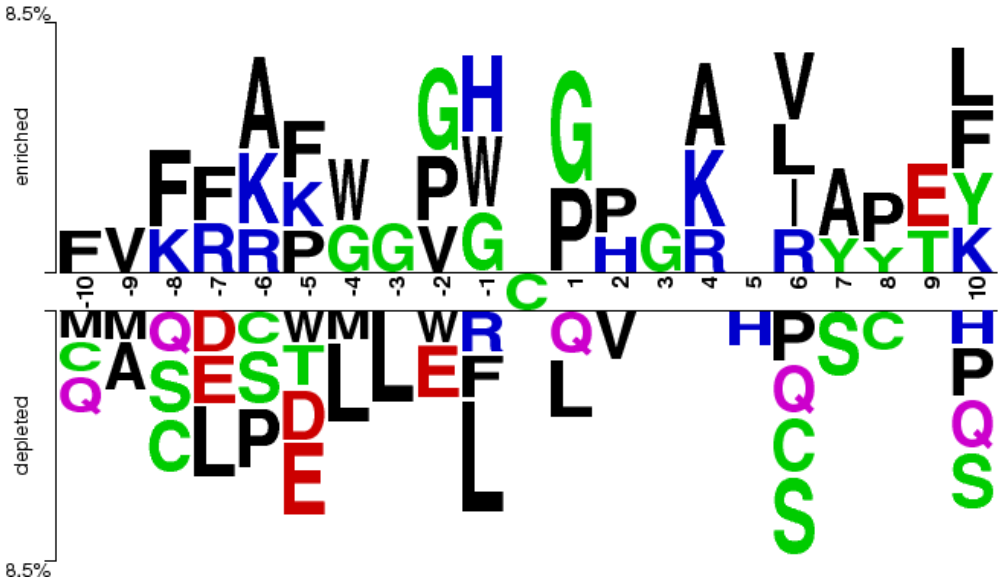

Figure 19: Sequence logo [28] of differences between the residues in the sequence neighborhood of Cys+ and Cys− in protein set 2. Only differences with a p-value < 0.05 according to the t-test are shown. Bonferroni correction was not applied. Sequence position in relation to cysteine are shown on the 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.

A higher abundance of phosphorylated residues was detected in the sequence neighborhood of Cys+ than Cys− in protein set 2 according to Uniprot [25] annotations. On average, each Cys+ had around 0.089 phosphorylated serines, threonines and tyrosines in its neighborhood of ten residues upstream and downstream, while only 0.046 phosphorylated residues around Cys− were observed, making phosphorylation about 1.91 times more common around Cys+ (data not shown). Modulation and crosstalk between phosphorylation sites and redox modifiable cysteines has been found in previous studies [62]. Similar results were found for ubiquitination near cysteines. I located an average of 0.14 ubiquitination sites around modified cysteines, while I only detected an average of 0.06 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 a p-value < 0.001 according to the Mann-Whitney U test. I did not find significant differences in the rate of acetylation near cysteines.

It appears likely that highly accessible cysteines would be easier to reach for and thus more reactive to ROS. I explored the accessible surface area as predicted by the algorithm DSSP [60] 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 $< 3 * 10^{-6}$), despite much overlap in the range of values, see Figure 20. This appears reasonable, as an exposed cysteine should be assumed to be more easily accessible to ROS. HSE showed similar results, see Figure 21 and Figure 22. I 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.

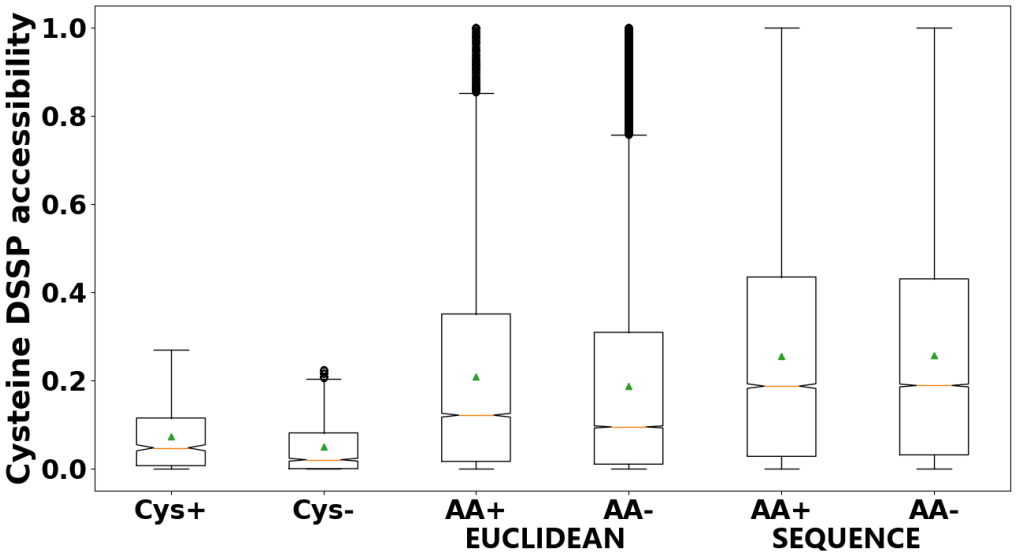

Figure 20: DSSP accessibility of Cys+ and Cys− as well as their Euclidean and sequence amino acid neighborhood. Orange bar signifies the median, green triangle the mean. Central rectangle spans the first quartile to the third quartile. Whiskers show the minimum and maximum, individual points are outliers.

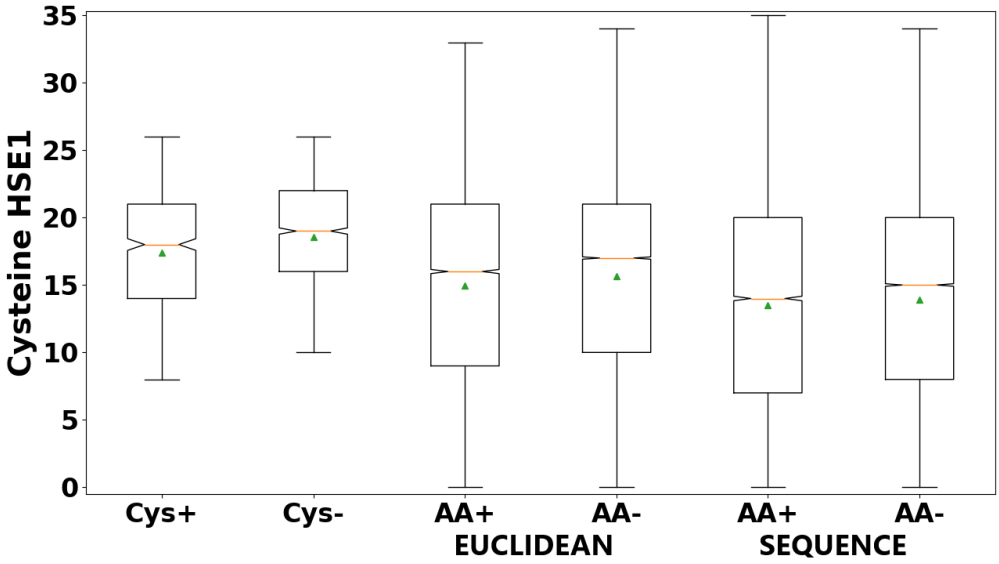

Figure 21: HSE1 of Cys+ and Cys− as well as their Euclidean and sequence amino acid neighborhood. Orange bar signifies the median, green triangle the mean. Central rectangle spans the first quartile to the third quartile. Whiskers show the minimum and maximum, individual points are outliers.

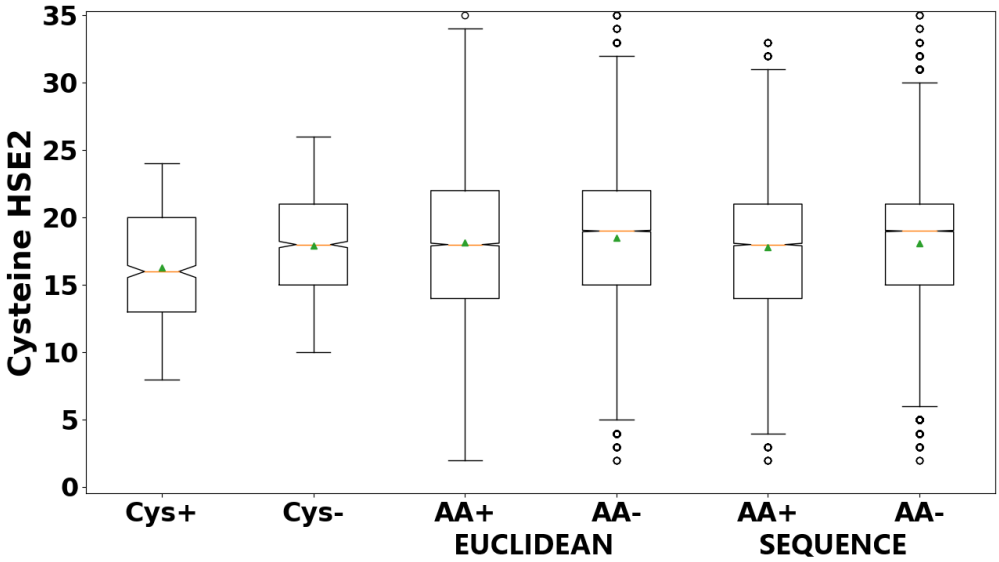

Figure 22: HSE2 of Cys+ and Cys− as well as their Euclidean and sequence amino acid neighborhood. Orange bar signifies the median, green triangle the mean. Central rectangle spans the first quartile to the third quartile. Whiskers show the minimum and maximum, individual points are outliers.

I examined the relative frequencies of SSEs in the neighborhood of Cys+ in protein set 1, as SSEs can significantly affect the structure and function of active sites in proteins. Some SSEs have been found to occur more frequently near redox modifiable cysteines. In the 3D neighborhood of Cys+, I found the frequencies of bends and unstructured loop regions to be significantly elevated and α-helices reduced, see Figure 23. 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 & Gladyshev [73], who found a marked preference for both α-helical and loop geometries around thiol oxidoreductases, testing a more limited dataset of 75 structures. I 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.* [41].

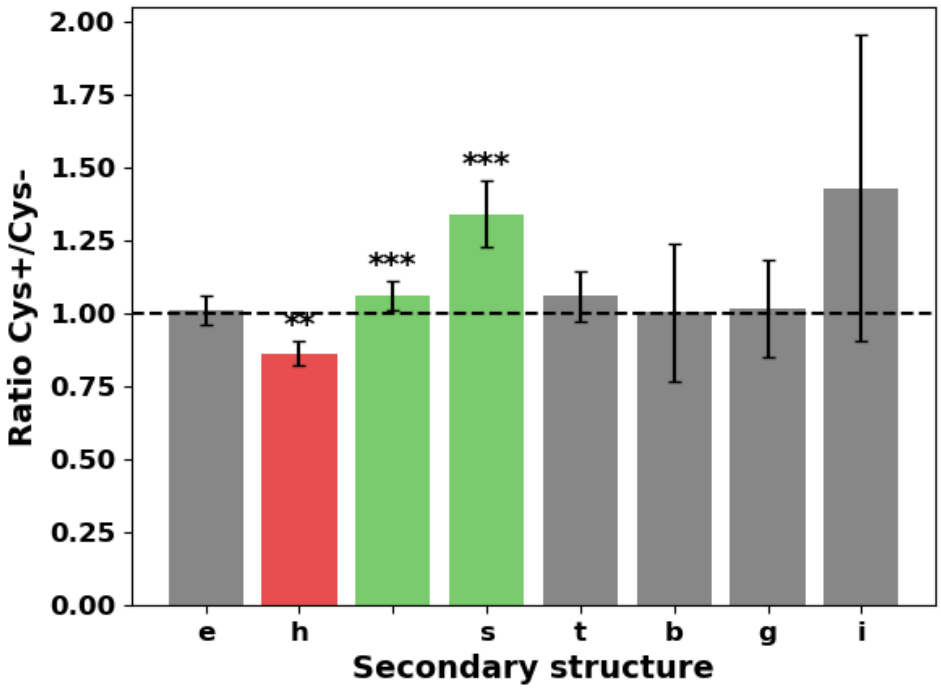

Figure 23: Relative frequencies of SSEs in the Euclidean neighborhood of Cys+. Green bars indicate significantly elevated frequencies of occurrences, red significantly reduced. SSEs were predicted by the DSSP [60] algorithm. In the Euclidean neighborhood of Cys+, the frequencies of the loop and bend structures are significantly elevated, α-helix structures reduced.

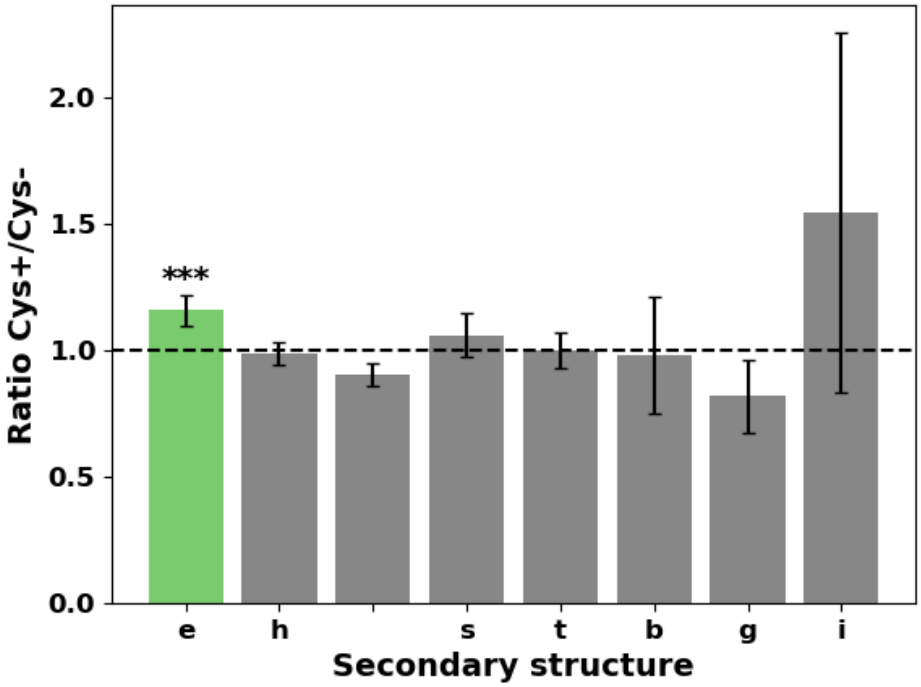

Figure 24: Relative frequencies of SSEs in the sequence neighborhood of Cys+. Green bars indicate significantly elevated frequencies of occurrences, red significantly reduced. SSEs were predicted by the DSSP [60] algorithm. In the sequence neighborhood of Cys+, the frequency of the β-strands is significantly elevated.

I created a heatmap of the Pearson correlation coefficient between the features in protein set 1, see Figure 25. This shows which features are highly correlated with the target value, i.e. redox modifiability, as well as each other. Features that are highly correlated with another feature are less useful for the purposes of machine learning, as they are essentially duplicates with little new information. 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.

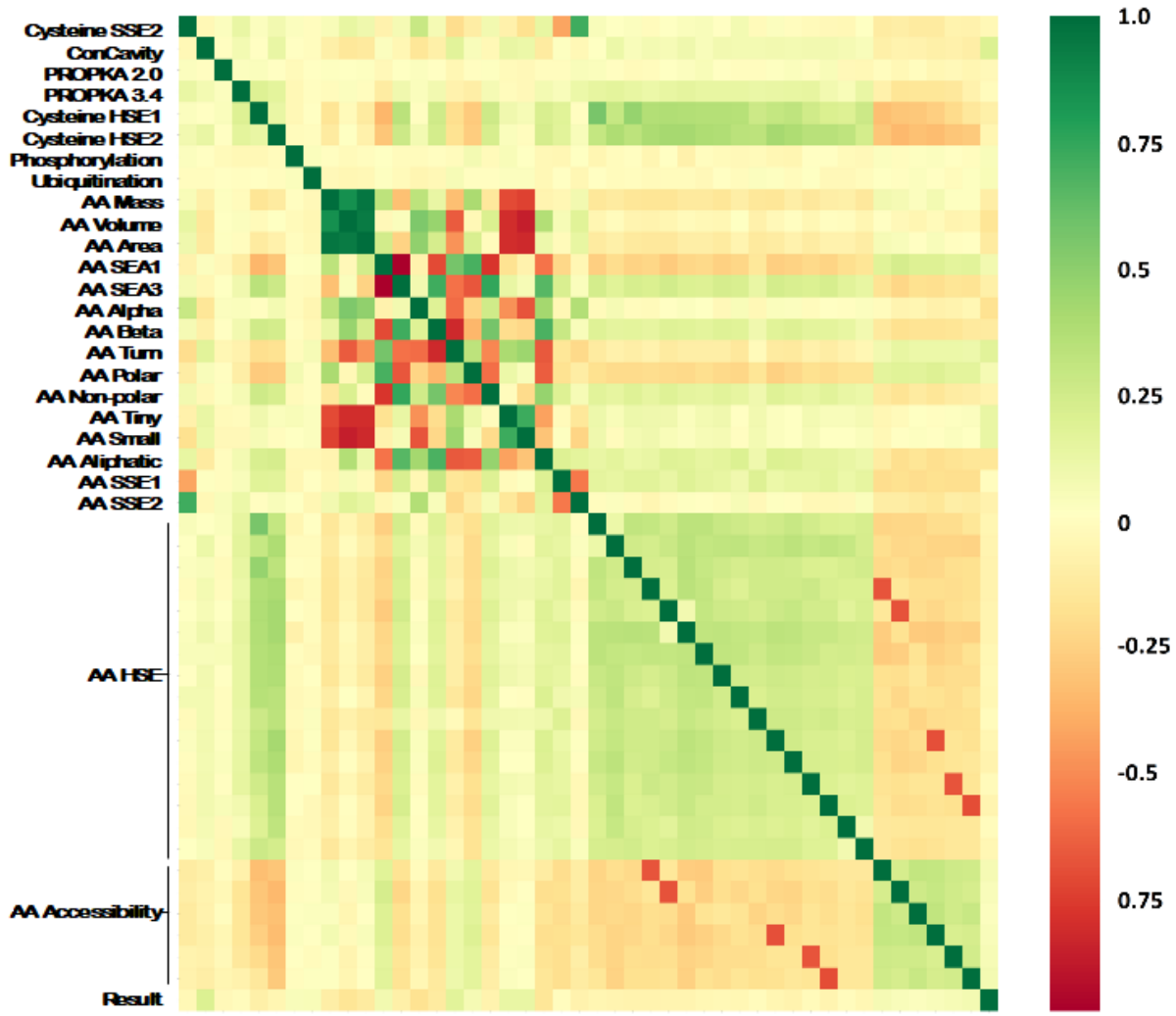

Figure 25: Heatmap of Pearson correlation coefficient between the features used in protein set 1. From top to bottom and left to right, cysteine features and number of PTMs, as well as physicochemical features, secondary structures, HSE and accessibility of neighboring amino acids are depicted. The last row and column indicate the correlation to the cysteine sensitivity. Green squares show highly correlated features, red squares show a high negative correlation, yellow squares are not highly correlated.

I used protein set 1 to compare the predictions of PSIPRED and

ASAquick to the SSE and ASA assignments by DSSP to justify the use of both tools when no PDB entry was available for a protein. In the case of PSIPRED, I found that 80.0% of its predictions agreed with the SSE assignments by DSSP. More specifically, 82.1% of helices, 79.5% of strands and 78.3% of unassigned (often coil) regions out of 5096 predictions were predicted correctly by PSIPRED. The proportion of SSEs in the full dataset were also very similar, with 40.0% and 40.2 % of AAs being assigned as coil or unassigned, 19.2% and 20.5% being assigned as strand, and 40.7% and 39.2% being assigned as helix by DSSP and PSIPRED, respectively. Comparing the ASA predictions by ASAquick to the assignments by DSSP, I calculated a Pearson correlation coefficient of 0.64 between the ASA of all AAs in the dataset. Mean ASA was similar for both tools, see Table 2.

Table 2: Statistical values for ASA

Tool	Mean	Median	St. Dev.	IQR
ASAquick	49.9	41.9	32.0	48.5
DSSP	46.0	32.0	47.1	70.0

NOTE: ASA values predicted/assigned by ASAquick and DSSP.

I compared the pK_a predictions by PROPKA 2.0 and PROPKA 3.4 between the 20 AAs closest in Euclidean space to Cys+ and Cys− to evaluate the usefulness of both tools, as pK_a is often seen as one of the main predictors of cysteine reactivity. I found significant differences according to the Mann-Whitney U test [72] between the distributions of pK_a predictions of most types of AAs, except for glutamic acid and histidine for PROPKA 2.0 and histidine for PROPKA 3.4. Arginine, cysteine and tyrosine showed the most significant differences, despite similar median values and interquartile ranges, see Table 3, Table 4, Figure 26 and Figure 27. PROPKA 3.4 showed more significant differences for all AAs except arginine.

Table 3: Median pK_a values predicted by PROPKA 2.0

Residue	Median Cys+	IQR Cys+	Median Cys−	IQR Cys−	p-value[a]
Arg	12.29	0.42	12.15	0.49	$7.1 \cdot 10^{-23}$
Asp	3.72	0.65	3.62	0.71	$1.4 \cdot 10^{-5}$
Cys	8.89	2.14	9.09	2.66	$6.2 \cdot 10^{-20}$
Glu	4.50	0.26	4.50	0.42	$1.8 \cdot 10^{-1}$
His	6.43	0.28	6.43	0.97	$1.6 \cdot 10^{-1}$
Lys	10.43	0.21	10.43	0.21	$8.3 \cdot 10^{-3}$
Tyr	10.08	1.26	10.56	2.08	$1.2 \cdot 10^{-16}$

NOTE: [a]p-value according to Mann-Whitney U test [72].

Table 4: Median pK_a values predicted by PROPKA 3.4

Residue	Median Cys+	IQR Cys+	Median Cys−	IQR Cys−	p-value[a]
Arg	12.33	0.39	12.39	0.53	$6.5 \cdot 10^{-13}$
Asp	3.77	0.62	3.86	0.64	$5.2 \cdot 10^{-9}$
Cys	11.23	2.85	11.77	1.99	$1.3 \cdot 10^{-24}$
Glu	4.58	0.47	4.58	0.46	$2.9 \cdot 10^{-2}$
His	5.81	0.94	5.82	1.12	$3.2 \cdot 10^{-1}$
Lys	10.37	0.29	10.40	0.29	$4.0 \cdot 10^{-8}$
Tyr	11.06	1.79	11.67	2.53	$7.9 \cdot 10^{-24}$

NOTE: [a]p-value according to Mann-Whitney U test [72].

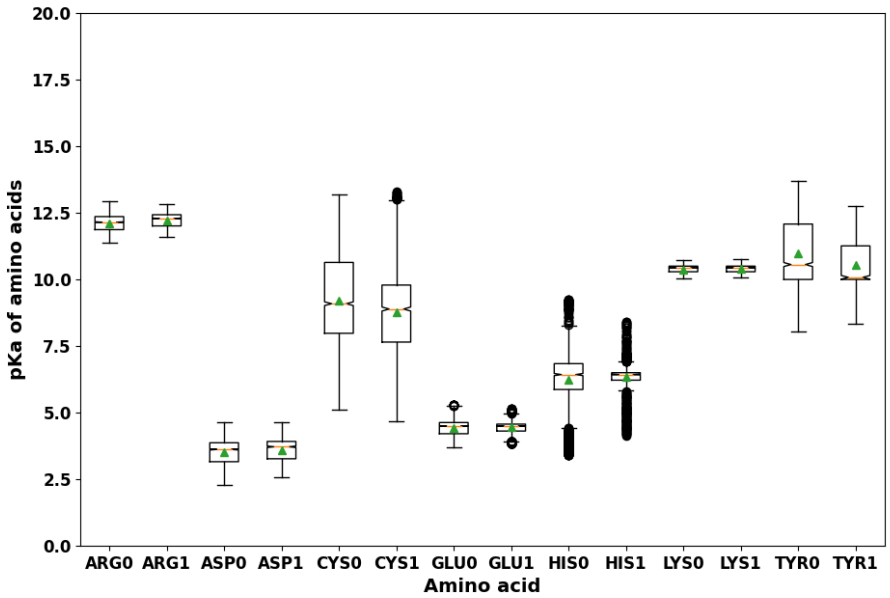

Figure 26: Distribution of pK_a values predicted by PROPKA 2.0. Data shown has had 25% of data points removed as outliers. Green triangle shows average value, orange line median. Boxes denote the second and third quartile, whiskers the first and fourth quartile, circles remaining outliers. X-axis shows AAs, with 0 denoting an AA close to a Cys− and 1 denoting an AA close to a Cys+. Y-axis shows pK_a values.

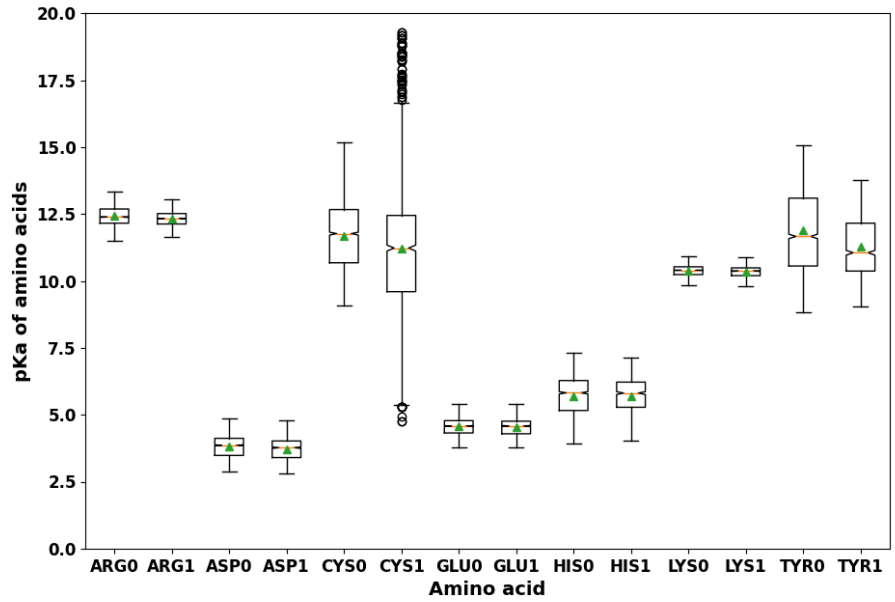

Figure 27: Distribution of pK_a values predicted by PROPKA 3.4. Data shown has had 25% of data points removed as outliers. Green triangle shows average value, orange line median. Boxes denote the second and third quartile, whiskers the first and fourth quartile, circles remaining outliers. X-axis shows AAs, with 0 denoting an AA close to a Cys− and 1 denoting an AA close to a Cys+. Y-axis shows pK_a values.

I evaluated the mutual information between the features and cysteine modifiability using the mutual_info_classif method to compare the feature importances based on the imputed protein set 2, see Table 5. A higher score signifies that the feature is more dependent in the target variable, i.e. the redox activity of the cysteine, and thus more useful for the purposes of machine learning. In general, features concerning the HSE and accessibility were found to show a high impact on redox modifiability, especially in AAs close to the central cysteine.

Table 5: Mutual information of the top 10 features

Rank	Feature	Sequence position	MI value
1	Accessibility	CYS	0.025639
2	Accessibility	-1	0.025442
3	Accessibility	-3	0.019240
4	Accessibility	-2	0.018683
5	Accessibility	-4	0.016576
6	Accessibility	-6	0.016560
7	AA mass	-10	0.015928
8	HSE1	4	0.015545
9	Accessibility	3	0.014753
10	Accessibility	6	0.014726

I calculated the ANOVA F-values and corresponding p-values for the CTD and autocovariance scores of the proteins in protein set 3. Out of 878 features, I found 86 having a p-value below 0.05 after Bonferroni correction, while 454 features possessed an ANOVA F-score above the critical value of 3.01. The distribution of SSEs, especially of strands, were overrepresented among predictive features. The composition of amino acids, especially from the AGV group (small sidechains with low dipole moment), as well as the distribution of cysteines, were also highly predictive, see Table 6.

Table 6: ANOVA F-values for the top 20 features

Rank	Feature	Residues/SSEs	Region (Distribution)	F-value	p-value[a]
0	Composition	FILP	1st quarter	43.11	$8.08 \cdot 10^{-8}$
1	Distribution	Strand	3rd quarter (75%)	39.50	$4.68 \cdot 10^{-7}$
2	Distribution	Strand	3rd quarter (100%)	39.44	$4.83 \cdot 10^{-7}$
3	Distribution	C	1st quarter (75%)	37.86	$1.04 \cdot 10^{-6}$
4	Distribution	Strand	1st quarter, (75%)	36.77	$1.78 \cdot 10^{-6}$
5	Distribution	C	1st quarter (100%)	36.49	$2.05 \cdot 10^{-6}$
6	Distribution	Strand	1st quarter (100%)	36.32	$2.23 \cdot 10^{-6}$
7	Distribution	Strand	1st quarter (50%)	34.59	$5.21 \cdot 10^{-6}$
8	Distribution	Strand	Center half (100%)	34.45	$5.57 \cdot 10^{-6}$
9	Distribution	Strand	3rd quarter (50%)	34.16	$6.44 \cdot 10^{-6}$
10	Transition	Strand to Strand	4th quarter	32.92	$1.18 \cdot 10^{-5}$
11	Composition	AGV	2nd quarter	32.72	$1.31 \cdot 10^{-5}$
12	Transition	Coil to Strand	4th quarter	32.71	$1.32 \cdot 10^{-5}$
13	Distribution	Strand	4th quarter (100%)	32.62	$1.37 \cdot 10^{-5}$
14	Transition	Strand to Strand	1st quarter	32.52	$1.45 \cdot 10^{-5}$
15	Distribution	C	1st quarter (50%)	31.99	$1.88 \cdot 10^{-5}$
16	Composition	FILP	3rd quarter	31.35	$2.58 \cdot 10^{-5}$
17	Composition	AGV	right half	31.20	$2.78 \cdot 10^{-5}$
18	Composition	AGV	1st quarter	30.95	$3.15 \cdot 10^{-5}$
19	Composition	AGV	3rd quarter	30.55	$3.84 \cdot 10^{-5}$

NOTE: [a]p-value according to Mann-Whitney U test [72], Bonferroni corrected.

I proceeded by looking at some of the more interesting features in detail. The group of residues F, I, L and P, which possess large side chains and low dipole moments, appear less frequently in redox-active proteins, see Figure 28. Residues A, G and V, possessing small side chains and low dipole moments, can be observed more frequently in several regions, see Figure 29. This may indicate that it is not the dipole moment of residues, but a different measure that is important to redox-activity in proteins.

I detected a lower number of strand-to-strand transitions in random Uniprot proteins than in RedoxDB proteins, see Figure 30. Distribution values often indicated a complete absence of strands in Uniprot protein regions, see Figure 31. Both results indicate a higher presence of β-strands in redox-active proteins, agreeing with the higher presence of strands close to redox-active cysteines that were found.

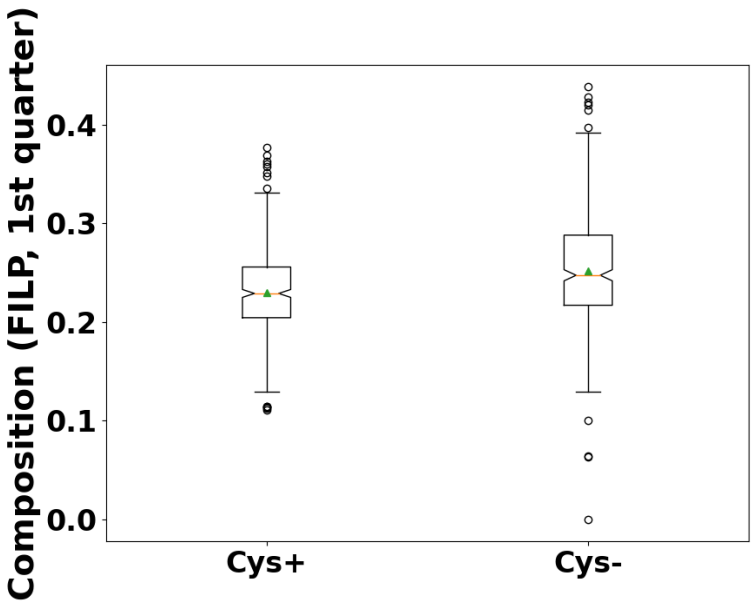

Figure 28: Amount of the residues F, I, L and P in RedoxDB and Uniprot proteins in the first quarter of the sequence. Orange bar signifies the median, green triangle the mean. Boxes denote the second and third quartile. Whiskers show the minimum and maximum, individual points are outliers.

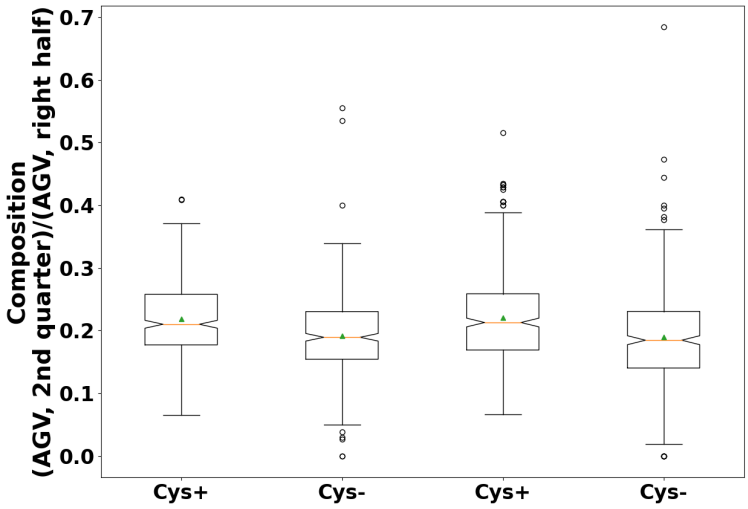

Figure 29: Amount of the residues A, G and V in RedoxDB and Uniprot proteins in the second quarter as well as the second half of the sequence. Orange bar signifies the median, green triangle the mean. Boxes denote the second and third quartile. Whiskers show the minimum and maximum, individual points are outliers.

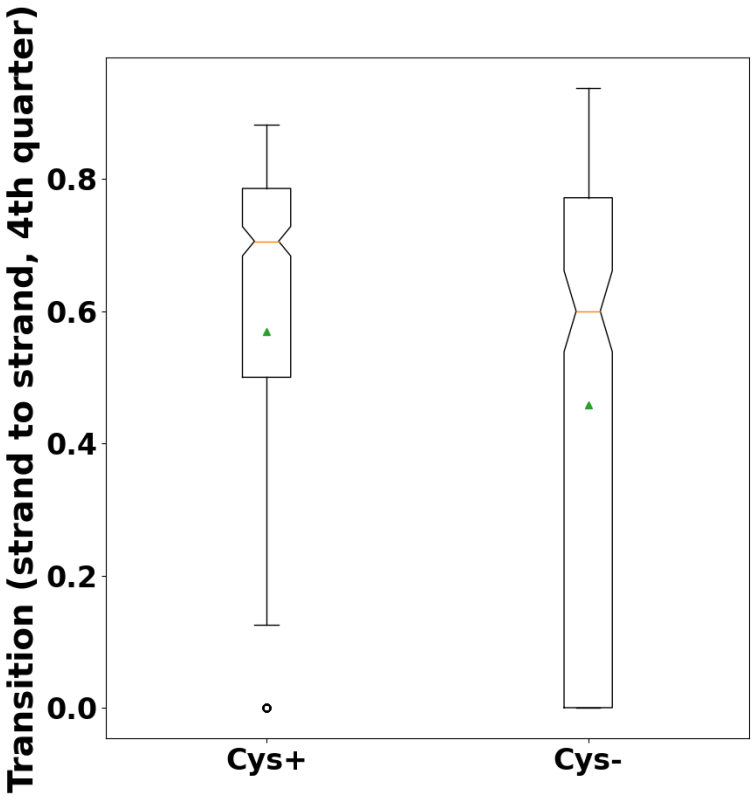

Figure 30: Amount strand-strand transitions in RedoxDB and Uniprot proteins in the fourth quarter of the sequence. Orange bar signifies the median, green triangle the mean. Boxes denote the second and third quartile. Whiskers show the minimum and maximum, individual points are outliers.

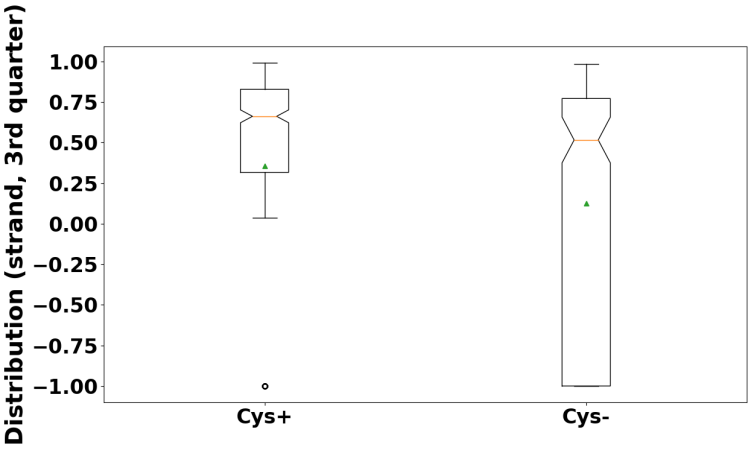

Figure 31: Distribution values of strands in RedoxDB and Uniprot proteins in the third quarter of the sequence. y-value shows if a certain number of strands could be found near the beginning ($y = 0$) or end ($y = 1$) of the protein region. $y = -1$ indicates no presence of strands. Orange bar signifies the median, green triangle the mean. Boxes denote the second and third quartile. Whiskers show the minimum and maximum, individual points are outliers.

4.2 Machine Learning

4.2.1 Proteins

I calculated CTD values and autocovariance of all proteins in protein set 3. After selecting the 30 best features with the scikit-learn SelectKBest function [85], a model for the prediction of redox-active proteins was produced by applying the ET algorithm, achieving an AUC value of 0.75, see Figure 32. The RedoxDB contains a higher percentage of mammalian proteins than Uniprot (60% and 34% in the data, respectively), leading to worries of either a potential bias in the model, or that the model may actually merely predict mammalian proteins. I investigated this potentiality, and found that the difference between prediction values for mammalian and non-mammalian proteins were smaller than the difference between redox-active and other proteins, see Table 7. Mammalian proteins did not display much higher prediction values than non-mammalian proteins, showing that the model does not erroneously learn to always predict them as redox-active. A difference in

average sequence length between datasets may also bias results. My investigations showed that this was also not a factor, as the average length for predicted redox-active proteins closely matched the average length of redox-active proteins in the training data, see Table 8. The methods used in this section will be applied to predict Cys+ containing proteins in chapter 4.3

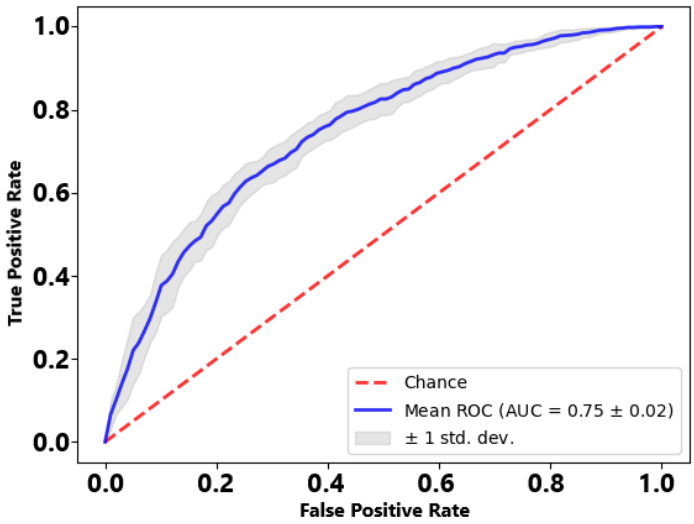

Figure 32: ROC curve of the ET algorithm for the protein set 3. Red line shows a completely random prediction.

Table 7: Average prediction for redox-active and random Uniprot proteins

	Redox-active	Random Uniprot
Mammal	0.59	0.51
Non-mammal	0.67	0.41

Table 8: Average length of redox-active and random Uniprot proteins

	Redox-active	Random Uniprot
Training	523.16	451.66
Predicted	518.13	461.55

We applied the Geometricus tool to assign shapemers to protein set 4. These shapemers were utilized as features to train another model with the ET and RF algorithms to predict redox-sensitive proteins. We used both radius-based and sequence based structural fragments as a foundation for the calculations, with a radius of 16 Å and a sequence length of 16 AAs, respectively. The method found 20,450 shapemers when using structural fragments based on sequence and 12,110 when based on radius with a resolution parameter of 0.5. The radius-based method achieved an AUC value of 0.78 for both the RF and ET algorithms, see Figure 33. The sequence-based method achieved an AUC of 0.81 and 0.82 for RF and ET, respectively, see Figure 34. See Figures 35 and 36 for the structure of one sequence-based shapemer typical for proteins containing Cys+, both on its own and as part of the full protein structures. We can see that Cys+ often appear to be clustered closely around the shapemer. See Figure 37 for the sequences of the shapemers.

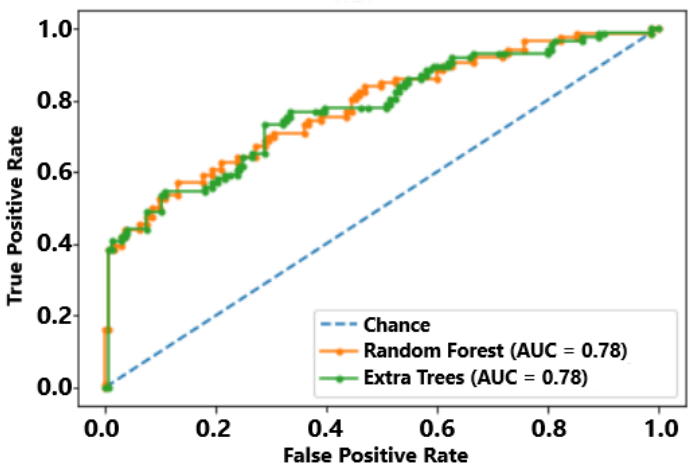

Figure 33: ROC curve of the ET algorithm for the protein set 4 using radius-based shapemers. Blue line shows a completely random prediction.

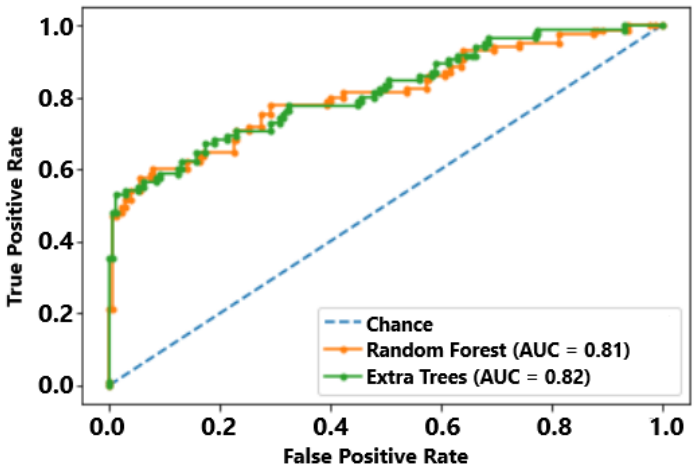

Figure 34: ROC curve of the ET algorithm for the protein set 4 using sequence-based shapemers. Blue line shows a completely random prediction.

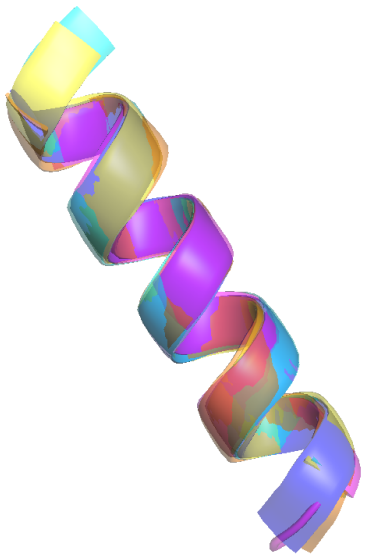

Figure 35: Aligned structure of a shapemer typical in proteins containing Cys+, found in PDB entries 2Q8K [64] (blue), 2HQM [113] (yellow), 1U8F [58] (magenta), 2QRJ [4] (cyan) and 6DFP [Kim *et al.*, to be published] (orange).

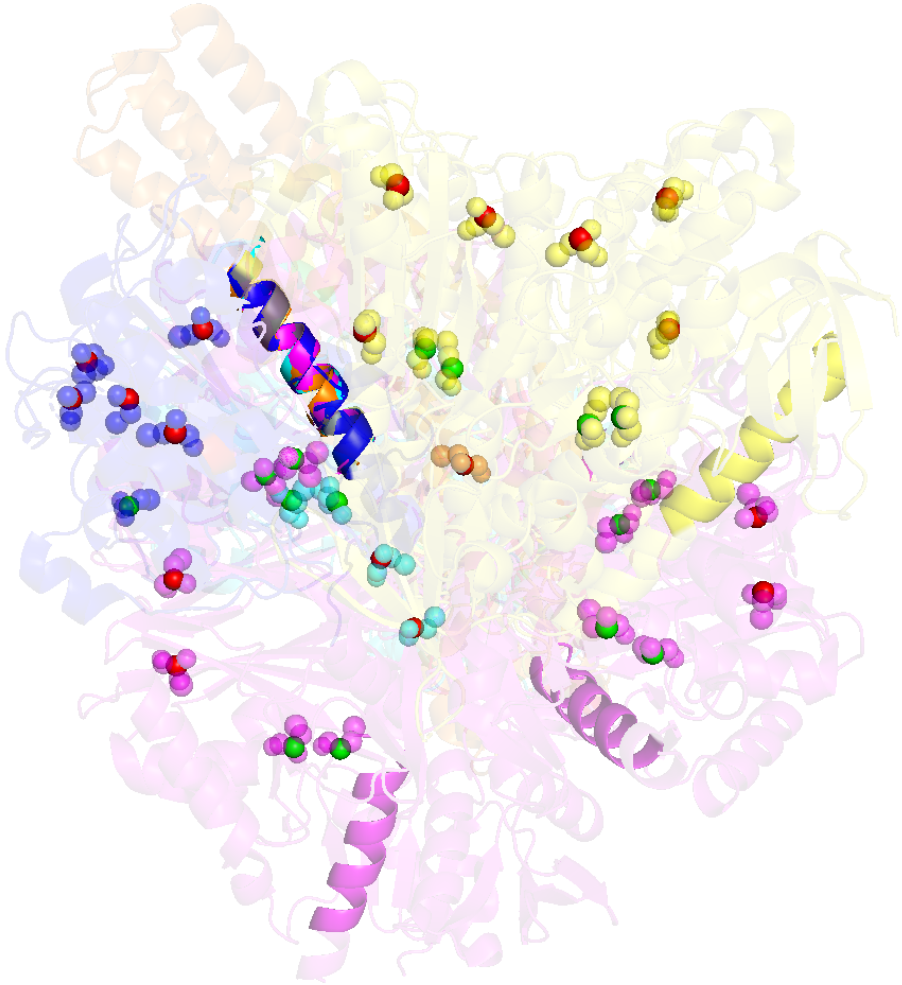

Figure 36: Structure of five proteins using PDB entries 2Q8K [64] (blue), 2HQM [113] (yellow), 1U8F [58] (magenta), 2QRJ [4] (cyan) and 6DFP [Kim *et al.*, to be published] (orange), with one shapemer typical for Cys+ aligned (multiple colors, upper left). Cysteine atoms are shown as spheres, with the C_α-atoms colored green for Cys+ and red for Cys−. The aligned shapemers are opaque, while the rest of the protein is transparent. When a protein contains multiple instances of the shapemer, unaligned shapemers are shown with lower transparency. We can see that Cys+ often seem clustered around the shapemer.

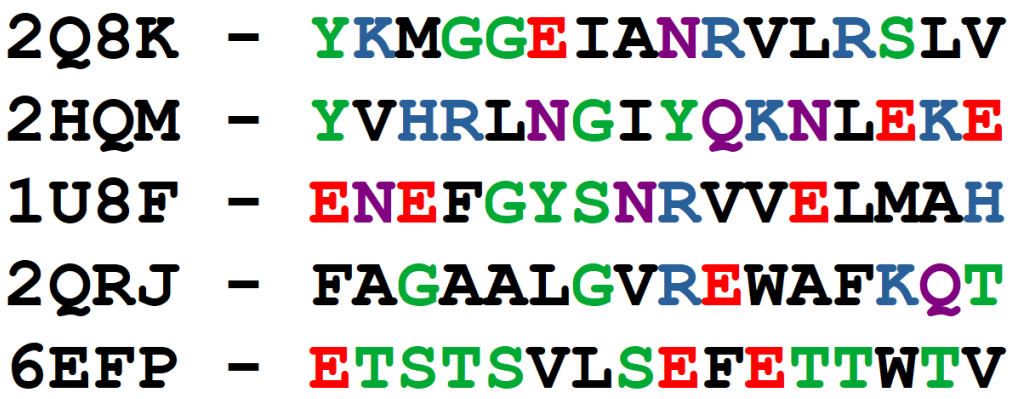

Figure 37: Sequence of a shapemer typical in proteins containing Cys+, found in PDB entries 2Q8K [64], 2HQM [113], 1U8F [58], 2QRJ [4] and 6DFP [Kim *et al.*, to be published]. Residues are colored according to chemical properties, with polar residues being colored in green, neutral in purple, basic in blue, acidic in red and hydrophobic in black.

4.2.2 Cysteines

We built an HMM by calculating matrices out of a random training subset of the data from protein set 1, using both sequence neighborhood and Euclidean space to determine the closest amino acids to Cys+ and Cys−, see Table 9 and 10. These matrices enabled us to assign a score to every cysteine dataset. We used this score as a feature for the RF and ET algorithms to build predictive models with a tree depth of 1. Applying these models to the remaining test set, we were able to achieve an AUC of 0.72 and 0.68 for the sequence method using RF and ET, respectively, and an AUC of 0.69 and 0.65 for the Euclidean method, likewise. See Figures 38 and 39 for the ROC curves and Figures 40 and 41 for the decision trees of the models.

Table 9: Probability matrix using sequence method

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
A	-0.02	-0.41	-0.21	-0.08	0.60	-0.17	-0.08	0.20	-0.13	0.19	-0.60	0.09	0.08	0.42	0.10	0.19	0.48	-0.02	0.22	-0.01
C	-0.39	-0.84	-0.99	0.44	-0.48	0.75	0.06	1.20	0.31	-0.41	-0.61	0.11	1.33	-0.06	0.54	-1.26	-0.14	-0.70	-0.22	-0.02
D	0.05	0.01	-0.07	-0.15	-0.01	-0.62	-0.18	0.07	-0.39	-0.08	0.11	0.07	-0.32	-0.18	-0.18	-0.28	0.15	-0.22	-0.06	-0.02
E	-0.17	0.04	-0.06	-0.53	-0.28	-0.62	-0.52	0.19	-0.44	-0.36	0.00	0.02	0.26	-0.07	-0.13	0.12	-0.20	0.34	0.21	-0.26
F	0.22	-0.01	0.31	0.32	0.05	0.64	0.44	-0.09	-0.55	-0.37	0.05	0.17	0.05	0.16	-0.06	0.14	-0.04	0.07	-0.05	0.68
G	0.34	-0.02	-0.11	0.12	-0.08	-0.07	0.21	0.15	0.48	0.23	0.49	0.09	0.45	-0.26	0.23	-0.61	0.16	0.00	-0.47	-0.02
H	0.08	-0.45	-0.23	-0.12	0.06	0.24	-0.09	0.06	0.03	0.71	-0.19	0.42	0.08	-0.06	-0.27	-0.26	0.02	0.08	0.15	-0.63
I	-0.17	0.21	0.10	0.03	0.42	0.11	0.17	-0.11	-0.06	-0.22	-0.38	0.19	-0.10	0.34	0.11	0.23	-0.01	-0.10	-0.09	0.26
K	0.14	0.00	0.29	0.21	0.28	0.13	-0.31	-0.67	-0.37	-0.50	0.39	0.11	-0.27	0.52	-0.12	0.10	0.09	0.33	0.22	0.25
L	-0.06	0.34	0.21	-0.22	0.06	-0.14	-0.29	-0.44	-0.10	-0.52	-0.16	-0.45	-0.11	-0.06	-0.09	0.33	-0.14	0.01	-0.22	0.24
M	-0.21	-0.35	0.10	0.32	-0.19	0.12	-0.96	-0.04	-0.10	-0.70	-0.31	0.05	-0.40	-0.14	0.42	0.51	-0.54	-0.07	-0.14	0.18
N	-0.24	-0.71	0.22	0.03	-0.35	0.22	-0.17	0.06	-0.17	0.04	-0.45	-0.03	-0.20	-0.33	-0.19	-0.36	-0.42	-0.10	0.45	-0.30
P	0.26	-0.13	0.11	0.11	-0.60	0.50	-0.10	-0.14	0.66	0.17	0.57	0.27	0.10	-0.09	0.17	-0.19	-0.20	0.67	-0.32	-0.65
Q	-0.57	0.28	-0.42	-0.05	0.15	-0.25	-0.28	-0.07	0.07	0.06	-0.24	-0.31	-0.33	0.04	-0.15	-0.68	0.08	-0.08	-0.27	-0.02
R	-0.14	0.13	0.05	0.36	0.03	0.03	0.08	0.06	-0.09	-0.49	-0.04	0.00	0.05	0.23	0.02	0.38	0.42	-0.20	0.21	0.27
S	0.08	-0.04	-0.20	-0.29	-0.11	0.30	0.59	-0.18	0.25	0.39	0.38	0.15	-0.29	0.16	0.21	-0.23	0.05	-0.28	0.06	-0.19
T	-0.15	0.25	-0.34	0.06	-0.22	-0.54	-0.17	0.19	0.08	0.21	0.18	-0.30	0.16	-0.07	0.09	0.21	-0.14	0.05	0.35	-0.26
V	0.34	0.42	0.29	0.18	0.02	0.07	-0.08	0.10	0.43	0.18	-0.06	-0.23	-0.21	-0.32	0.23	0.33	-0.24	-0.19	0.10	-0.32
W	0.31	0.07	0.06	-0.05	-0.29	-0.74	1.34	-1.01	-1.01	1.38	-0.37	0.16	-0.30	-0.07	-0.74	-0.17	0.04	-0.25	-0.12	0.18
Y	0.12	-0.21	0.03	0.00	0.03	0.11	0.36	-0.45	-0.04	-0.02	-1.03	0.44	-0.77	-0.24	-0.12	0.46	0.45	0.36	0.36	0.58
-	-0.53	-0.47	-0.61	-0.90	-1.15	-1.51	-1.59	0.00	0.00	0.00	0.00	0.00	-2.80	-1.82	-1.37	-0.84	-0.76	-0.65	-0.55	-0.56

NOTE: **Protein set 1 was used to calculate values. Top row: position of the AA in the primary structure, cysteine being situated between position 10 and 11. Left-most column: one letter codes for different types of AAs, - shows a missing AA. Numerical values are the natural logarithm of the ratio of the likelihoods to find the specific AA at the position around a Cys+ or Cys−.**

Table 10: Probability matrix using Euclidean method

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
A	-0.13	0.08	-0.33	-0.07	0.06	0.05	-0.03	0.25	0.11	0.05	-0.13	0.10	0.04	0.43	0.16	-0.30	0.03	0.03	0.12	0.04
C	0.29	0.32	-0.04	0.01	0.07	-0.07	-0.12	-0.02	0.03	0.02	-0.29	0.00	-0.05	-0.36	-0.55	0.20	0.16	0.75	-0.20	-0.51
D	0.50	0.03	0.13	-0.29	-0.26	-0.20	0.00	-0.47	0.25	0.10	-0.19	0.08	-0.13	-0.19	0.13	0.31	0.32	-0.25	-0.04	0.23
E	-0.39	-0.37	-0.08	-0.15	0.07	0.04	-0.09	0.37	0.23	-0.24	-0.09	0.06	0.14	-0.18	0.15	-0.24	-0.32	-0.34	0.17	0.04
F	-0.26	-0.70	-0.07	-0.24	-0.17	-0.21	-0.29	-0.10	0.11	-0.12	0.02	0.17	0.20	0.14	-0.05	-0.02	0.33	0.19	0.17	0.10
G	0.00	0.26	0.15	0.56	0.38	0.37	0.33	0.37	0.17	0.26	0.21	-0.13	0.15	0.02	0.19	0.14	0.27	0.23	0.15	0.24
H	0.18	0.58	0.19	0.00	0.25	0.79	0.63	0.27	0.28	0.02	0.32	-0.13	0.02	-0.01	-0.21	0.09	-0.50	0.09	-0.04	-0.06
I	-0.06	-0.36	-0.02	-0.10	0.19	-0.14	-0.32	-0.23	0.10	-0.02	-0.35	0.13	0.10	-0.10	0.08	-0.18	-0.19	-0.16	0.00	-0.02
K	0.04	-0.01	0.05	0.14	0.08	0.01	-0.01	0.39	0.08	-0.04	0.15	0.22	-0.02	-0.09	-0.11	0.06	-0.31	0.47	-0.04	0.44
L	-0.74	-0.26	-0.56	-0.39	-0.31	-0.40	-0.34	-0.23	-0.29	-0.36	-0.44	-0.12	-0.23	-0.02	-0.33	-0.18	-0.20	-0.41	-0.09	-0.42
M	-0.31	-0.21	-0.42	-0.14	0.02	0.10	-0.59	-0.42	-0.19	-0.49	-0.09	-0.18	-0.29	0.47	0.16	-0.59	-0.22	0.42	-0.55	0.16
N	0.22	0.32	0.27	-0.21	0.05	-0.48	-0.16	0.15	0.27	0.13	-0.01	0.01	0.07	-0.20	-0.17	-0.07	-0.14	0.16	0.37	-0.28
P	0.32	0.26	0.83	0.25	0.19	0.31	0.51	0.38	0.05	-0.10	0.22	0.45	-0.66	-0.17	0.12	-0.19	-0.05	-0.19	-0.31	-0.08
Q	-1.52	0.00	-0.34	0.16	-0.65	0.03	0.06	-0.14	-0.11	0.32	0.52	-0.24	-0.55	-0.04	0.13	0.08	-0.11	0.23	-0.33	-0.13
R	0.37	-0.26	-0.26	0.08	0.15	-0.05	0.05	-0.62	-0.14	-0.02	0.19	-0.31	0.05	-0.23	0.15	0.20	0.28	-0.16	0.05	-0.41
S	0.23	-0.03	0.54	0.42	0.35	0.06	0.30	-0.08	0.25	0.53	0.16	-0.15	0.34	-0.14	0.04	-0.06	0.35	-0.09	0.06	-0.08
T	0.21	0.19	0.23	0.15	0.03	0.52	0.01	-0.11	-0.45	-0.02	-0.11	0.13	0.17	-0.01	0.09	-0.05	0.06	-0.54	-0.06	0.12
V	-0.62	-0.36	-0.09	-0.02	-0.34	-0.15	0.13	-0.06	-0.06	0.17	0.08	-0.39	0.02	-0.14	-0.08	0.06	-0.12	-0.07	-0.21	0.03
W	-0.77	0.33	0.45	-0.22	0.30	0.65	-0.21	-0.21	-0.08	0.43	0.75	0.29	0.21	0.86	0.08	0.46	0.25	-0.33	0.31	-0.16
Y	-0.25	0.33	-0.15	-0.28	0.02	-0.21	0.13	-0.02	-0.40	-0.31	-0.06	0.13	0.10	-0.01	-0.09	0.46	-0.35	0.21	0.26	0.40

NOTE: **Protein set 1 was used to calculate values. Top row: position of the AA in order of Euclidean distance from the investigated cysteine, with AA 01 being closest. Left-most column: one letter codes for different types of AAs. Numerical values are the natural logarithm of the ratio of the likelihoods to find the specific AA at the position around a Cys+ or Cys−.**

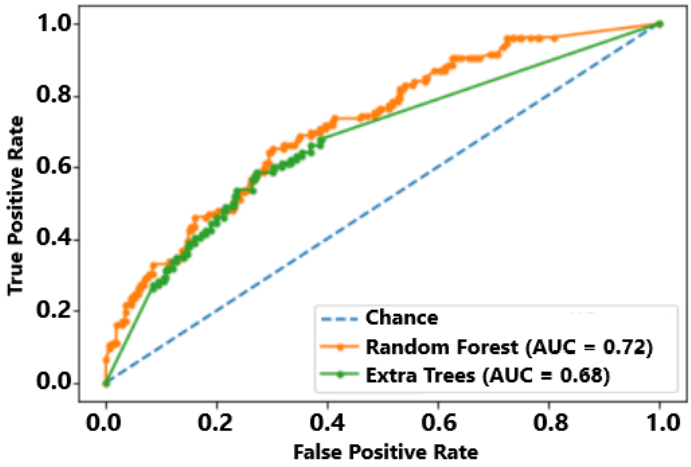

Figure 38: ROC curve of the RF and ET algorithms for the protein set 1 using HMM data applying the sequence method. Blue line shows a completely random prediction.

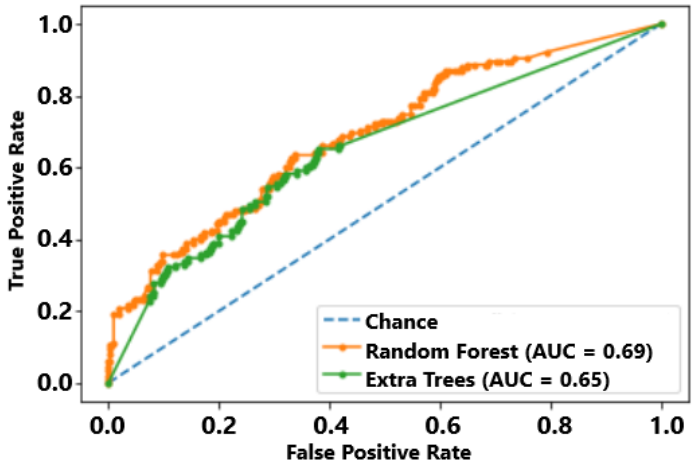

Figure 39: ROC curve of the RF and ET algorithms for the protein set 1 using HMM data applying the Euclidean method. Blue line shows a completely random prediction.

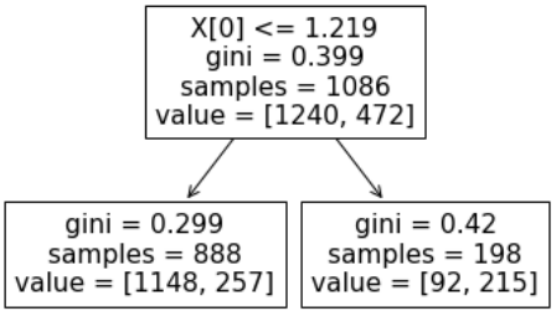

Figure 40: Decision tree of the RF algorithm for the protein set 1 using HMM data applying the sequence method. X[0] shows the cutoff value for the prediction of redox-activity, gini the Gini coefficient, samples the number of cysteines in the group and value the number of experimentally verified Cys− and Cys+ among the samples. The upper box shows the full dataset, the lower left box shows the cysteines predicted by the model as Cys−, the lower right box as Cys+.

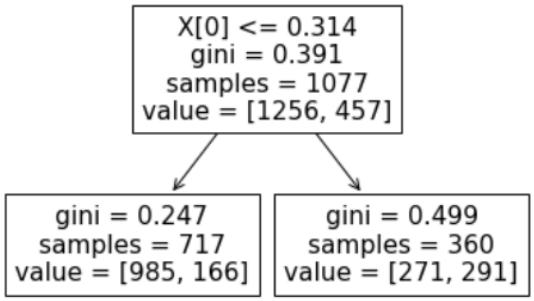

Figure 41: Decision tree of the RF algorithm for the protein set 1 using HMM data applying the Euclidean method. X[0] shows the cutoff value for the prediction of redox-activity, gini the Gini coefficient, samples the number of cysteines in the group and value the number of experimentally verified Cys− and Cys+ among the samples. The upper box shows the full dataset, the lower left box shows the cysteines predicted by the model as Cys−, the lower right box as Cys+.

I applied the SVM, RF, ET and GB algorithms after preprocessing and

feature selection to protein set 1 and 2 to be able to infer redox-sensitive cysteine sites in proteins and compared their results. The best AUC value for the ROC curve was 0.72 for the imputed dataset using the ET and SVM algorithms, see Figure 42 and Table 11. The ROC curves depict average values over the different cross-validation folds. The models created in this step will be used for the prediction of redox cysteines in Chapter 4.3: Use Cases.

Table 11: AUC of the three different algorithms

	SVM	RF	ET	GB	$\bar{x}^b$ dataset
Structure	0.66	0.69	0.7	0.67	0.68
Sequence	0.69	0.70	0.71	0.69	0.70
Imputation (Seq.)	$\underline{0.72}^a$	0.71	$\underline{0.72}^a$	0.71	0.72
$\bar{x}^b$ algorithm	0.69	0.70	0.71	0.69	0.70

NOTE: ^{a}highest result underlined, ^{b} average.

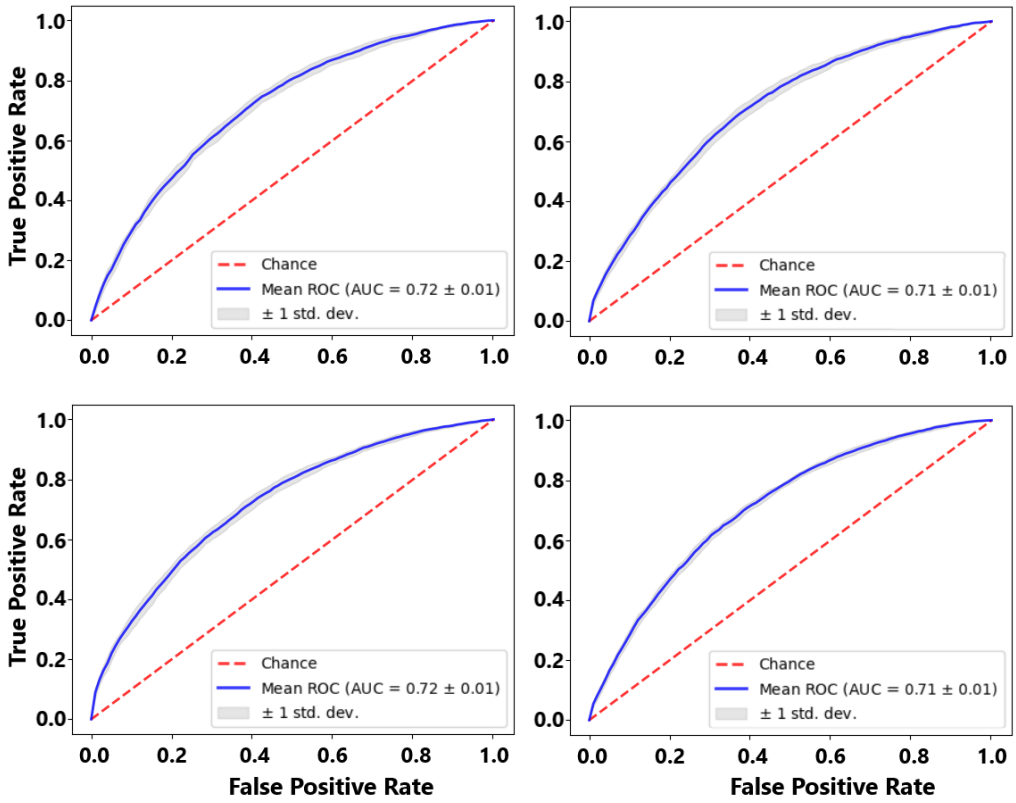

Figure 42: ROC curves of the SVM (upper left), RF (upper right), ET (lower left) and GB (lower right) algorithms for the imputed dataset. Red line shows a completely random prediction. The ROC curve shows average values over the different cross-validation folds.

I trained models using different numbers of neighboring AAs to assess which AAs were the most relevant for training and prediction. For the imputation method, I investigated the closest ten to twenty AAs found in the primary structure. For the Euclidean method, I investigated the closest seven to nineteen AAs in Euclidean space. I repeated this assessment using different proportions of protein set 1 and 2 to understand if a larger amount of data would lead to better results. For both methods, I investigated 50% to 100% of the protein set. I applied the SVM algorithm with the imputation method and the ET algorithm with the Euclidean method, as they previously resulted in the most accurate results. Predictions were made both on

a random test set and on complex I. The random test set was used due to its large size, resulting in more reliable results. The complex I set was used to test the models on data originating from a different source than the rest of the data. Due to its small size, results may be less reliable and more prone to outliers.

I found that, using the imputation method, a larger number of AAs considered would slightly improve results using the test set, while predictions stayed roughly the same for the Euclidean method. I found strong improvements for the predictions of complex I using the imputation method for a higher number of AAs, while the Euclidean method showed the most promising results for 13 or 15 AAs, see Figures 43 and 44.

The Euclidean methods showed strong improvements when considering a larger proportion of the protein sets when predicting complex I cysteines. but no improvement on the test set. The imputation method showed small improvements predicting the test set, but no improvement on complex I, see Figures 45 and 46.

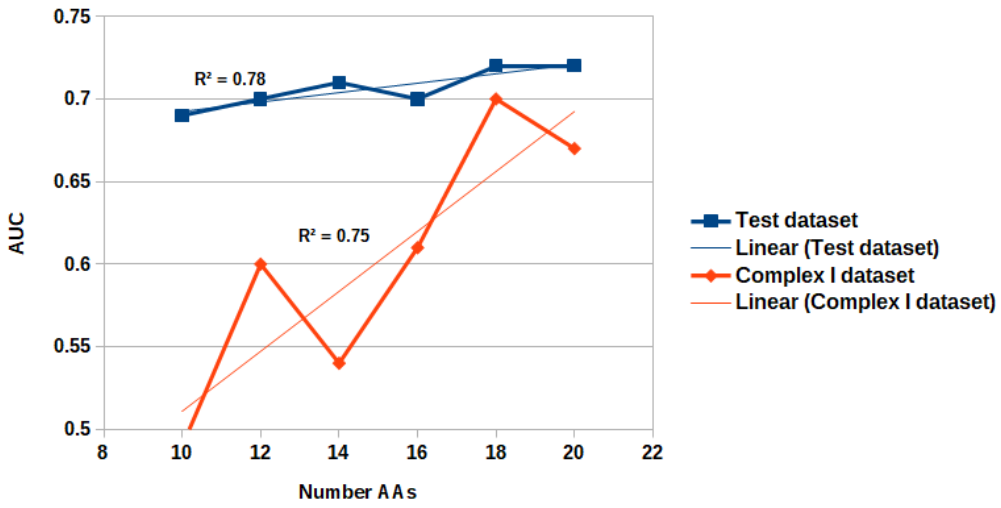

Figure 43: AUC values for different numbers of closest AAs in the sequence around the central cysteine, applying the SVM algorithm. Thick blue line shows AUC for test dataset, thick orange line for complex I. Thin lines are trendlines. R^2 the coefficient of determination, i.e. how well the trendline corresponds to the data.

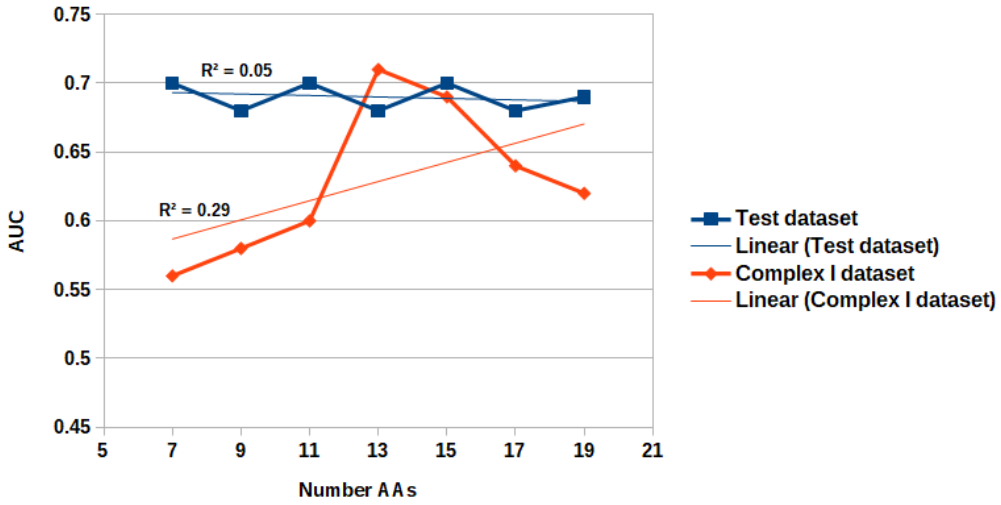

Figure 44: AUC values for different numbers of closest AAs in Euclidean space around the central cysteine, applying the SVM algorithm. Thick blue line shows AUC for test dataset, thick orange line for complex I. Thin lines are trendlines. R^2 the coefficient of determination, i.e. how well the trendline corresponds to the data.

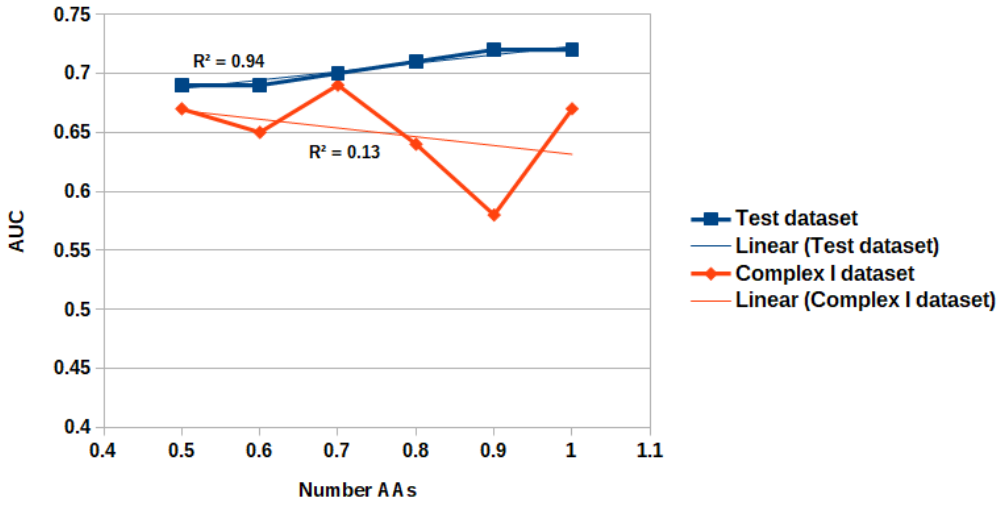

Figure 45: AUC values using the imputation method for different amounts of data, with x = 1.0 denoting the full dataset and x = 0.5 50% of the dataset, applying the SVM algorithm. Thick blue line shows AUC for test dataset, thick orange line for complex I. Thin lines are trendlines. R^2 the coefficient of determination, i.e. how well the trendline corresponds to the data.

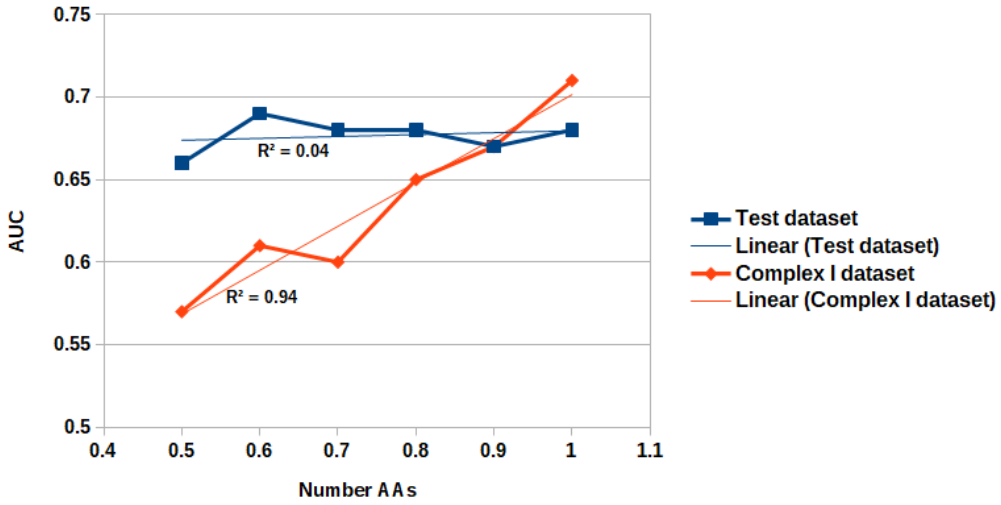

Figure 46: AUC values using the Euclidean method for different amounts of data, with x = 1.0 denoting the full dataset and x = 0.5 50% of the dataset, applying the SVM algorithm. Thick blue line shows AUC for test dataset, thick orange line for complex I. Thin lines are trendlines. R^2 the coefficient of determination, i.e. how well the trendline corresponds to the data.

We utilized the Geometricus tool to turn the local neighborhood of cysteines into structural moment invariants, basing the calculations on either the sequence neighborhood of 16 AAs or the Euclidean neighborhood with a radius of 16 Å. These invariants were then used as features for the ET and RF algorithms, creating models for the predictions of Cys+. Models created with the RF algorithm were able to produce superior predictions with the test dataset. The sequence-based method showed better performance than the radius-based method, which was only marginally better than chance, see Figure 47 and 48. It may be beneficial to incorporate sequence-based invariants as a feature into other models.

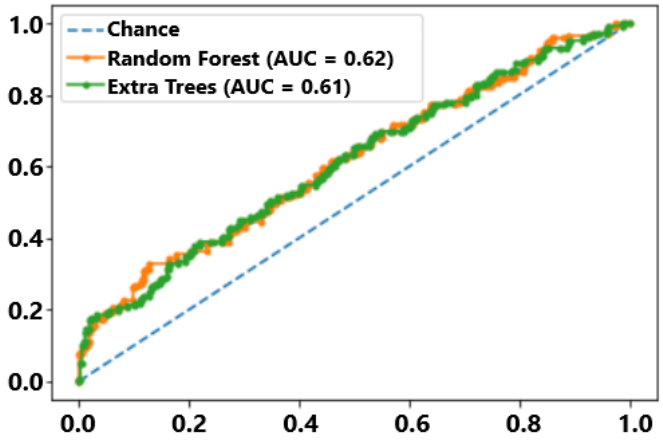

Figure 47: ROC curve of the RF and ET algorithms for the protein set 1 using sequence-based moment invariants. Blue line shows a completely random prediction.

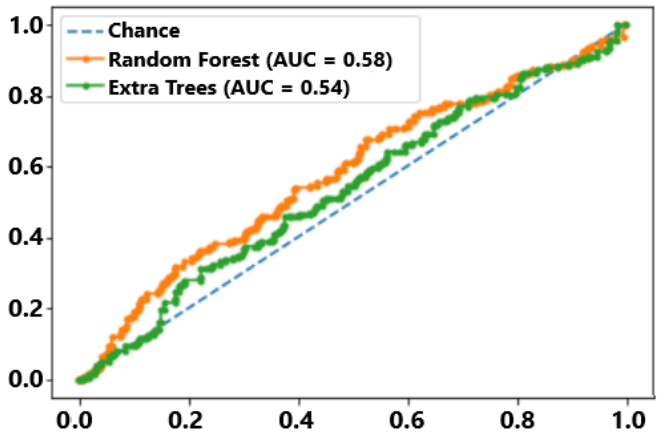

Figure 48: ROC curve of the RF and ET algorithms for the protein set 1 using radius-based moment invariants. Blue line shows a completely random prediction.

4.3 Use Cases

4.3.1 Complex I

After using different algorithms to create models for the prediction of redox-sensitive cysteines, I applied the models to generate predictions for the *NDUFS1*, *MT-ND3* and *NDUFA2* subunits of mammalian respiratory complex I, based on the structural data from PDB entries 6G2J [2], 6G72 [2], 5LC5 [114], 5LNK [38] and 5XTD [49].

Due to their high AUC value on the test set, I applied the models trained with the ET and SVM algorithms to complex I, using the parameters and training data detailed in the previous sections. Training was accomplished using the protein sets 1 and 2, respectively. For illustration, I used the PDB entry with the accession number 5XTD of mitochondrial complex I [49], as it is the most complete structure of human origin.

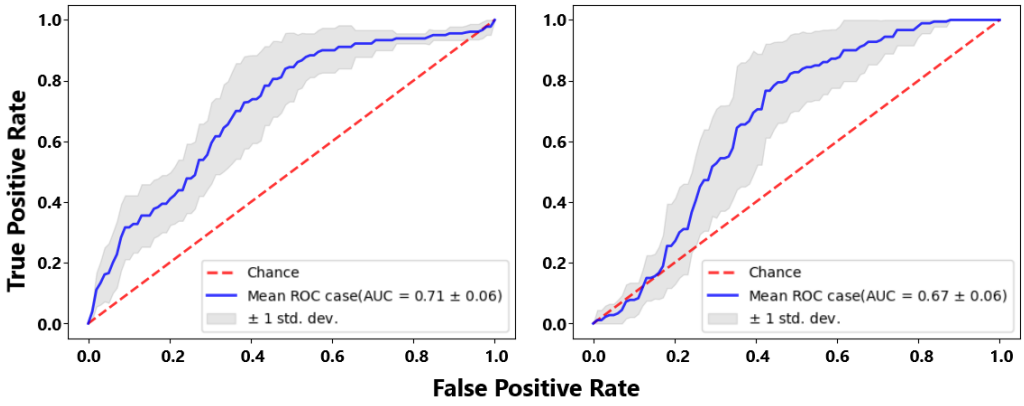

Figure 49: ROC curves of the ET (left) and SVM (right) algorithms for complex I trained on the Euclidean and imputed dataset, respectively. Red line shows a completely random prediction. The ROC curve shows average values over the different cross-validation folds.

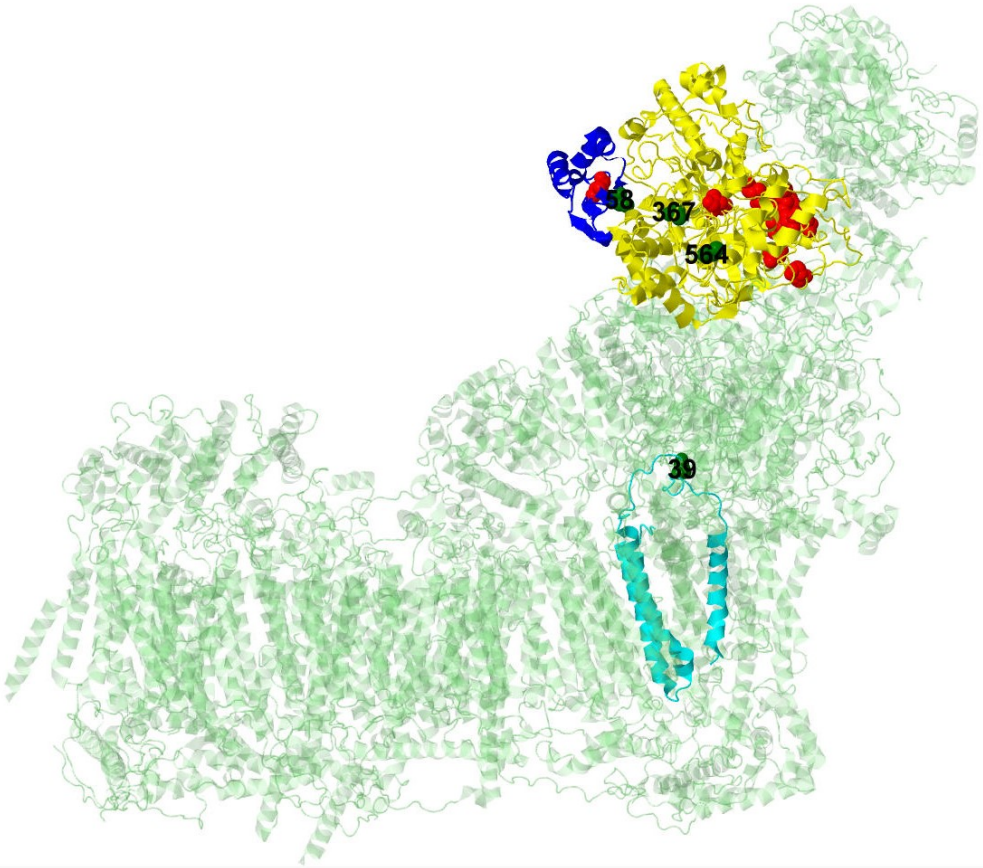

Figure 50: 3D structure of human complex I using PDB entry 5XTD [49]. The subunits *NDUFS1*, *MT-ND3* and *NDUFA2* are colored yellow, cyan and blue, respectively. Green dots indicate the positions of experimentally verified [21] Cys+. Red dots indicate the positions of Cys−.

I compared the predictions of the SVM and ET algorithms to the experimental results from Chouchani *et al.* [21]. 77% of the cysteines of the relevant subunits were predicted correctly by ET and 72% by SVM. For the predictions of PDB entry 5XTD, see Table 12. For the 3D structure of complex I with cysteines experimentally validated to be modified, see Figure 50. For a closer look at the structural features of CYS367, CYS554 and CYS564 in subunit *NDUFS1*, see Figure 51.

Table 12: Experimental data of Cys+ in human complex I compared to predictions of PDB entry 5XTD

Validation[a]	ET[b]	SVM[c]	SU/CYS[d]
-	-	-	*NDUFS1*/53
-	-	-	*NDUFS1*/64
-	-	+	*NDUFS1*/75
-	-	-	*NDUFS1*/78
-	-	-	*NDUFS1*/92
-	-	-	*NDUFS1*/128
-	+	-	*NDUFS1*/131
-	-	-	*NDUFS1*/137
-	-	-	*NDUFS1*/176
-	-	-	*NDUFS1*/179
-	-	-	*NDUFS1*/182
-	-	-	*NDUFS1*/226
+	-	-	*NDUFS1*/367
-	-	+	*NDUFS1*/554
+	+	-	*NDUFS1*/564
-	-	-	*NDUFS1*/710
+	+	-	*MT-ND3*/39
-	+	+	*NDUFA2*/24
+	-	+	*NDUFA2*/58

NOTE: "+" for Cys+, "-" for Cys−, [a]experimental validation, [b]ET trained on protein set 1, [c]SVM trained on protein set 2, [d]subunit/cysteine sequence position.

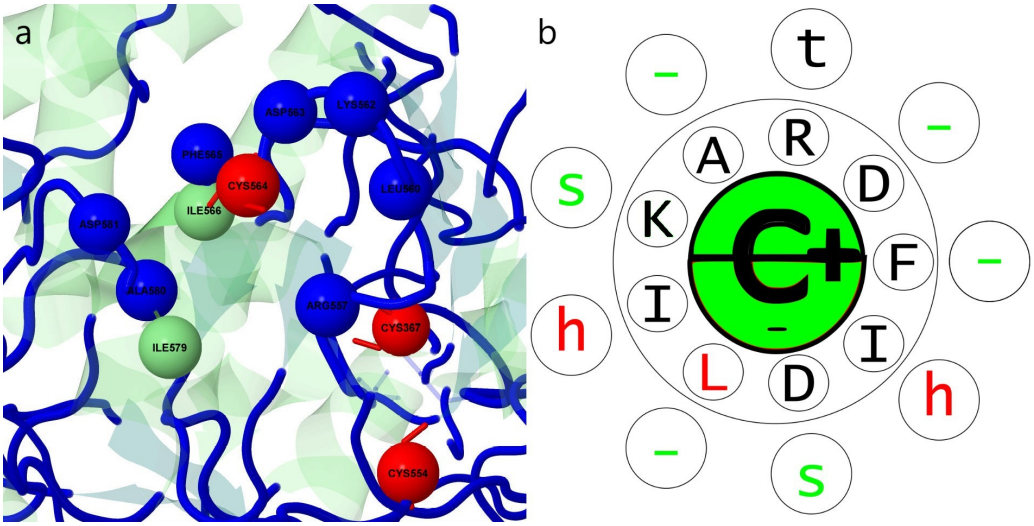

Figure 51: a: Three Cys+, CYS367, CYS554 and CYS564, in subunit *NDUFS1* of human complex I represented by red balls. All three cysteines are found in loop regions (blue). Other SSEs are shown in green. The nine closest amino acids to CYS564 are also shown as large balls. PDB entry 5XTD [49] was utilized for visualization of the structure. b: Structural features of CYS564 and its surrounding amino acids for comparison. Innermost circle shows the cysteine (C) with its secondary structure below and the experimental validation to the right. The half spheres show HSE1 (upper) and HSE2 (lower), a green half sphere signifies a lower number of proximate residues and a red one a higher number. The middle circle shows the surrounding residues, in the order of closeness, starting at the top, going clockwise. The outer circle shows the corresponding SSE assignments. A green symbol signifies a feature which is statistically more likely to occur in or near a Cys+, a red symbol in or near a Cys−, and a black symbol has a roughly equal likelihood for both.

4.3.2 NKG2E Natural Killer Cell Receptor

I trained a model on dataset 2 using the imputation method and used it to predict redox-activity in NKG2E to examine the impact of SNPs on my method.

I investigated four different variants for this purpose, which are combinations of wild type (WT) and alternative (ALT) forms:

- WT rs2682489 and WT rs286266404
- ALT rs2682489 and ALT rs28626640
- ALT rs2682489 and WT rs28626640
- WT rs2682489 and ALT rs28626640

These variants differ at two positions, only one of which is close enough to a cysteine to be relevant in my model, see Figure 52.

Only the cysteine at sequence position 235 was predicted to be more likely to be redox-active than not. All cysteines received very similar prediction values for both variants, see Table 13. The cysteine at position 129, which is located close to the SNP, showed by far the greatest difference between prediction values.

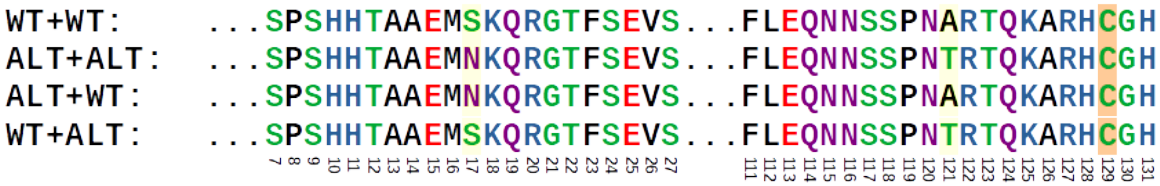

Figure 52: Amino acid sequence around the two SNPs of four NKG2E variants. Residues are colored according to chemical properties, with polar residues being colored in green, neutral in purple, basic in blue, acidic in red and hydrophobic in black. SNPs highlighted in yellow, cysteine in orange. First line shows wildtype/wildtype variant, second shows alternative/alternative, third shows alternative/wildtype and fourth shows wildtype/alternative.

Table 13: Average prediction values of NKG2E variants

Sequence position	Prediction value	Difference
76	0.29	$3.31 \cdot 10^{-4}$
95	0.17	$1.17 \cdot 10^{-4}$
129	**0.35**	$\mathbf{1.36} \cdot 10^{-2}$
132	0.36	$2.08 \cdot 10^{-3}$
143	0.44	$2.66 \cdot 10^{-3}$
160	0.29	$3.15 \cdot 10^{-4}$
170	0.25	$2.91 \cdot 10^{-4}$
222	0.35	$4.12 \cdot 10^{-4}$
235	0.54	$4.11 \cdot 10^{-4}$
280	0.45	$1.46 \cdot 10^{-3}$
282	0.41	$7.56 \cdot 10^{-4}$

NOTE: [a]prediction values range from 0 (not redox-active) to 1 (redox-active), [b]difference between the predictions of variant 1/3 and 2/4.

4.3.3 Investigation of Proximal Tubule Cell Proteins

Using a model trained with the ET algorithm on the protein set 2, I predicted the redox modifiability of all cysteines in the imputed proximal tubule cell dataset to test the models and provide novel predictions for this dataset, see Table 14. The dataset was provided and is currently being researched by a research group led by Dr. Flávia Rezende to find the physiological function of the NADPH oxidase 4 in the kidney. The average and median prediction scores were 0.33 and 0.32, respectively, see Figure 53. I repeated the process for any proteins with a PDB entry, using the Euclidean method instead of imputation. The average and median prediction scores were 0.40 and 0.41, respectively, see Figure 54. The Pearson correlation coefficient between the two methods for the same set of cysteines was 0.29. See Figure 57 for the differences between prediction scores between the Euclidean and imputation-based methods.

Table 14: Cysteines predicted as redox-active in proximal tubule cells using ET

Protein name	ID	Cys	P	Molecular function
Peroxisomal acyl-coenzyme A oxidase 1	Q15067	198	0.82	Acyl-CoA oxidase activity
Guanylate cyclase activator 2B	Q16661	66	0.68	Guanylate cyclase activator
Folate receptor alpha	P15328	64	0.68	Drug binding, folic acid binding
Glycine N-acyltransferase	Q6IB77	292	0.68	Glycine N-acyltransferase
Guanylate cyclase activator 2B	Q16661	79	0.67	Guanylate cyclase activator
Folate receptor alpha	P15328	65	0.67	Drug binding, folic acid binding
Transmembrane protein 174	Q8WUU8	228	0.65	
Solute carrier family 22 member 6	Q4U2R8	48	0.64	Anion, chloride ion binding
Sodium-dependent neutral AA transporter	G5E8X1	161	0.64	
3-hydroxybutyrate dehydrogenase type 2	Q9BUT1	92	0.64	3-hydroxybutyrate dehydrogenase
Ribonuclease P protein subunit p21	Q9H633	81	0.63	Metal ion binding, ribonuclease P
Ribonuclease P protein subunit p21	Q91WH2	282	0.63	Glucuronosyltransferase
Sodium-dependent neutral AA transporter	Q96N87	11	0.63	AA transmembrane transporter
Growth hormone receptor	P10912	439	0.63	cytokine binding, cytokine receptor
Solute carrier family 7 member 13	Q8TCU3	126	0.63	L-AA transmembrane transporter
Transmembrane protein 150A	Q86TG1	262	0.62	
Somatotropin	P01241	214	0.62	Growth factor
Ribonuclease P protein subunit p21	Q9H633	74	0.62	metal ion binding, ribonuclease P
Guanylate cyclase activator 2B	Q16661	110	0.62	Guanylate cyclase activator
DnaJ homolog subfamily C member 12	Q9UKB3	40	0.62	
Folate receptor alpha	P15328	151	0.62	drug binding, folic acid binding
Glutaryl-CoA dehydrogenase, mitochondrial	Q8C3M5	83	0.61	
Beta-defensin 129	Q9H1M3	37	0.61	
Glutaryl-CoA dehydrogenase, mitochondrial	Q8C3M5	89	0.61	
Glycerol-3-phosphate dehydrogenase	P21695	167	0.61	Glycerol-3-phosphate dehydrogenase
Coiled-coil domain-containing protein 107	Q8WV48	211	0.61	
Bifunctional epoxide hydrolase 2	P34913	231	0.61	Phosphatase
Transmembrane protein 150A	Q86TG1	43	0.61	
Guanylate cyclase activator 2B	Q16661	107	0.60	Guanylate cyclase activator
Transmembrane protein 106A	Q96A25	246	0.60	
Growth hormone receptor	P10912	55	0.60	cytokine binding, cytokine receptor
3-ketoacyl-CoA thiolase B, peroxisomal	P07871	25	0.60	acetyl-CoA C-acetyltransferase

NOTE: Model trained on protein set 2, only proteins with P $>$ 0.6 are shown. [a]Uniprot ID, [b]cysteine sequence position, [c]prediction value.

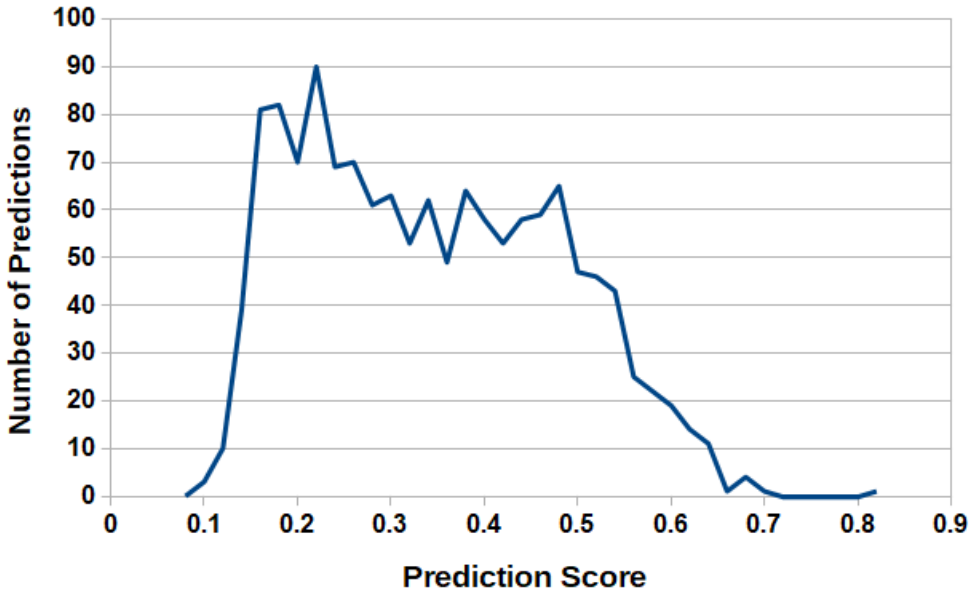

Figure 53: Frequency of predictions by the ET algorithm for the proximal tubule cell dataset. The model was trained on the protein set 2. A higher prediction score corresponds to a higher likelihood of a cysteine being a Cys+.

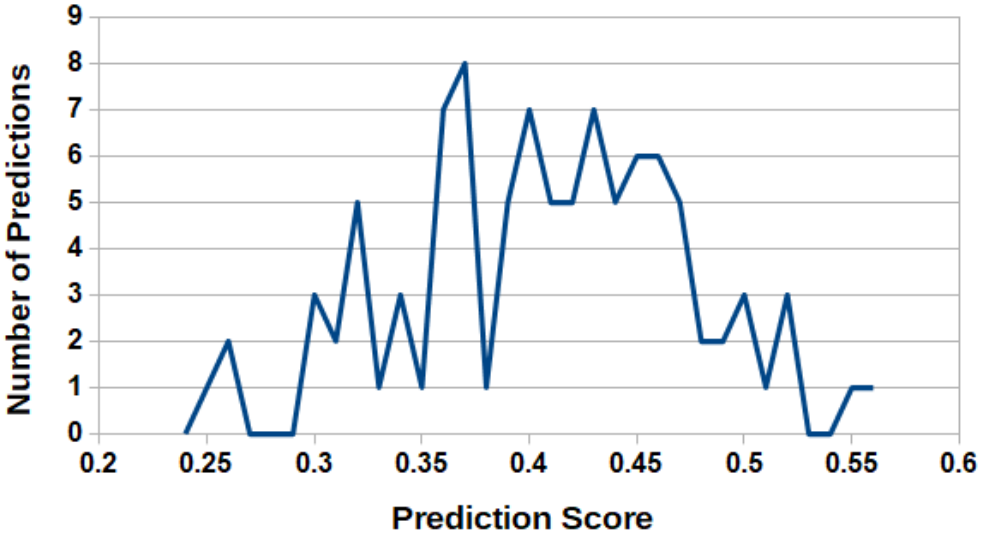

Figure 54: Frequency of predictions by the ET algorithm for the proximal tubule cell dataset. The model was trained on the protein set 2. Only proteins with PDB entry are included. A higher prediction score corresponds to a higher likelihood of a cysteine being a Cys+.

I took a closer look at some specific proteins to gain a better understanding of the reasons behind the different predictions by the two methods I used to analyze proteins with a PDB entry. Folate receptor alpha with the PDB ID 4KM6 [107] shows a relatively complete structure, with 208 out of 257 amino acid positions known. Cysteines in loop regions on the surface of the protein tended to receive a very high prediction score using the PDB and DSSP files, while cysteines that were part of an SSE or in the interior of the protein got lower scores, see Figure 55. In this case, the method using PDB and DSSP data is likely superior to the imputed method.

Low-density lipoprotein receptor-related protein 2 with the PDB ID 2M0P [29] shows a very incomplete structure, with only 52 out of 4655 amino acid positions known. All cysteines appear on the surface of this incomplete structure. As they are also found in loop regions, prediction values for all cysteines are very high, see Figure 56. These predictions are based on false data in respect to close AAs, SSEs and accessibility, so the imputed method is likely superior for very incomplete structures as well as cysteines close to a missing part of the structure in the PDB entry. After removing all

proteins with less than 90% of the sequence being represented in the PDB entry, the Pearson correlation coefficient between the two methods rose to 0.42. See Figure 58 for the differences between prediction scores between the PDB-based and imputation-based methods when lower-quality PDB files were removed.

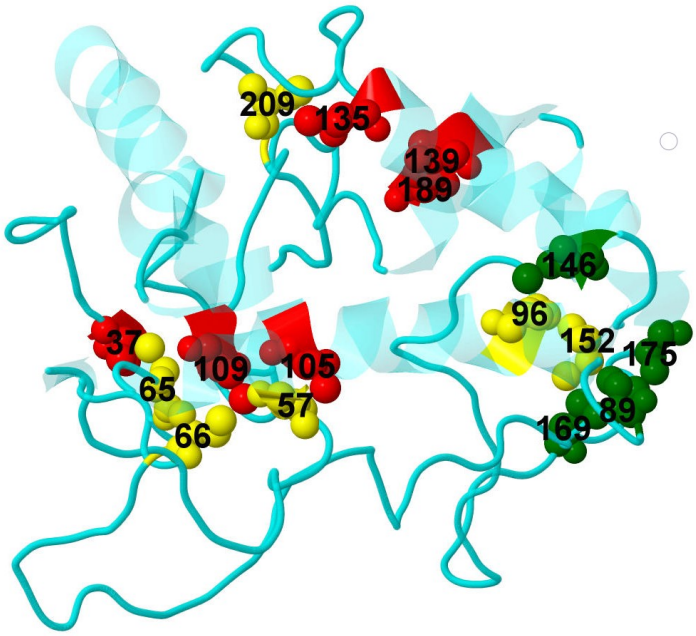

Figure 55: Structure of PDB entry 4KM6 [107] in cartoon representation. Cysteines with a prediction value higher than 0.7 are green, between 0.5 and 0.7 are yellow, below 0.5 are red. It appears that cysteines with a high prediction value tend to be on the surface of the amino acid. They also tend to be part of loop regions.

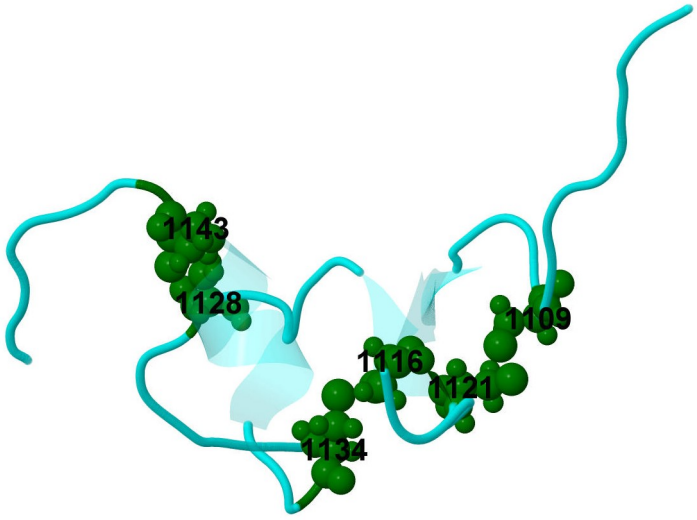

Figure 56: Structure of PDB entry 2M0P [29] in cartoon representation. All cysteines had a prediction value higher than 0.7 and are shown in green. Since the PDB entry is missing most of the structure, all cysteines appear to be on the surface, leading to high prediction values.

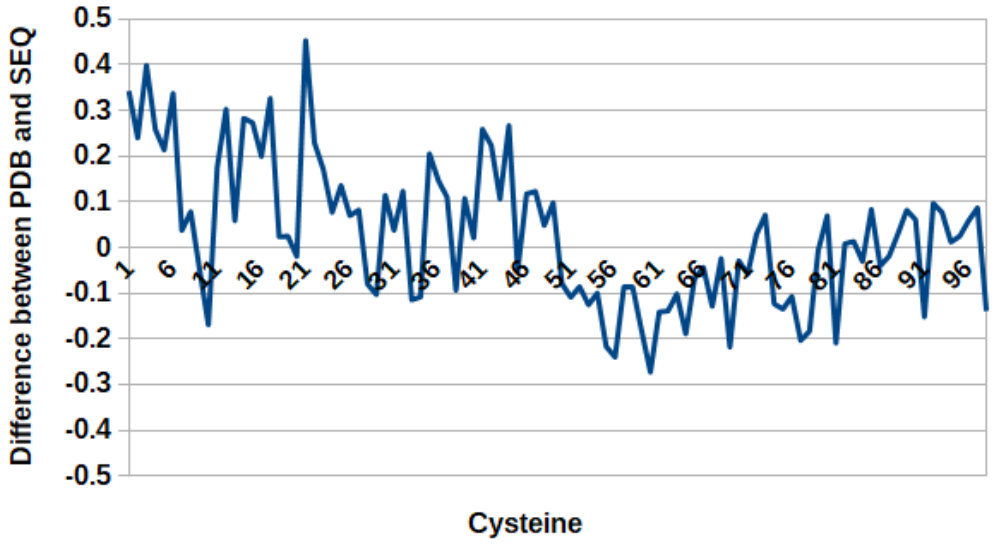

Figure 57: Difference between prediction scores for cysteines using the PDB-based (PDB) and imputation-based (SEQ) methods.

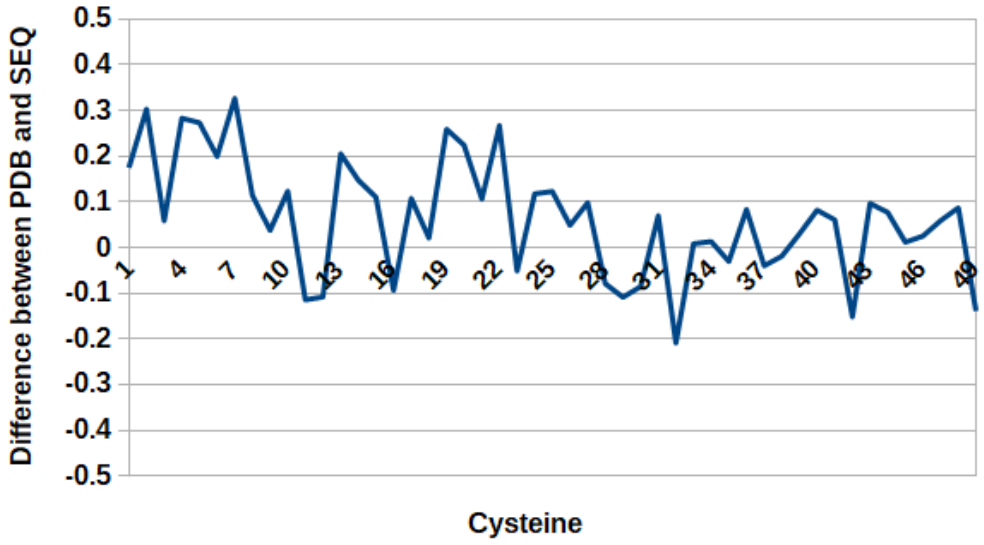

Figure 58: Difference between prediction scores for cysteines using the PDB-based (PDB) and imputation-based (SEQ) methods when lower-quality PDB files are removed.

I applied the model created using the ET algorithm trained with protein set 3 to the proximal tubule cell dataset to predict the likelihood of redox-activity of the proteins. Out of 119 proteins, 60 received a prediction value greater than 0.5, see Table 15. The mean prediction was 0.48 and the median 0.50 with a standard deviation of 0.14. For the structures of PDB entry 5AM2 of the bifunctional epoxide hydrolase 2 (EPHX2), which was predicted to likely be redox-active, and PDB entry 6M17 of the solute carrier family 6 member 19 (SLC6A19), which was predicted to likely not be redox-active, see Figures 59 and 60. It may be possible to combine predictions for redox-active proteins and cysteines for more accurate results.

ID	P	ID	P	ID	P	ID	P
LDHD	0.77	ACY3	0.58	4931406C07RIK	0.49	UGT2B38	0.39
FBP1	0.74	SLC4A4	0.57	GHR	0.49	HSPE1	0.39
LAP3	0.74	KCNJ15	0.57	SLC13A3	0.48	DEFB29	0.38
SORD	0.72	PIPOX	0.56	ASS1	0.48	TMEM106A	0.38
PDZK1	0.70	CALML4	0.56	SLCO1A6	0.48	TTC36	0.38
ACAA1B	0.70	LRP2	0.56	CAR14	0.48	SLC7A7	0.38
AKR1A1	0.69	HAO2	0.56	CELA1	0.47	GSTZ1	0.38
UPB1	0.69	NEU1	0.56	FOLR1	0.47	GUCA2B	0.35
KHK	0.69	UGT3A2	0.56	TCN2	0.46	MIOX	0.35
AKR1C21	0.69	PRSS8	0.55	THEM7	0.46	SLC47A1	0.34
GPX1	0.68	HSD3B2	0.55	UGT3A1	0.45	SLC22A30	0.33
SCP2	0.67	SLC22A12	0.54	TMEM174	0.45	CDA	0.33
SLC27A2	0.66	RAB11FIP3	0.54	SLC5A8	0.45	SNHG11	0.32
ACOX3	0.66	PECR	0.54	GLYAT	0.45	GM11128	0.32
ACSM2	0.66	PCK1	0.53	HRSP12	0.44	CML1	0.32
AKR7A5	0.65	MEP1A	0.53	CES1F	0.44	NAT8	0.32
ALDH6A1	0.65	GM10804	0.53	CES1D	0.43	SLC37A4	0.32
DNASE1	0.63	0610011F06RIK	0.52	SLC6A19	0.43	SLC17A3	0.30
GPD1	0.62	SLC6A18	0.52	G6PC	0.43	FTH1	0.30
FMO2	0.62	ECI3	0.52	NUDT19	0.42	SLC22A18	0.30
GCDH	0.61	HYKK	0.52	SLC34A1	0.42	SLC7A13	0.28
MACROD2	0.61	SLC6A20B	0.51	SLC22A6	0.42	CYP2D26	0.28
AK4	0.60	CAT	0.51	CCDC107	0.41	CYP2E1	0.27
CNDP2	0.60	FUT9	0.51	SLC22A28	0.41	CYP2J5	0.23
PRODH2	0.59	INMT	0.51	ERRFI1	0.41	SLC22A8	0.23
PRODH	0.59	BDH2	0.50	DNAJC12	0.40	TNFAIP8	0.21
ASPDH	0.59	FAH	0.50	4833439L19RIK	0.40	SLC17A1	0.19
EPHX2	0.59	ACOX1	0.50	SLC22A1	0.40	KEG1	0.14
TRIM7	0.59	NOX4	0.50	TMEM150A	0.39	D630029K05RIK	0.12
XYLB	0.59	ASL	0.50	MPV17L	0.39		

Table 15: Uniprot ID (ID) and prediction value (P) of proteins predicted as redox-active using ET. Model was trained on protein set 3.

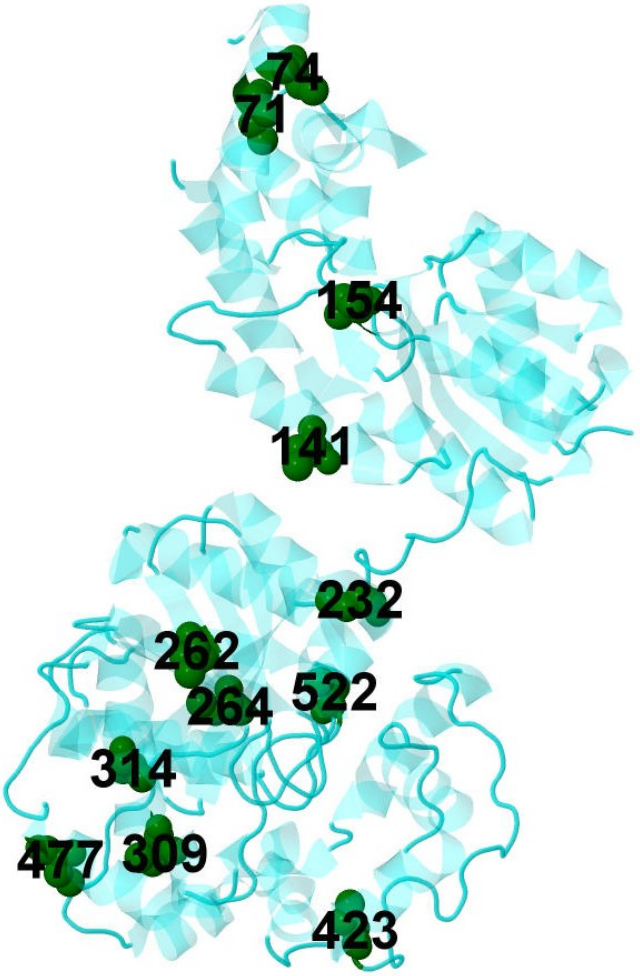

Figure 59: Structure of PDB entry 5AM2 [115] in cartoon representation. Cysteines shown in green. This structure was predicted to likely be redox-active.

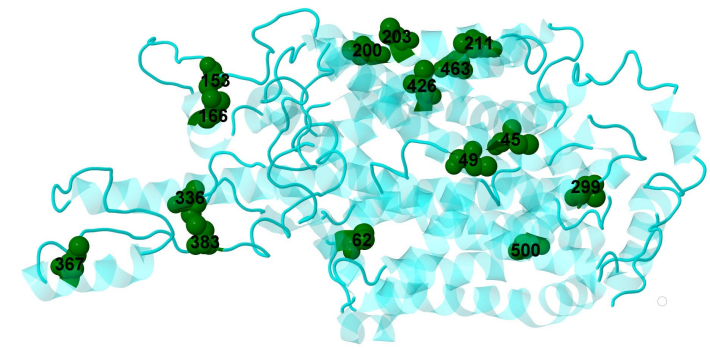

Figure 60: Structure of PDB entry 6M17 [111] in cartoon representation. Cysteines shown in green. This structure was predicted to likely not be redox-active.

4.3.4 Modifiable Cysteines in Redox Proteins

The ET algorithm was trained on protein set 2, using imputation to make novel predictions for cysteine sensitivity in a new protein set. These proteins were collected from dataset B of the RedoxDB and are suspected to contain redox-sensitive cysteines. They were not used for training as the sequence positions of the redox-sensitive cysteines are not known. The list included 130 proteins with 620 cysteines. I ran the predictions ten times for more robust results. Out of the 130 proteins, 62 included cysteines with a prediction score larger than 0.45, which I applied as a threshold for a cysteine likely being a Cys+. I 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 the dataset. It is likely that some of the investigated proteins in the dataset did not contain any redox-active cysteines in their PDB entries due to incomplete data. 15% of the cysteines were predicted as Cys+. For the full list of positively predicted cysteines, see Table 17 to Table 19. I compared the predictions with known data of post-translational modifications from the UniProt [25] database. I found that two of my predictions, CYS37 and CYS40 in 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 cysteines were known locations of redox activity according to UniProt. These findings reflect positively on the reliability of my

methods. Redox activity in disulfide bridges not known to be redox-active were predicted in several more proteins, such as CYS120 and CYS127 in PDB file 1B56 [53] for fatty acid-binding protein 5, where CYS120 was predicted as CYS+, as well as CYS105 and CYS137 in PDB file 3POW [22] 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 [44]) and the cytochrome b6-f complex iron-sulfur subunit (1RFS [17]).

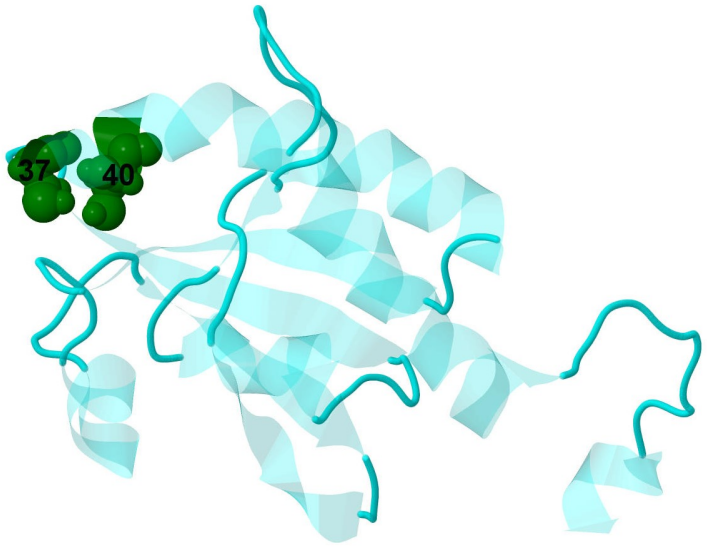

Figure 61: Structure of PDB entry 1X5D (available at https://www.rcsb.org/structure/1x5d) in cartoon representation. Both cysteines had a prediction value higher than 0.7 and are shown in green.

Table 16: !!

!! *NOTE:* [a]!!

Table 17: Cysteine prediction scores for RedoxDB proteins

Protein name	PDB/Chain ID	AA ID	Prediction[a]	Molecular function
Fructose-bisphosphate aldolase B	1QO5/A	289	0.469	Lyase
14-3-3 protein zeta/delta	2V7D/A	189	0.490	Protein domain specific binding
Ribulose bisphosphate carboxylase	1IR1/B	77	0.468	Photosynthesis
Nucleoside diphosphate-linked moiety	3H95/A	239	0.453	Hydrolase
Amine oxidase [flavin-containing] A	1O5W/A	321	0.675	Oxidoreductase
Amine oxidase [flavin-containing] A	1O5W/A	323	0.673	Oxidoreductase
Enoyl-CoA hydratase, mitochondrial	2HW5/A	111	0.473	Lyase
Multifunctional protein ADE2	2H31/A	151	0.490	Decarboxylase, Ligase, Lyase
Multifunctional protein ADE2	2H31/A	374	0.487	Decarboxylase, Ligase, Lyase
Serine/threonine-protein phosphatase	1B3U/A	316	0.554	Chromosome partition
Serine/threonine-protein phosphatase	1B3U/A	328	0.481	Chromosome partition
Serine/threonine-protein phosphatase	1B3U/A	389	0.686	Chromosome partition
Transportin-1	2Z5K/A	289	0.490	Host-virus interaction, transport
Sucrose synthase 1	3S28/A	349	0.456	Sucrose synthase activity
Sucrose synthase 1	3S28/A	434	0.493	Sucrose synthase activity
Inosine-5'-monophosphate dehydrogenase	1NF7/A	26	0.467	DNA-binding, oxidoreductase
Inosine-5'-monophosphate dehydrogenase	1NF7/A	331	0.500	DNA-binding, oxidoreductase
Inosine-5'-monophosphate dehydrogenase	1NF7/A	339	0.463	DNA-binding, oxidoreductase
Drebrin-like protein	1X67/A	104	0.465	Actin-binding
14-3-3 protein beta/alpha	4DNK/A	191	0.475	Host-virus interaction
Adenylosuccinate lyase	2J91/A	99	0.474	Lyase
Adenylosuccinate lyase	2J91/A	113	0.525	Lyase
Aflatoxin B1 aldehyde reductase member	1GVE/A	66	0.506	Oxidoreductase
Cysteine synthase 1	1Z7W/A	28	0.528	Cysteine synthase activity
Ribulose bisphosphate carboxylase	1GK8/B	83	0.468	Monooxygenase activity
Adenylosuccinate synthetase, chloroplastic	1DJ3/A	111	0.479	Ligase
Neuroligin-1	3B3Q/A	342	0.482	Biological rhythms, cell adhesion
D-3-phosphoglycerate dehydrogenase	2G76/A	233	0.517	electron transfer activity
Fructose-bisphosphate aldolase A	1ALD/A	177	0.461	Lyase
Fructose-bisphosphate aldolase A	1ALD/A	239	0.524	Lyase
Triosephosphate isomerase	2JK2/A	126	0.472	Isomerase, lyase

NOTE: [a]Only cysteines with scores greater than 0.45 are shown.

Table 18: Cysteine prediction scores for RedoxDB proteins

Protein name	PDB/Chain ID	AA ID	Prediction[a]	Molecular function
Triosephosphate isomerase	2JK2/A	217	0.459	Isomerase, lyase
Uroporphyrinogen decarboxylase	1J93/A	45	0.468	Decarboxylase, lyase
Alcohol dehydrogenase 1A	1U3T/A	174	0.463	Alcohol dehydrogenase activity
Peptidyl-prolyl cis-trans isomerase A	4DGD/A	52	0.661	Isomerase, rotamase
Peptidyl-prolyl cis-trans isomerase A	4DGD/A	62	0.562	Isomerase, rotamase
UTP–glucose-1-phos. uridylyltransferase	2ICY/A	95	0.515	Nucleotidyltransferase, transferase
L-lactate dehydrogenase B chain	1I0Z/A	163	0.680	L-lactate dehydrogenase activity
Homocysteine methyltransferase 1	1U1J/A	649	0.495	Methionine synthase activity
Homocysteine methyltransferase 1	1U1J/A	733	0.521	Methionine synthase activity
Nucleoside diphosphate kinase B	3BBB/A	145	0.512	Activator, DNA-binding, kinase
Protease Do-like 1, chloroplastic	3QO6/A	409	0.534	Hydrolase, serine protease
Adenylate kinase 2, mitochondrial	2C9Y/A	92	0.500	Adenylate kinase activity
Fatty acid-binding protein 5	1B56/A	120	0.494	Transport
Glyoxylate/succinic semiald. reductase	3DOJ/A	36	0.468	Glyoxylate reductase activity
Glyoxylate/succinic semiald. reductase	3DOJ/A	47	0.487	Glyoxylate reductase activity
Ran-specific GTPase-activating protein	1K5D/B	132	0.496	GTPase activator activity
Threonine synthase 1, chloroplastic	1E5X/A	455	0.469	Allosteric enzyme, lyase
Bifunctional epoxide hydrolase 2	3I28/A	81	0.472	Hydrolase, multifunctional enzyme
Bifunctional epoxide hydrolase 2	3I28/A	141	0.453	Hydrolase, multifunctional enzyme
Dihydropyrimidinase-related protein	2VM8/A	248	0.523	Developmental protein
Dihydropyrimidinase-related protein	2VM8/A	323	0.485	Developmental protein
Formate dehydrogenase	3JTM/A	81	0.453	Oxidoreductase
Formate dehydrogenase	3JTM/A	215	0.481	Oxidoreductase
Profilin-1	2PBD	16	0.464	Actin-binding
Ubiquitin-protein ligase E3A	1C4Z/D	137	0.461	Transferase
Beta-amylase	2XFR/A	93	0.474	Glycosidase, hydrolase
Non-specific lipid-transfer protein	1BWO/A	27	0.453	Transport
Non-specific lipid-transfer protein	1BWO/A	48	0.531	Transport
Non-specific lipid-transfer protein	1BWO/A	50	0.477	Transport
Catechol O-methyltransferase	3U81/A	69	0.599	Methyltransferase, transferase
Catechol O-methyltransferase	3U81/A	95	0.622	Methyltransferase, transferase

NOTE: [a]Only cysteines with scores greater than 0.45 are shown.

Table 19: Cysteine prediction scores for RedoxDB proteins

Protein name	PDB/Chain ID	AA ID	Prediction[a]	Molecular function
Catechol O-methyltransferase	3U81/A	191	0.516	Methyltransferase, transferase
Ribulose bisphosphate carboxylase	8RUC/A	459	0.499	Monooxygenase, oxidoreductase
Enoyl-CoA hydratase, mitochondrial	1MJ3/A	111	0.454	Lyase
Ubiquitin thioesterase otubain-like	4DHI/D	87	0.555	Hydrolase, protease, thiol protease
Ubiquitin carboxyl-terminal hydrolase 16	2I50/A	34	0.519	Activator, chromatin regulator
Glycine–tRNA ligase	2ZT5/A	157	0.496	Aminoacyl-tRNA synthetase
Glycine–tRNA ligase	2ZT5/A	177	0.476	Aminoacyl-tRNA synthetase
Glycine–tRNA ligase	2ZT5/A	390	0.453	Aminoacyl-tRNA synthetase
Thiamine thiazole synthase, chloroplastic	1RP0/A	172	0.524	Transferase
Thiamine thiazole synthase, chloroplastic	1RP0/A	187	0.500	Transferase
Eukaryotic translation initiation factor 5A-2	3HKS/A	80	0.492	Initiation factor
Mitotic checkpoint ser/thr-protein kinase	2WVI/A	132	0.466	Kinase, ser/thr-protein kinase
3-hydroxyacyl-CoA dehydrogenase type-2	2O23/A	91	0.527	Oxidoreductase
DNA-directed RNA polymerase	3PO3/S	188	0.484	Nucleotidyltransferase, transferase
DNA-directed RNA polymerase	3PO3/S	231	0.467	Nucleotidyltransferase, transferase
DNA-directed RNA polymerase	3PO3/S	271	0.558	Nucleotidyltransferase, transferase
DNA-directed RNA polymerase	3PO3/S	274	0.545	Nucleotidyltransferase, transferase
DNA-directed RNA polymerase	3PO3/S	302	0.522	Nucleotidyltransferase, transferase
Protein disulfide-isomerase A6	1X5D/A	37	0.794	Chaperone, isomerase
Protein disulfide-isomerase A6	1X5D/A	40	0.723	Chaperone, isomerase
Crk-like protein	2BZY/A	13	0.497	
Cysteine synthase, mitochondrial	4AEC/A	136	0.532	Transferase
Moesin	1EF1/A	117	0.563	Host-virus interaction
Phosphoglycerate kinase 1	2WZB/A	98	0.458	Kinase, transferase
Adenylate kinase	1AKY/A	82	0.514	Kinase, transferase
Calreticulin	3POW/A	137	0.473	Chaperone
Ezrin	1NI2/A	117	0.527	Cell shape
Serotransferrin	1RYO/A	39	0.458	Ion transport, iron transport
Serotransferrin	1RYO/A	118	0.455	Ion transport, iron transport
Serotransferrin	1RYO/A	179	0.516	Ion transport, iron transport

NOTE: [a]Only cysteines with scores greater than 0.45 are shown.

5 Discussion

Statistics

I applied statistical methods to characterize the close neighborhood of Cys+, both in the sequence and in Euclidean space. I found that the occurrence of several AAs were significantly enhanced or depleted. In the Euclidean neighborhood, small AAs like proline, glycine and the polar cysteine and serine were found more frequently. In the sequence neighborhood, the frequencies of alanine, glycine, isoleucine and lysine were increased. The aliphatic AA leucine was depleted in both datasets. Similar findings could be repeated by an investigation of the physicochemical properties of close AAs to Cys+, where I found a high enrichment of small and a depletion of aliphatic AAs in the Euclidean neighborhood of Cys+, while the sequence neighborhood showed an enhanced frequency of aromatic and positively charged AAs.

Through the use of sequence logos, I found that leucine is mainly depleted in the upstream sequence of Cys+. The relative scarcity of leucine may simply be explained by the fact that non-polar residues tend to be found more often on the inside of proteins, while polar residues occur more on the surface. Positively charged AAs like lysine and arginine are enriched more than four positions away from the central cysteine, confirming the results of Chen *et al.* [19], who found a higher abundance of positively charged residues around *S*-nitrosylation sites, which may aid in the regulation of redox modification. Polar residues like glycine and serine showed an enhanced frequency in the close Euclidean neighborhood of Cys+.

Significant differences were found in the makeup of SSEs near Cys+. In the Euclidean neighborhood, helices appeared at a lowered frequency, while disordered regions and bends were more common. In the sequence neighborhood, β-strands were found to be enriched. A high incidence of β-strands upstream and α-helices downstream from Cys+ was detected, confirming the findings of Fomenko *et al.* [41], who found that this property could be used to differentiate redox cysteines from metal-binding cysteines. They theorized that the downstream α-helix may be used to stabilize the reactive thiolate.

Taken together, the results concerning the frequency of AAs and SSEs confirm the position that the Euclidean neighborhood of Cys+ differs in unique and important ways from Cys− and can be treated as another class of features for machine learning methods, as it is not just redundant data when already considering sequence neighborhood.

I investigated the influence of accessibility, both of the cysteine and its neighboring AAs, on redox modifiability, comparing the two measures RSA and HSE. In both cases, I found that, while there was much overlap between Cys+ and Cys−, Cys+ its neighborhood tended to be significantly more accessible in the known PDB structures. It should be noted that these structures are static do not show the full range of shapes a protein may take *in vivo*. These results may indicate that interactions with other proteins or PTMs, which are known to often change the 3D structure of proteins, could have a significant regulatory effect on redox modifications. HSE provided slightly better differentiation than RSA, especially for the AA sequence neighborhood.

I predicted the pK_a of cysteines and other AAs by applying PROPKA 2.0 and PROPKA 3.4. I found significant differences in the predicted pK_a values of most AAs in the Euclidean neighborhood of cysteines between Cys+ and Cys−, except for glutamic acid and histidine. A low pK_a value leads to higher reactivity in thiols, facilitating the establishment of redox modifications. PROPKA 3.4 provided better differentiation between the datasets for most AAs. There was overlap in the range of values between Cys+ and Cys− for all AAs, confirming earlier studies showing that, while pK_a may play a large role in redox modifiability, many other factors appear to have a stronger predictive value [97], at least when using pK_a values predicted by computational tools based on static PDB structures.

I compared SSEs and RSA predicted using the PSIPRED and ASAquick tools to assignments by the DSSP algorithm in order to supplement structural data with sequence-based predictions without having to rely too heavily on traditional imputation methods. Both tools showed high accuracy, with PSIPRED agreeing with 80% of DSSP assignments, while ASAquick and DSSP had a Pearson correlation coefficient of 0.64, justifying their inclusion in my methods.

I calculated the CTD values for AAs and SSEs as well as the autocovariance scores of physicochemical properties for a protein set consisting of both Cys+ proteins from the RedoxDB and a random set of proteins from Uniprot. I was able to confirm that a large number of these values differed significantly enough between the two parts of the dataset to be useful features for the classification of proteins containing Cys+ according to the Bonferroni-corrected p-value and the ANOVA F-score. Among the most useful features were the distribution of SSEs, the composition of AAs and the distribution of cysteines.

Machine Learning

Utilizing the CTD values and autocovariance scores as features, I built and tested a model for the prediction of Cys+ containing proteins by applying the ET algorithm. The model achieved an AUC score of 0.75. I tested the model to see if the dataset biased it in terms of length or taxonomy, and found only a very small influence at most.

We applied the recently developed Geometricus tool to find shapemers in a dataset consisting of Cys+ containing proteins from the RedoxDB and a random set of proteins collected through PISCES from the PDB. We utilized the shapemers as features to build models by applying the RF and ET algorithms. RF achieved an AUC value of 0.78 and 0.81 using radius-based shapemers and sequence-based shapemers, respectively. ET achieved an AUC value of 0.78 and 0.82 using radius-based shapemers and sequence-based shapemers, respectively.

Both the approach based on CTD and autocovariance values and the approach based on shapemers produced reasonably succesful models and could be used to help researchers identify proteins of interest for further experimentation. It may prove useful in the future to combine both approaches for higher accuracy.

We implemented a profile HMM with AA data from the Euclidean and sequence environment of cysteines in the RedoxDB. We calculated probability matrices for each position and AA to build a model for the prediction of Cys+ in proteins. The models achieved an AUC of 0.69 for the Euclidean method and 0.72 for the sequence method.

Data from the AA environment, SSEs, RSA, HSE, pK_a and PTMs was

used together as features to build robust models by applying the SVM, ET, RF and GB algorithms, comparing their performance in the process. We utilized a Euclidean and a sequence method while enhancing the sequence dataset using imputation. AUC ranged from 0.66 to 0.72, with the most promising results being achieved using an imputed sequence dataset, showing that the payoff for using imputed and predicted data was worthwhile. The algorithms all showed similar performance on the test dataset.

I tested the SVM algorithm for the use of different amounts of AAs considered for the feature set as well as different amounts of cysteines in the training set. The tests were run using both a set of proteins taken randomly from the training set as well as data from mitochondrial complex I as a use case. I found that the number of AAs that was chosen showed the most favorable performance for both the Euclidean and the imputed sequence method among the tested values, although it may be possible to improve sequence-based models further by testing for a higher amount of AAs. The imputation method showed only marginally worse performance for lower amounts of training data, while the Euclidean method showed strong improvements for the use case, but not for the larger test set. It appears that a greater amount of training data will, on average, not result in much improvement for the performance of this method.

We calculated structural moment invariants of cysteines with the Geometricus tool to use them as a redox cysteine classification feature. either using a sequence-based or radius-based neighborhood. We produced models by applying the RF and ET algorithms to the data. The sequence-based method resulted in an AUC of 0.62 and 0.61, the radius-based method in an AUC of 0.58 and and 0.54, for RF and ET, respectively. Moment invariants and HMMs, especially using the sequence-based method, may prove to be a useful additional feature for redox models.

Use Cases

I applied the models I had developed earlier to several different use cases. After predicting the Cys+ for mitochondrial complex I, I compared the output to the results of earlier experimental research, achieving an AUC of 0.71 and 0.67 by utilizing using the Euclidean method with ET and the imputed sequence method with SVM, repectively. For human

mitochondrial complex I, the majority of Cys− and a large percentage of Cys+ were assigned correct predictions. Most notably, Cys39 of the ND3 subunit was predicted as redox modifiable by the ET algorithm, but not the SVM algorithm. This cysteine has been shown to become exposed under conditions of hypoxia, resulting in low complex I activity [21], and can inhibit complex I function. The static PDB structure underlying the predictions did not contain the protein structure under hypoxic conditions. It appears that, despite this, the algorithms were able to identify Cys39 as a possible point of interest, as they were trained on a dataset that may have also contained cysteines that similarly only become exposed under specific conditions.

I compared different variants of the NKGE2 natural killer cell receptor, differing only in two SNPs. One cysteine was close enough to the SNPs to be affected by the difference in residues in their local environment in the context of my methods. Applying the imputed sequence method with SVM, the difference between the prediction values between the variants was around 0.01 (for a value between 0 and 1), more than an order of magnitude larger than the usual inaccuracy due to the randomness of machine learning methods that were observed for the other cysteines. One single SNP appears to rarely cause a major difference in prediction values, unless the SNP causes a major reorganization regarding the structure of the protein. I found only one cysteine that was predicted to be redox modifiable in NKG2E, which was not close to the SNPs.

I made predictions for a dataset of proteins from proximal tubule cells. These predictions are currently being used by another group of researchers lead by Dr. Flávia Rezende to find the physiological function of the NADPH oxidase 4 in the kidney, a potential pharmacological target in kidney fibrosis. Their results may also be able to confirm the validity of my methods. I compared the predictions for this dataset using the imputed sequence method to the Euclidean method. Some proteins showed very different prediction values depending on the method used. After some investigation, I found that many of the greatest differences occurred for proteins with very incomplete PDB entries. It may be important in future studies to automatically or manually curate PDB datasets to contain only PDB entries displaying very high completeness.

I applied the imputation-based sequence method with the ET algorithm to make predictions on a dataset containing redox-active proteins from the RedoxDB, where the exact location of the Cys+ was not yet known. About 15% of the cysteines in the dataset were predicted as Cys+. I compared my predictions to annotations from the Uniprot database. I found only one cysteine which was annotated as containing a redox-active disulfide bridge in the dataset. Both of the cysteines participating in the disulfide bridge were predicted as Cys+ by my method.

Outlook

In future studies, we would like to add many of our newer features, like moment invariants and profile HMM values, to the general Cys+ prediction model to improve results. We would similarly like to combine the CTD and autocovariance method with shapemers as a feature for improved predictions of Cys+ containing proteins. We would like to explore additional features, such as AA conservation, subcellular location, taxonomy or redox stimuli.

We will perform more statistical analyses of new and existing features, such as shapemers, CTD values and autocovariance, to gain a better understanding of how exactly the environment of Cys+ differs from Cys− on a molecular level, identifying specific structural characteristics.

Applying the prediction of Cys+ in tandem with the prediction of Cys+ containing proteins may yield superior performance for both on datasets without any confirmed redox activity.

We would like to continue working together with experimental researchers to help identify redox-active targets, in turn testing the validity of the machine learning methods.

In this study, all types of redox modifications were treated as equivalent. As the amount of data grows, it will almost certainly be advantageous to categorize different redox modifications into different groups, developing models that can differentiate between them. Training data was collected from many different species. While this didn't appear to make a large difference in the quality of the predictions, it may still prove beneficial to separate data from different sources. Eventually, different datasets for a variety of redox stimuli as well as cellular compartments could be developed, shining further light on the very specific conditions that enable redox modification.

The machine learning and data analysis tools exist in the form of a Python script executable from the command line. In the future, this script could be

developed into a program with a graphical user interface for easier use by researchers without a programming background.

My results indicate that a machine learning approach may be a valuable tool for the prediction and analysis of redox-sensitive cysteines, while further research may be able to improve the robustness and performance of my models.

6 Acknowledgements

I would like to express my thanks to my research supervisors, Prof. Ina Koch and Dr. Ilka Wittig, for their continued support and guidance and for giving me the opportunity to undertake this research project. Many thanks must also go to Dr. Jörg Ackermann, who contributed many ideas, valuable and constructive criticism, and who was always ready to lend a helping hand. Furthermore, I would like to thank Prof. Ralf Brandes for taking the time to review this dissertation.

I would also like to thank Brigitte Scheidemantel-Geiß, who showed great patience when having to deal with any unnecessary administrative complications caused by my own forgetfulness, as well as the entire MolBI group, the Institute of Informatics, and my fellow PhD students. I have made many friends during my studies and spent many enjoyable hours, creating memories that will last for a lifetime.

My thanks also go to Dr. Flávia Rezende, who provided important data for me to test my models on, and who was always a joy to work with. Further thanks go to my Masters student Nils Nover for the opportunity to supervise his thesis, for his work, his motivation, and for being a great guy.

Furthermore, I would like to say thanks to my parents, for their continued support during my studies.

Last but not least, I must extend my thanks to my group of friends, because my friends are the friendliest friends, and I love all of you.

Funding: This study was supported by the Deutsche Forschungsgemeinschaft: SFB 815/Z1 (I. W.).

7 Bibliography

References

[1] G. Abrusán and J. A. Marsh. Alpha helices are more robust to mutations than beta strands. *PLoS Computational Biology*, 12(12):1–16, 2016.

[2] A. A. Agip, J. N. Blaza, H. R. Bridges, C. Viscomi, S. Rawson, S. P. Muench, and J. Hirst. Cryo-EM structures of complex I from mouse heart mitochondria in two biochemically defined states. *Nature Structural & Molecular Biology*, 25:548–556, 2018.

[3] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[4] B. Andi, H. Xu, P. F. Cook, and A.H. West. Crystal Structures of Ligand-Bound Saccharopine Dehydrogenase from Saccharomyces cerevisiae. *Biochemistry*, 46:12512–12521, 2007.

[5] F. Åslund, K. D. Berndt, and A. Holmgren. Redox Potentials of Glutaredoxins and Other Thiol-Disulfide Oxidoreductases of the Thioredoxin Superfamily Determined by Direct Protein-Protein Redox Equilibria. *Journal of Biological Chemistry*, 272(49):30780–30786, 1997.

[6] M. A. Baker, A. Weinberg, L. Hetherington, A. I. Villaverde, and T. Velkov. Analysis of protein thiol changes occurring during rat sperm epididymal maturation. *Biology of reproduction*, 92(1), 11:1–10, 2015.

[7] L. E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[9] L. Bleier, I. Wittig, H. Heide, M. Steger, U. Brandt, and S. Dröse. Generator-specific targets of mitochondrial reactive oxygen species. *Free Radical Biology and Medicine*, 78:1–10, 2015.

[10] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, 4(6):1633–1649, 2004.

[11] F. Bock, L.A. Blaga, and B. Klusemann. Mechanical Performance Prediction for Friction Riveting Joints of Dissimilar Materials via Machine Learning. *Procedia Manufacturing*, 47:615–622, 2020.

[12] N. Brandes, S. Schmitt, and U. Jakob. Thiol-Based Redox Switches in Eukaryotic Proteins. *Antioxidants & Redox Signaling*, 11(5):997–1014, 2009.

[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[14] D. Buchan and D. T. Jones. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research*, 47(W1):W402–W407, 2019.

[15] D. P. Byrne, S. Shrestha, M. Galler, M. Cao, L. A. Daly, A. E. Campbell, C. E. Eyers, E. A. Veal, N. Kannan, and P. A. Eyers. Aurora A regulation by reversible cysteine oxidation reveals evolutionarily conserved redox control of Ser/Thr protein kinase activity. *Science Signaling*, 13(639), 2020.

[16] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Computational Biology*, 5(12):1–18, 2009.

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

[18] L. Chen. towards data science basic ensemble learning (random forest, adaboost, gradient boosting)- step by step explained. https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725. Accessed: 2021-10-03.

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

[20] E. T. Chouchani, T. R. Hurd, S. M. Nadtochiy, P. S. Brookes, Fearnley I. M., K. S. Lilley, R. A. J. Smith, and M. P. Murphy. 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. *Biochemical Journal*, 430(1):49–59, 2010.

[21] E. T. Chouchani, C. Methner, S. M. Nadtochiy, A. Logan, V. R. Pell, S. Ding, A. M. James, H. M. Cochemé, J. Reinhold, K. S. Lilley, L. Partridge, I. M. Fearnley, A. J. Robinson, R. C. Hartley, R. A. J. Smith, T. Krieg, P. S. Brookes, and M. P. Murphy. Cardioprotection by S-nitrosation of a cysteine switch on mitochondrial complex I. *Nature Medicine*, 19:753–759, 2013.

[22] A. Chouquet, H. Paidassi, W. L. Ling, P. Frachet, G. Houen, G. J. Arlaud, and C. Gaboriaud. X-ray structure of the human calreticulin globular domain reveals a Peptide-binding area and suggests a multimolecular mechanism. *PLoS One*, 6:e17886–e17886, 2011.

[23] H. S. Chung, S. B. Wang, V. Venkatraman, C. I. Murray, and J. E. Van Eyk. Cysteine Oxidative Posttranslational Modifications - Emerging Regulation in the Cardiovascular System. *Circulation Research*, 112(2):382–392, 2013.

[24] Y. Collins, E. T. Chouchani, A. M. James, K. E. Menger, H. M. Cochemé, and M. P. Murphy. Mitochondrial redox signalling at a glance. *Journal of Cell Science*, 125(4):801–806, 2012.

[25] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47:D506–D515, 2019.

[26] G. Coqueret and T. Guida. *Machine Learning for Factor Investing: R Version*. Chapman and Hall/CRC, 2020.

[27] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[28] G. E. Crooks, G. Hon, J. Chandonia, and S. E. Steven E. Brenner. WebLogo: A sequence logo generator. *Genome Research*, 14:1188–1190, 2004.

[29] R. Dagil, C. O'Shea, A. Nykjar, A. M. Bonvin, and B. B. Kragelund. Gentamicin binds to the megalin receptor as a competitive inhibitor using the common ligand binding motif of complement type repeats: insight from the nmr structure of the 10th complement type repeat domain alone and in complex with gentamicin. *Journal of Biological Chemistry*, 288:4424–4435, 2013.

[30] T. Dansen, L. Smits, M. van Triest, P. L. J. de Keizer, D. van Leenen, M. G. Koerkamp, A. Szypowska, A. Meppelink, A. B. Brenkman, J. Yodoi, F. C. P. Holstege, and Burgering B. M. T. Redox-sensitive cysteines bridge p300/cbp-mediated acetylation and foxo4 activity. *Nature Chemical Biology*, 5:664–672, 2009.

[31] S. A. Dogan, R. Cerutti, C. Benincá, G. Brea-Calvo, H. T. Jacobs, M. Zeviani, M. Szibor, and C. Viscomi. Perturbed redox signaling exacerbates a mitochondrial myopathy. *Cell Metabolism*, 28(5):764–775, 2018.

[32] S. Dröse, U. Brandt, and I. Wittig. Mitochondrial respiratory chain complexes as sources and targets of thiol-based redox-regulation. *Biochimica et Biophysica Acta*, 1844(8):1344–1354, 2014.

[33] J. Durairaj, M. Akdel, D. de Ridder, and A. D. J. van Dijk. Geometricus represents protein structures as shape-mers derived from moment invariants. *bioRxiv*, 2020.

[34] S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[35] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[36] E. Faraggi, Y. Zhou, and A. Kloczkowski. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins*, 82(11):3170–3176, 2014.

[37] M. Ferez, C. J. Knudson, A. Lev, E. B. Wong, P. Alves-Peixoto, L. Tang, C. Stotesbury, and L. J. Sigal. Viral infection modulates Qa-1^b in infected and bystander cells to properly direct NK cell killing. *Journal of Experimental Medicine*, 218(5):1–13, 2021.

[38] K. Fiedorczuk, J. A. Letts, G. Degliesposti, K. Kaszuba, M. Skehel, and L. A. Sazanov. Atomic structure of the entire mammalian mitochondrial complex I. *Nature*, 538(7625):406–410, 2016.

[39] J. Flusser, J. Boldys, and B. Zitova. Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):234–246, 2003.

[40] D. E. Fomenko, S. M. Marino, and V. N. Gladyshev. Functional diversity of cysteine residues in proteins and unique features of catalytic redox-active cysteines in thiol oxidoreductases. *Molecules and Cells*, 26(3):228–235, 2008.

[41] D. E. Fomenko, W. Xing, B. M. Adair, D. J. Thomas, and V. N. Gladyshev. High-Throughput Identification of Catalytic Redox-Active Cysteine Residues. *Science*, 315(5810):387–389, 2007.

[42] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 1999.

[43] M. Y. Galperin, A. N. Nikolskaya, and E. V. Koonin. Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiology Letters*, 203:11–21, 2001.

[44] P. S. Gandhi, M. J. Page, Z. Chen, L. Bush-Pelc, and E. Di Cera. Mechanism of the Anticoagulant Activity of Thrombin Mutant W215A/E217A. *Journal of Biological Chemistry*, 284:24098–24105, 2009.

[45] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[46] D. Gough and T. Cotter. Hydrogen peroxide: a jekyll and hyde signalling molecule. *Cell Death & Disease*, 2:1–8, 2011.

[47] T. M. Greco, R. Hodara, I. Parastatidis, H. F. G. Heijnen, M. K. Dennehy, D. C. Liebler, and H. Ischiropoulos. Identification of S-nitrosylation motifs by site-specific mapping of the S-nitrosocysteine proteome in human vascular smooth muscle cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(19):7420–7425, 2006.

[48] B. Groitl and U. Jakob. Thiol-based redox switches. *Biochimica et Biophysica Acta*, 1844:1335–1343, 2014.

[49] R. Guo, S. Zong, M. Wu, J. Gu, and M. Yang. Architecture of Human Mitochondrial Respiratory Megacomplex $I_2III_2IV_2$. *Cell*, 170(6):1247–1257, 2017.

[50] V. Gupta and K. S. Carroll. Sulfenic acid chemistry, detection and cellular lifetime. *Biochimica et biophysica acta*, 1840(2):847–875, 2014.

[51] D. T. Hess, A. Matsumoto, S. Kim, H. E. Marshall, and J. S. Stamler. Protein S-nitrosylation: purview and parameters. *Nature Reviews Molecular Cell Biology*, 6:150–166, 2005.

[52] T. K. Ho. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, page 278–282, 1995.

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

[54] A. Holmgren. Thioredoxin and glutaredoxin systems. *Journal of Biological Chemistry*, 264(24):13963–13966, 1989.

[55] T. R. Hurd, T. A. Prime, M. E. Harbour, K. S. Lilley, and M. P. Murphy. Detection of reactive oxygen species-sensitive thiol proteins by redox difference gel electrophoresis: implications for mitochondrial redox signaling. *Journal of Biological Chemistry*, 282(30):22040–22051, 2007.

[56] M. Ikemoto, T. Nikawa, M. Kano, K. Hirasaka, T. Kitano, C. Watanabe, R. Tanaka, T. Yamamoto, M. Kamada, and K. Kishi. Cysteine supplementation prevents unweighting-induced ubiquitination in association with redox regulation in rat skeletal muscle. *Biological Chemistry*, 383(3-4), 2005.

[57] K. Jakubczyk, K. Dec, J. Kałduńska, D. Kawczuga, J. Kochman, and K. Janda. Reactive oxygen species - sources, functions, oxidative damage. *Polski Merkuriusz Lekarski*, 48(284):124–127, 2020.

[58] J. L. Jenkins and J. J. Tanner. High-resolution structure of human D-glyceraldehyde-3-phosphate dehydrogenase. *Acta Crystallographica Section D: Structural Biology*, 62:290–301, 2006.

[59] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

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

[61] B. K. Kaiser, J. C. Pizarro, J. Kerns, and R. K. Strong. Structural basis for NKG2A/CD94 recognition of HLA-E. *Proceedings of the National Academy of Sciences of the United States of America*, 105(18):6696–6701, 2008.

[62] C. Klomsiri, P. A. Karplus, and L. B. Poole. Cysteine-based redox switches in enzymes. *Antioxidants & Redox Signaling*, 14(6):1065–1077, 2011.

[63] T. Kortemme and T. E. Creighton. Ionisation of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pK_a values in proteins of the thioredoxin family. *Journal of Molecular Biology*, 253(5):799–812, 1995.

[64] E. Kowalinski, G. Bange, B. Bradatsch, E. Hurt, K. Wild, and I. Sinning. The crystal structure of Ebp1 reveals a methionine aminopeptidase fold as binding platform for multiple interactions. *FEBS Letters*, 581:4450–4454, 2007.

[65] A. Kumar, I. S. Yadav, S. Hussain, B. C. Das, and M. Bharadwaj. Identification of immunotherapeutic epitope of E5 protein of human papillomavirus-16: An in silico approach. *Biologicals*, 43(5):344–348, 2015.

[66] V. Larosa and C. Remacle. Insights into the respiratory chain and oxidative stress. *Bioscience Reports*, 38:1–14, 2018.

[67] L. I. Leichert, F. Gehrke, H. V. Gudiseva, T. Blackwell, M. Ilbert, A. K. Walker, J. R. Strahler, P. C. Andrews, and U. Jakob. Quantifying changes in the thiol redox proteome upon oxidative stress in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 105(24):8197–8202, 2008.

[68] C. Lin, K. Lin, C. Yang, I. Chung, C. Huang, and Y. Yang. Protein Metal Binding Residue Prediction Based on Neural Networks. *International Journal of Neural Systems*, 15(1-2):71–84, 2011.

[69] F. Magnani and A. Mattevi. Structure and mechanisms of ros generation by nadph oxidases. *Current Opinion in Structural Biology*, 59:91–97, 2019.

[70] R. J. Mailloux, X. Jin, and W. G. Willmore. Redox regulation of mitochondrial function with emphasis on cysteine oxidation reactions. *Redox biology*, 2:123–139, 2013.

[71] A. G. Mamistvalov. N-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):819–831, 1998.

[72] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[73] S. M. Marino and V. N. Gladyshev. 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 Computational Biology*, 5(5):1–13, 2009.

[74] S. M. Marino and V. N. Gladyshev. Structural Analysis of Cysteine S-Nitrosylation: A Modified Acid-Based Motif and the Emerging Role of Trans-Nitrosylation. *Journal of Molecular Biology*, 395(4):844–859, 2010.

[75] S. M. Marino and V. N. Gladyshev. Analysis and Functional Prediction of Reactive Cysteine Residues. *Journal of Biological Chemistry*, 287:4419, 2012.

[76] P. Martìnez-Acedo, E. Núñez, F. J. Gómez, M. Moreno, E. Ramos, A. Izquierdo-Álvarez, E. Miró-Casas, R. Mesa, P. Rodriguez, A. Martínez-Ruiz, D. G. Dorado, S. Lamas, and J. Vázquez. A Novel Strategy for Global Analysis of the Dynamic Thiol Redox Proteome. *Molecular & Cellular Proteomics*, 11(9):800–813, 2015.

[77] A. Mata-Cabana, M. García-Domínguez, F. J. Florencio, and M. Lindahl. Thiol-based redox modulation of a cyanobacterial eukaryotic-type serine/threonine kinase required for oxidative stress tolerance. *Antioxidants & Redox Signaling*, 17(4):521–533, 2012.

[78] Z. M. Moghadam, P. Henneke, and J. Kolter. From Flies to Men: ROS and the NADPH Oxidase in Phagocytes. *Frontiers in Cell and Developmental Biology*, 9:1–16, 2021.

[79] E. Moutevelis and J. Warwicker. Prediction of pk_a and redox properties in the thioredoxin superfamily. *Protein Science*, 13(10):2744–2752, 2004.

[80] M. P. Murphy. How mitochondria produce reactive oxygen species. *Biochemical journal*, 417(1):1–13, 2009.

[81] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, 2011.

[82] J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, and K. Suszták. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, 360(6390):758–763, 2018.

[83] A. Passerini and P. Frasconi. Learning to discriminate between ligand-bound and disulfide-bound cysteines. *Protein Engineering Design & Selection*, 17(4):367–373, 2004.

[84] A. Passerini, M. Punta, A. Ceroni, B. Rost, and P. Frasconi. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins*, 65(2):305–316, 2006.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] V. R. Pell, A. Spiroski, J. Mulvey, N. Burger, A. S. H. Costa, A. Logan, A. V. Gruszczyk, T. Rosa, A. M. James, C. Frezza, M. P. Murphy, and T. Krieg. Ischemic preconditioning protects against cardiac ischemia reperfusion injury without affecting succinate accumulation or oxidation. *Journal of Molecular and Cellular Cardiology*, 123:88–91, 2018.

[87] B. Poljsak, D. Šuput, and I. Milisav. Achieving the Balance between ROS and Antioxidants: When to Use the Synthetic Antioxidants. *Oxidative Medicine and Cellular Longevity*, 2013:1–11, 2013.

[88] P. D. Ray, B. Huang, and Y. Tsuji. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cellular Signalling*, 24:981–990, 2012.

[89] R. Requejo, E. T. Chouchani, A. M. James, T. A. Prime, K. S. Lilley, I. M. Fearnley, and M. P. Murphy. Quantification and identification of mitochondrial proteins containing vicinal dithiols. *Archives of Biochemistry and Biophysics*, 504(2):228–235, 2010.

[90] D. J. Rigden, M. J. Jedrzejas, and M. Y. Galperin. An extracellular calcium-binding domain in bacteria with a distant relationship to efhands. *FEMS Microbiology Letters*, 221:103–110, 2003.

[91] G. Roos, N. Foloppe, and J. Messens. Understanding the pK(a) of Redox Cysteines: The Key Role of Hydrogen Bonding. *Antioxidants & Redox Signaling*, 18:94–127, 2012.

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

[93] Jr Salsbury, F. R., S. T. Knutson, L. B. Poole, and J. S. Fetrow. Functional site profiling and electrostatic analysis of cysteines modifiable to cysteine sulfenic acid. *Protein science : a publication of the Protein Society*, 17(2):299–312, 2008.

[94] R. Sanchez, M. Riddle, J. Woo, and J. Momand. Prediction of reversibly oxidized protein cysteine thiols using protein structure properties. *Protein science : a publication of the Protein Society*, 17(3):473–481, 2008.

[95] B. K. Sarma and G. Mugesh. Redox Regulation of Protein Tyrosine Phosphatase 1B (PTP1B): A Biomimetic Study on the Unexpected Formation of a Sulfenyl Amide Intermediate. *Journal of the American Chemical Society*, 129(28):8872–8881, 2007.

[96] J. Song, H. Tan, K. Takemoto, and T. Akutsu. HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*, 24(13):1489–1497, 2008.

[97] C. M. Spickett and A. R. Pitt. Protein oxidation: role in signalling and detection by mass spectrometry. *Amino Acids*, 42:5–21, 2012.

[98] L.B. Sullivan and N. S. Chandel. Mitochondrial reactive oxygen species and cancer. *Cancer & Metabolism*, 2(17):1–12, 2014.

[99] M. Sun, Y. Wang, H. Cheng, Q. Zhang, W. Ge, and D. Guo. RedoxDB - a curated database of experimentally verified protein redox modification. *Bioinformatics*, 28(19):2551–2552, 2012.

[100] M. Sun, Q. Zhang, Y. Wang, W. Ge, and D. Guo. Prediction of redox-sensitive cysteines using sequential distance and other sequence-based features. *BMC Bioinformatics*, 17(316):1–10, 2016.

[101] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, and J. H. Jensen. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, 2011.

[102] J. J. Tanner, Z. D. Parsons, A. H. Cummings, H. Zhou, and K. S. Gates. Redox Regulation of Protein Tyrosine Phosphatases: Structural and Chemical Aspects. *Antioxidants & Redox Signaling*, 15(1):77–97, 2011.

[103] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, , and C. O. Wilke. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One*, 8(11):1–8, 2013.

[104] J. F. Turrens. Mitochondrial formation of reactive oxygen species. *The Journal of Physiology*, 552(2):335–344, 2003.

[105] J. van der Reest, S. Lilla, L. Zheng, S. Zanivan, and E. Gottlieb. Proteome-wide analysis of cysteine oxidation reveals metabolic sensitivity to redox stress. *Nature Communications*, 9(1581):1–16, 2018.

[106] G. Wang and R. L. R. L. Dunbrack Jr. Pisces: a protein sequence culling server. *Bioinformatics*, 19:1589–1591, 2003.

[107] A. S. Wibowo, M. Singh, K. M. Reeder, J. J. Carter, A. R. Kovach, W. Meng, M. Ratnam, F. Zhang, and C. E. Dann. Structures of human folate receptors reveal biological trafficking states and diversity in folate and antifolate recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 110:15180–15188, 2013.

[108] C. C. Winterbourn and M. B. Hampton. Thiol chemistry and specificity in redox signaling. *Free radical biology & medicine*, 45(5):549–561, 2008.

[109] C. Wirth, U. Brandt, C. Hunte, and V. Zickermann. Structure and function of mitochondrial complex I. *Biochimica et Biophysica Acta*, 1857:902–914, 2016.

[110] M. A. Wouters, S. W. Fan, and N. L. Haworth. Disulfides as redox switches: From molecular mechanisms to functional significance. *Antioxidants & Redox Signaling*, 12(1):53–91, 2010.

[111] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou. Structural basis for the recognition of sars-cov-2 by full-length human ace2. *Science*, 367(6485):1444–1448, 2020.

[112] Z. H. You, Y. K. Lei, L. Zhu, J. Xia, and B. Wang. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics*, 14(8):S10, 2013.

[113] J. Yu and C. Z. Zhou. Crystal structure of glutathione reductase Glr1 from the yeast Saccharomyces cerevisiae. *Proteins*, 68:972–979, 2007.

[114] J. Zhu, K. R. Vinothkumar, and J. Hirst. Structure of mammalian respiratory complex I. *Nature*, 536(7616):354–358, 2016.

[115] L. Öster, S. Tapani, Y. Xue, and H. Käck. Successful generation of structural information for fragment-based drug discovery. *Drug Discovery Today*, 20(9):1104–1111, 2015.

8 Curriculum Vitae

Personal Data

Full Name:	Marcus Dominik Keßler
Date of Birth:	March 15, 1988
Place of Birth:	Heidelberg
Nationality:	Germany
Address:	Heisterstraße 34
	60594 Frankfurt am Main
Phone Number:	0179/9815899
E-Mail:	damarc@gmx.net
Marital Status:	unmarried

Education

School:	
1994-1998	Waldparkschule Heidelberg
1998-2007	Helmholtz Gymnasium Heidelberg
School Diploma:	2007 Abitur (GPA 2.0)
University:	
2008-2010	Hochschule Mannheim Bachelor Biotechnologie
2010-2013	Goethe-Universität Frankfurt Bachelor Bioinformatik
2013-2017	Goethe-Universität Frankfurt Master Bioinformatik
University Diploma:	
2013	Bachelor Bioinformatics (GPA 2.2)
Bachelor-Thesis:	Bioinformatics investigations of protein complexes using graph-theoretical methods(Grade: 1.3)
2017	Master Bioinformatik (GPA 1.5)
Master-Thesis:	Development of a Plugin for NeuroBox for Modeling and Simulation of Calcium Dynamics in Neuronal Spines(Grade:1.3)