Article

TRANSPARENT PROCESS

OPEN ACCESS

THE EMBO JOURNAL

Protein structure, amino acid composition and sequence determine proteome vulnerability to oxidation-induced damage

Roger L Chang[1,2,*], Julian A Stanley[1], Matthew C Robinson[1], Joel W Sher[1], Zhanwen Li[3], Yujia A Chan[1,2], Ashton R Omdahl[1], Ruddy Wattiez[4], Adam Godzik[3] & Sabine Matallana-Surget[5,**]

Abstract

Oxidative stress alters cell viability, from microorganism irradiation sensitivity to human aging and neurodegeneration. Deleterious effects of protein carbonylation by reactive oxygen species (ROS) make understanding molecular properties determining ROS susceptibility essential. The radiation-resistant bacterium *Deinococcus radiodurans* accumulates less carbonylation than sensitive organisms, making it a key model for deciphering properties governing oxidative stress resistance. We integrated shotgun redox proteomics, structural systems biology, and machine learning to resolve properties determining protein damage by γ-irradiation in *Escherichia coli* and *D. radiodurans* at multiple scales. Local accessibility, charge, and lysine enrichment accurately predict ROS susceptibility. Lysine, methionine, and cysteine usage also contribute to ROS resistance of the *D. radiodurans* proteome. Our model predicts proteome maintenance machinery, and proteins protecting against ROS are more resistant in *D. radiodurans*. Our findings substantiate that protein-intrinsic protection impacts oxidative stress resistance, identifying causal molecular properties.

Keywords *Deinococcus radiodurans*; oxidative stress; protein carbonyl; radioresistance; structural systems biology

Subject Categories Computational Biology; Post-translational Modifications & Proteolysis; Proteomics

DOI 10.15252/embj.2020104523 | Received 9 March 2020 | Revised 16 September 2020 | Accepted 22 September 2020 | Published online 19 October 2020

The EMBO Journal (2020) 39: e104523

Introduction

Proteome oxidation caused by reactive oxygen species (ROS) is a primary determinant of cellular sensitivity to desiccation and irradiation (Daly *et al*, 2007; Krisko & Radman, 2010) and is involved in the progression of age-related human diseases (Krisko & Radman, 2019), including neurodegeneration and cancer (Hohn *et al*, 2017). ROS toxicity is a common antibiotic mechanism (Belenky *et al*, 2015) and presents challenges in biotechnology including metabolic engineering (Ruenwai *et al*, 2011; Chin *et al*, 2017; Sun *et al*, 2018) and synthetic systems involving the high expression of fluorescent proteins (Ganini *et al*, 2017).

Prior to the previous decade, the dogma surrounding biological sensitivity to ionizing radiation focused primarily on DNA damage, but this changed as key experiments substantiated the role of protection from protein oxidation in the extreme radioresistance of the bacterium *Deinococcus radiodurans* (Daly, 2006). *Deinococcus* is a crucial model for investigating resistance to ROS because of its notorious tolerance of extreme oxidative stress, even prolonged cosmic doses of γ-radiation (Yamagishi *et al*, 2018). This tolerance stems from the evolution of *D. radiodurans* to tolerate desiccation, which also induces oxidative stress (Slade & Radman, 2011). *D. radiodurans* accumulates less protein oxidation than more sensitive species such as *Escherichia coli* (Krisko & Radman, 2010). Resistance in *D. radiodurans* is due partly to highly active ROS-detoxifying systems providing protein-extrinsic protection against ROS (i.e., not a property of the oxidation targets themselves; Daly *et al*, 2004, 2007).

Foundational work hypothesized that differential rates of protein oxidation and subsequent degradation also play a key role in stress response phenotypes (Stadtman, 1986) and broadly established that bacteria exhibit protein-specific patterns of susceptibility to oxidation under oxidative conditions leading to cellular senescence (Dukan & Nystrom, 1998). More recently, it was observed that pathogenic bacteria, which have evolved mechanisms to combat host immune responses that utilize ROS, are less sensitive to protein oxidation than non-pathogenic species and that certain physicochemical properties broadly differentiate the proteomes of pathogenic versus non-pathogenic bacteria, hypothesizing a causal link to

[1] Department of Systems Biology, Blavatnik Institute at Harvard Medical School, Boston, MA, USA
[2] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA
[3] Division of Biomedical Sciences, University of California Riverside School of Medicine, Riverside, CA, USA
[4] Department of Proteomics and Microbiology, Research Institute for Biosciences, University of Mons, Mons, Belgium
[5] Division of Biological and Environmental Sciences, Faculty of Natural Sciences, University of Stirling, Stirling, UK
*Corresponding author. E-mail: roger_chang@hms.harvard.edu
**Corresponding author. E-mail: sabine.matallanasurget@stir.ac.uk

susceptibility to protein oxidation (Vidovic *et al*, 2014). However, the extent to which protein-intrinsic properties (i.e., specific to individual protein species) contribute to ROS resistance and how such properties are distributed across distinct protein species has not been well-established. This comparative study of *D. radiodurans* and *E. coli* proteomes reveals proteins with distinguished vulnerability to ROS, thereby discovering mechanisms that contribute to the survival of oxidative stress following irradiation.

Reactive oxygen species damage proteins by the oxidation of side chains and backbones generally resulting in loss of function due to misfolding, aggregation, and proteolysis. Several types of protein oxidation can result upon reaction with ROS (Stadtman & Levine, 2003). In this study, we have focused exclusively on protein carbonylation, which has also been the focus of most experimental methods and foundational work on protein oxidation to date. Protein carbonyl sites (CS) on arginine, lysine, proline, and threonine (RKPT) sidechains (Appendix Fig S1) are seen as the most severe oxidative damage due to their irreversibility and frequency of occurrence. Furthermore, these carbonyls themselves are also highly reactive leading subsequently to additional damaging downstream reactions, such as non-enzymatic backbone cleavage via the proline oxidation pathway (Uchida *et al*, 1990; Cabiscol *et al*, 2000; Nystrom, 2005). In this way, RKPT carbonylation can be thought of as a committed step initiating a cascade of protein damage. Site-specific susceptibility to carbonylation differs across amino acid types and structural location, extending to the whole-molecule scale to distinguish ROS vulnerability across protein species (Fig 1A). However, specific molecular properties responsible for this vulnerability remain poorly understood.

Previous work provided evidence that there is a difference in carbonylation susceptibility between distinct protein species in bacteria through observation of banding patterns on carbonyl assay gels (Daly, 2009), but this work did not provide protein identification, quantification, nor residue specificity of carbonylation events. Identification of proteins prone to carbonylation and their specific sites is vital to understanding the molecular manifestation of deleterious oxidative stress phenotypes. This goal has motivated the development of mass spectrometry for direct proteome-wide CS identification and concomitant relative abundance changes, termed shotgun redox proteomics (Matallana-Surget *et al*, 2013). However, these experiments provide limited coverage of modified sites, a common problem in proteomics of post-translational modifications. Chemical derivatization during these experiments helps to stabilize the inherently transient, highly reactive protein carbonyls to promote their detection, but interference from derivatized adducts with proteolytic sites can also limit CS sampling capabilities. Computational methods for CS prediction are intended to learn shared features across modified sites in redox proteomic datasets and generalize to unknown sites on other proteins. Existing methods (Maisonneuve *et al*, 2009; Lv *et al*, 2014; Weng *et al*, 2017) are not ideal because they rely on linear sequence motifs and local homology; such a correlative basis for predicting structure–function relationships can require a very large number of example sequences before very strong predictors can be trained (Kamisetty *et al*, 2013), which are not yet available in the context of redox proteomic data. Furthermore, the exclusion of molecular structure features beyond simple sequence motifs provides limited understanding of causal mechanisms for protein carbonylation.

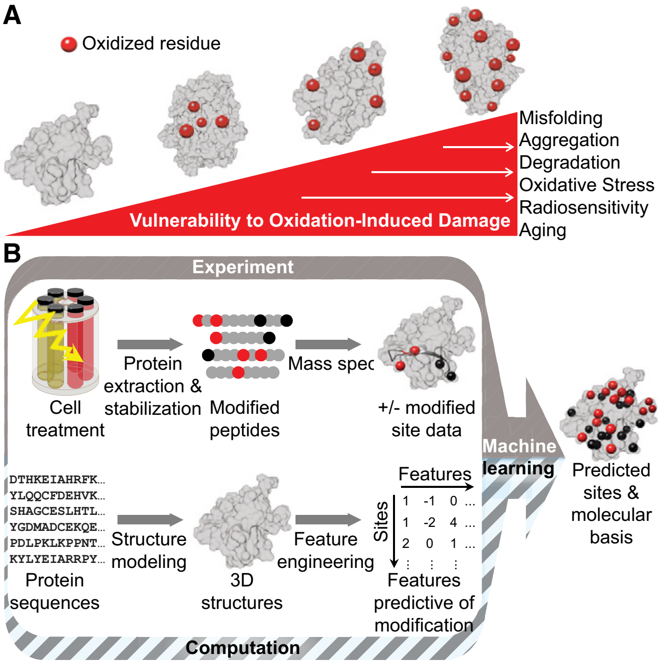

Figure 1. Study concept and workflow.

A Relationship between carbonylation site distribution, protein vulnerability to reactive oxygen species, and stress phenotypes.

B Structural systems biology workflow for proteome-wide carbonyl site prediction. Red circles = carbonyl sites (CS); black circles = non-oxidized RKPT residues; gray protein regions = non-RKPT residues.

In response to the limitations of conventional techniques, the field of structural systems biology offers approaches based on protein 3D molecular properties to investigate multi-scale proteomic questions, including mechanisms of physicochemical stress (Chang *et al*, 2013a,b). These approaches are empowered by the expansion of experimentally determined protein structures and advances in protein fold prediction (Yang *et al*, 2015). Our robust experimental design combined for the first time redox proteomics performed on cells exposed to an acute dose of γ-radiation with structural systems biology and machine learning (Fig 1B), generating a predictive model for protein carbonylation. This interdisciplinary workflow enabled proteome-wide characterization of susceptibility to carbonylation in *E. coli* and *D. radiodurans*, identifying phenotypically important protein targets, providing molecular explanations for target susceptibility, and supporting the role of protein-intrinsic properties in the survival of extreme oxidative stress.

Results

Gamma-irradiation causes more targeted protein damage in *D. radiodurans* than *E. coli*

To investigate oxidative damage to bacterial proteins, cultures were exposed to an acute dose of γ-radiation (6.7 kGy) lethal to *E. coli* but yielding 55–70% survival of *D. radiodurans*, and protein carbonyls and relative abundance changes were measured by mass spectrometry (Figs 1B, 2, and EV1). Based on previous work (Krisko & Radman, 2010), a dosage of radiation lethal to *E. coli* is required in order to observe any deleterious impact on *D. radiodurans* survival. Furthermore, our selected dosage approximates the highest reported dosage (7 kGy) used in bulk protein carbonylation measurements from whole cell lysate and dialyzed samples from both species (Krisko & Radman, 2010), providing a basis to model the impact of extrinsic protection of proteins by small molecule antioxidants. In order to limit *de novo* protein synthesis throughout and following irradiation, bacterial cultures were maintained near 0°C using a custom rack design (Dataset EV1 and EV2). Importantly, this resulted in differential relative protein abundances due specifically to oxidative damage (Materials and Methods), distinguishing our results from previous proteomic studies. Protein concentrations upon extraction were similar regardless of irradiation for each species (Appendix Table S1), and SDS–PAGE banding patterns were also qualitatively similar across protein samples extracted from the same species (Appendix Fig S2). Altogether, these results suggest that cell membrane integrity was preserved upon radiation.

As expected (Krisko & Radman, 2010), we observed carbonylation of more proteins in *E. coli* (~700 CS in 102 of 1,373 identified proteins) than in *D. radiodurans* (~400 CS in 70 of 1,264 identified proteins) under either unirradiated or irradiated conditions

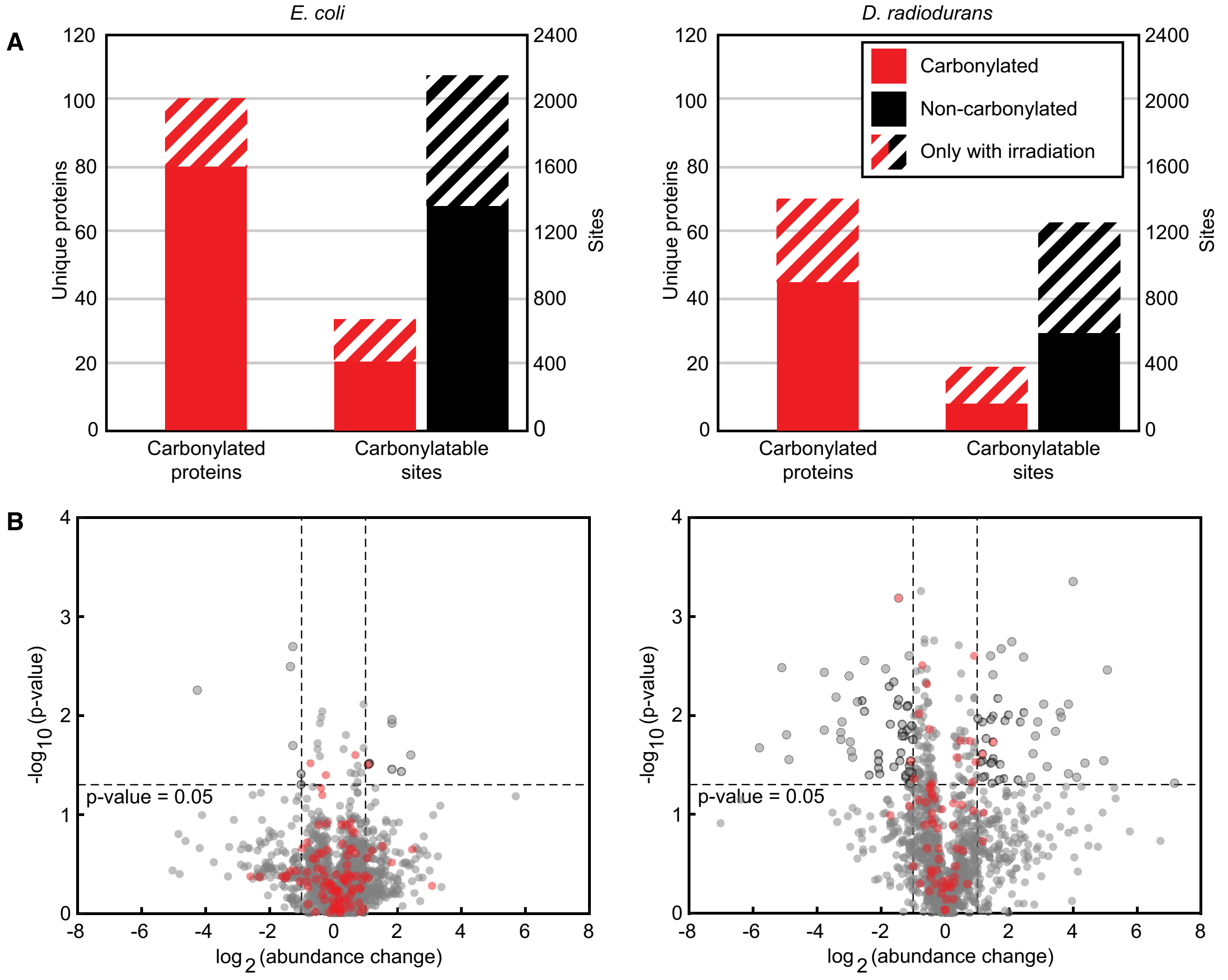

Figure 2. Summary of shotgun redox proteomic data.

A Total carbonyl-bearing proteins detected by shotgun redox proteomic measurement in three biological replicates each of *E. coli* and *D. radiodurans* with and without irradiation. The left axis is the number of sequence-unique proteins detected as carbonylated. The right axis is the number of sites in total detected as carbonylated (red) or not oxidized (black) in peptides bearing at least one carbonyl. Stripes indicate carbonylated proteins and carbonylatable sites detected only in irradiated samples. See also Appendix Fig S1.

B Volcano plots for relative protein abundance changes measured by mass spectrometry in *E. coli* (left) and *D. radiodurans* (right) after irradiation using the same biological replicates as in Fig 2A. Black-circled points are those proteins with significant changes (paired, 2-sided *t*-test *P*-value < 0.05) of > 2-fold or < 0.5-fold. Red points are proteins with at least one carbonylated peptide detected. Fold change and *P*-value cutoffs considered for significance are indicated by dashed lines. See also Fig EV1.

(Fig 2A and Table EV1). *D. radiodurans* showed similar detection rates to that in *Photobacterium angustum* exposed to UVB (62 carbonylated proteins of 1,221 identified) using the same redox proteomic technique (Matallana-Surget *et al*, 2013). The lesser total protein carbonylation in *D. radiodurans* was likely due to its effective ROS detoxification mechanisms (Slade & Radman, 2011). CS saturation curves suggest the fewer detected carbonylation events in *D. radiodurans* account for a greater percent coverage of all *in vivo* events than is the case for *E. coli* (85 and 27%, respectively; Fig EV1B), in agreement with the difference in oxidative stress sensitivity between these species. Slightly more unique proteins were detected as carbonylated in a radiation-dependent manner in *D. radiodurans* (25) than in *E. coli* (20; Fig 2A). Based on the much lower estimated coverage of all *in vivo* carbonylation in *E. coli*, we suggest that extensive damage to the *E. coli* proteome—leading to more degraded and aggregated proteins—hindered identification of some carbonylated peptides by mass spectrometry.

Relative protein quantification provided clear evidence of contrasting differential protein damage distinguishing these organisms (Fig 2B and Table EV2). Although in *E. coli* only six proteins showed significant > 2-fold differential relative abundance (paired *t*-test *P*-value < 0.05), 163 proteins overall showed > 2-fold changes albeit with higher variability across replicates. In *D. radiodurans*, 81 proteins significantly changed in relative abundance by > 2-fold; the magnitude of change was greater on average with lower variability than in *E. coli*. Proteins for which we detected at least one CS decreased in relative abundance more than other proteins in *D. radiodurans* (unpaired *t*-test *P*-value = 0.031), illustrating the expected relationship between carbonylation and degree of protein degradation. However, this relationship was less prominent in our *E. coli* data (unpaired *t*-test *P*-value = 0.104). Hence, although *E. coli* accumulated more protein carbonyls overall, their distribution is broader across distinct protein species, providing evidence of more protein-specific mechanisms for protection against ROS in *D. radiodurans* that are absent in *E. coli*.

Analogous relative peptide quantification was also performed. For *D. radiodurans*, 148 peptides representing 134 unique proteins significantly increased in relative abundance (fold change > 2, satisfying Benjamini–Hochberg criteria with false discovery rate of 0.05) after irradiation, and one peptide significantly decreased (fold change < 0.5, satisfying Benjamini-Hochberg criteria). For *E. coli*, 26 peptides representing 25 unique proteins significantly decreased in relative abundance after irradiation, and no peptides significantly increased. No individual carbonylated peptides significantly changed in relative abundance in either species. These observations generally parallel the anticipated contrasting response upon irradiation of these species. However, greater statistical power is achieved when pooling peptides to evaluate abundance changes at the whole-protein level. This is partly because stochastically missed tryptic sites and post-translational modifications lead to imperfect peptide identity when quantifying at the peptide level.

Broad functional characterization of proteins with substantial relative abundance change (<0.5-fold or > 2-fold) was carried out by Gene Ontology (GO) biological process term enrichment analysis with protein abundance correction (Scholz *et al*, 2015). These proteins in *E. coli* exhibited no significantly over- or underrepresented GO annotations. In contrast, *D. radiodurans* proteins with > 2-fold relative increase were overrepresented by proteins involved in translation and broader protein metabolism (Table 1), including many ribosomal subunits. Additionally, *D. radiodurans* proteins with < 0.5-fold change underrepresented proteins involved in nitrogen compound biosynthesis, indirectly implicating the importance of amino acid and nucleotide synthesis. Therefore, resistance to protein oxidation in *D. radiodurans* preferentially protects the critical process of proteome regeneration under oxidative stress.

Amino acid composition protects against oxidative damage

Although the relative frequency of carbonylated RKTP residues generally confirmed previous studies (Rao & Moller, 2011; Matallana-Surget *et al*, 2013), we found lysine to be as susceptible as proline to carbonylation under γ-irradiation (Fig 3A) in *D. radiodurans* (ratio 1.77 versus 1.66) and to a lesser extent in *E. coli* (ratio 1.17 versus 1.43). Protein carbonylation by natively generated ROS in eukaryotes (Rao & Moller, 2011) and UV irradiation in *P. angustum* (Matallana-Surget *et al*, 2013) both indicated proline as the most ROS susceptible of RKPT and lysine as not especially or least susceptible, respectively. Proline carbonylation often leads to polypeptide self-cleavage, which may explain the relatively low proline content of bacterial ribosomal versus non-ribosomal

Table 1. Gene Ontology terms enriched among *D. radiodurans* proteins with high relative abundance change.

Retained or lost	GO ID	Over-/ underrepresented	% foreground	% background	Fold enrichment	Foreground count	Background count	P-value	GO biological process
Retained > 2-fold	GO:0006412	O	16.67	7.58	2.20	22	10	0.037	Translation
	GO:0006518	O	19.70	9.09	2.17	26	12	0.022	Peptide metabolic process
	GO:0044267	O	21.97	11.36	1.93	29	15	0.031	Cellular protein metabolic process
	GO:0009059	O	24.24	12.88	1.88	32	17	0.026	Macromolecule biosynthetic process
	GO:0019538	O	28.03	15.91	1.76	37	21	0.025	Protein metabolic process
	GO:0009987	O	73.49	61.36	1.20	97	81	0.049	Cellular process
Lost < 0.5-fold	GO:0044271	U	12.12	27.27	0.44	8	18	0.048	Cellular nitrogen compound biosynthetic process

proteins (Lott *et al*, 2013), an evolutionary adaptation contributing to protection of translation against oxidative stress. In contrast, lysine, found incorporated into proteins much more frequently, lacks a similar mechanism for self-cleavage upon carbonylation. The more complex role of lysine in oxidative stress is discussed below.

Selective amino acid composition is a major adaptation organisms have evolved to thrive in diverse environmental niches (Brbic *et al*, 2015). Comparing compositions between expressed proteomes of *E. coli* and *D. radiodurans* under permissive conditions (Fig 3B) revealed significant differences among oxidizable amino acids. Lysine and arginine, both positively charged at physiological pH, differ in ROS susceptibility and exhibited significant usage differences. While highly susceptible lysine was found to be less frequently used in *D. radiodurans*, less susceptible arginine was overrepresented instead (0.71-fold and 1.57-fold, respectively). Reversibly oxidizable sulfur-containing amino acids, cysteine and methionine, were rare in both species, but significantly less prevalent in *D. radiodurans* under permissive conditions (0.53-fold and 0.17-fold, respectively). Surface methionines and cysteines help protect proteins from oxidative damage in many organisms due to their own reversible oxidation (Stadtman & Levine, 2003). However, cysteine and methionine are metabolically expensive (i.e., stoichiometrically consume the most ATP) for bacterial synthesis (Kaleta *et al*, 2013), and *D. radiodurans* is auxotrophic for methionine (Zhou *et al*, 2017), which may explain their significantly lower prevalence in slower-growing *D. radiodurans* despite expected benefits for resistance. Tryptophan and tyrosine, two metabolically inexpensive amino acids that function as integrated antioxidants in some proteins (Moosmann & Behl, 2000), were significantly more abundant in *D. radiodurans* than in *E. coli* (both ~1.3-fold).

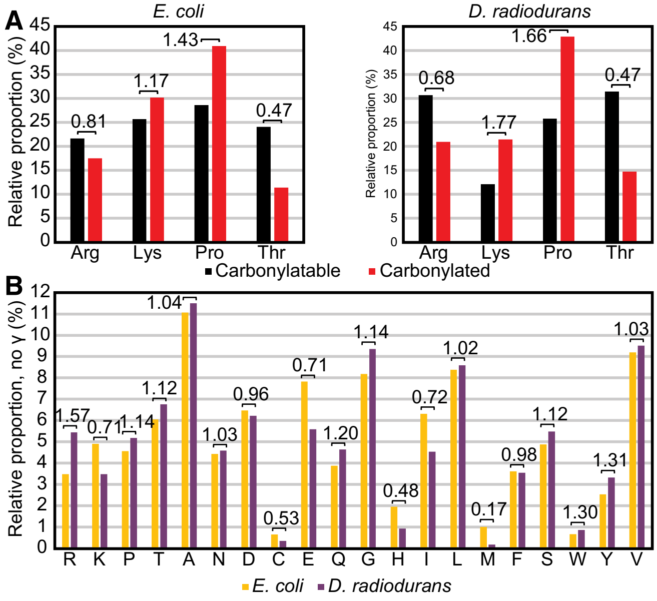

Figure 3. Amino acid prevalence in proteomic data before and after irradiation.

A Prevalence of individual RKPT residues and prevalence of carbonylated form in experimentally measured peptides combining all three biological replicates of both conditions for each organism. Ratios are given above each pair of bars. All proportions are significantly different between each RKPT and their respective carbonylation state by two-tailed z-test of two proportions (*P*-values < 0.01; see Materials and Methods), and meaning carbonylated proportions are not determined simply by relative prevalence of RKPT. See also Appendix Fig S1.

B Prevalence of all canonical amino acids before irradiation of *E. coli* and *D. radiodurans*, combining all three biological replicates for each condition. Ratios are given above each pair of bars. All proportions are significantly different between species by two-tailed *z*-test of two proportions (*P*-values < 0.01). See also Figs EV1 and EV2.

To evaluate the impact of oxidative stress on amino acid prevalence in identified proteins, we compared changes in amino acid composition after γ-irradiation of *E. coli* and *D. radiodurans* (Fig EV2). While only seven amino acids significantly changed in *E. coli*, 16 significantly changed in *D. radiodurans* and to a greater magnitude. The greatest decrease among RKPT was lysine in both species, further supporting that incorporated lysine is an important mediator of protein oxidative damage under γ-irradiation. Lysine can sometimes be exchanged for histidine in proteins and still preserve protein function as shown in synthetic mutational studies (Yampolsky & Stoltzfus, 2005). Notably, relative histidine prevalence increased modestly (+2%) in *E. coli* and significantly (+11%) in *D. radiodurans* after irradiation, suggesting that *D. radiodurans* has evolved proteins that are more composed of non-carbonylatable histidine rather than lysine as another protein-intrinsic protection mechanism. Indeed, across sequences of functional orthologs and isozymes in these species (Appendix Fig S3) we found 10% greater histidine composition in *D. radiodurans* than in *E. coli* as a fraction of total histidine and lysine (paired *t*-test *P*-value $< 6 \times 10^{-60}$). Following irradiation, tyrosine prevalence significantly increased in *E. coli* (+4%) and in *D. radiodurans* (+8%), and cysteine increased significantly (+18%) only in *D. radiodurans*. The most significant decrease in *E. coli* (−13%) and increase in *D. radiodurans* (+45%) was for methionine. This contrast suggests a more efficient methionine sulfoxide reductase system under oxidative stress in *D. radiodurans*. All together, these results establish that protein-intrinsic properties, even in primary structure, differ between *E. coli* and *D. radiodurans* and affect which proteins withstand the onslaught of ROS-induced oxidative damage.

Structure- and sequence-based model predict protein vulnerability to carbonylation

Structure-based molecular feature engineering

The computational phase of this study (Fig 1B) involved proteome-wide derivation of 3D structures to investigate molecular properties contributing to ROS susceptibility (Fig 4A, Table EV3, and Materials and Methods). Due to incomplete proteome coverage by crystal structures (<3% for *D. radiodurans* proteins), computation of molecular features required high-throughput modeling of single-chain proteins, which we performed *de novo* for *D. radiodurans* and used published models for *E. coli* (Xu & Zhang, 2013b; Yang *et al*, 2015). The challenge of deriving *D. radiodurans* proteins by available modeling strategies is summarized in Fig EV3A. The best representative model from alternative methods (Appendix Table S2) for each protein was selected using multiple structure quality metrics (Appendix Table S3). Models generally evaluated comparably to crystal structures for *D. radiodurans* proteins by these metrics

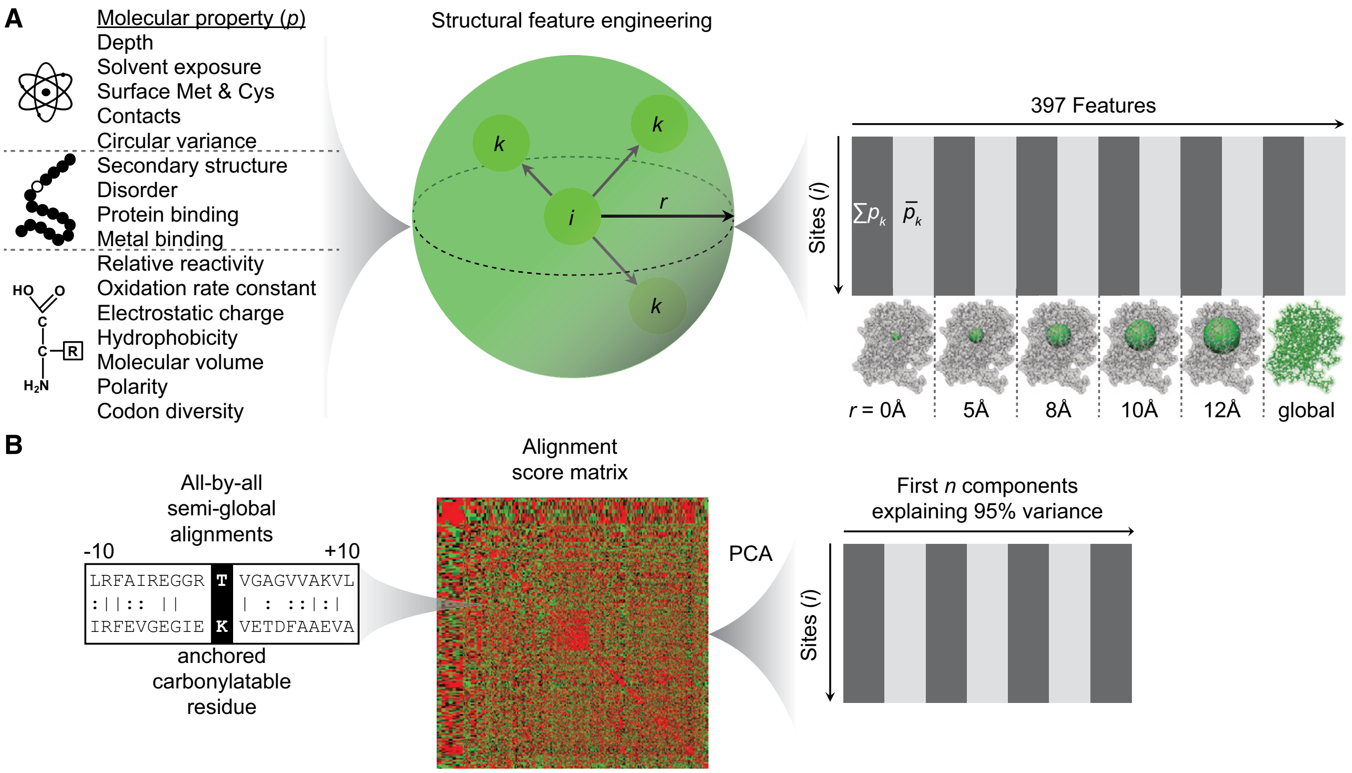

Figure 4. Feature engineering.

A Three-dimensional feature engineering from molecular properties. Initial properties that can be determined only with an atomic resolution structure, in the context of an amino acid sequence, or that depend only on amino acid identity are denoted at left. This property list is a non-redundant abbreviated set of all properties considered (see Appendix Table S4 and Materials and Methods for full detail). Columns of the feature matrix at right are alternating property sums and means at spatial scales denoted below matrix. p = a molecular property; i = RKPT residue; k = neighbor residues of i; r = radius length. See also Fig EV3.

B Sequence homology-based features for machine learning were derived by performing sequence alignments of all RKPT sites (± 10 residues) anchored at the central residue to compute alignment scores that were then reduced to a computationally manageable number of features by principal component analysis (PCA).

(Fig EV3B and Table EV4). Best representative models were obtained for >95% of *D. radiodurans* proteins (Fig EV3C), most commonly resulting from I-TASSER (Yang *et al*, 2015) or ProtMod (http://protmod.godziklab.org/protmod-cgi/protModHome.pl). Future replacement with higher quality models or experimentally determined structures could improve the performance of our algorithm.

We engineered for the first time molecular features at multiple spatial scales using 3D structures (Fig 4A, Table EV3, Appendix Table S4, and Materials and Methods) to predict carbonylation. Features were computed with respect to all RKPT across *D. radiodurans* and *E. coli* proteomes. These features quantitatively summarize the molecular environment of carbonylatable sites. Statistical summaries of local structural properties were computed as the sums and means of canonical property values for neighboring residues within multiple radii to account for a gradient of scales. This feature engineering strategy enabled incorporation of more molecular properties and with spatial dimensionality than possible using sequences alone to represent proteins.

Combining structure- and sequence-based approaches for machine learning

In addition to structure-derived features, we implemented simple sequence alignment-based feature engineering to predict CS (Fig 4B). We defined a local neighborhood centered on each RKPT covered by carbonylated peptides in our proteomic data and performed all-by-all pairwise sequence alignments of these regions, using the alignment score matrix as potential predictive features. This alignment-based approach is agnostic to specific sequence motifs while still leveraging any useful local sequence homology across CS.

All RKPT from carbonylated peptides were mapped to respective protein structure and sequence to assign carbonylated and non-carbonylated residues. Unlike previous CS prediction efforts (Maisonneuve *et al*, 2009; Lv *et al*, 2014; Weng *et al*, 2017), we did not assume that any given RKPT is deterministically carbonylated or not. Protein carbonylation is an inherently stochastic process. Therefore, we took a probabilistic approach and used all of the carbonylated peptide data regardless of site redundancy or occurrence as carbonylated in one peptide but non-carbonylated in another. Previous approaches also often sampled unmodified RKPT across all detected peptides, carbonylated or not, to define negatives for training. Compared to non-carbonylated peptides, unmodified RKPT on peptides bearing a carbonyl on another residue better-represent negative data because it is certain that those molecules were directly exposed to ROS yet did not react with ROS.

Independent probability estimators for CS were trained by logistic regression using structure-based features and sequence-based features and then combined into a stacked model. Each independent model and the stacked model were evaluated by leave-1-out validation and their performance quantified by receiver operating characteristic (ROC) analysis (Fig 5A from data in Tables EV5 and EV6). At the residue scale, our stacked model outperformed (AUC_{norm} = 0.73) each of its structure- and sequence-based components. Shuffling each feature before training yielded random performance (AUC_{norm} = 0.54), strongly supporting the predictive power of our engineered features. We also evaluated performance of our model for predicting protein-scale vulnerability to oxidation (Fig 5B) by calculating a CS enrichment metric. Predicted carbonylation enrichments for training set proteins strongly rank correlate with enrichments derived from measured carbonylated peptides (Spearman ρ = 0.82, permutation test P-value = 1.3×10^{-22} for *E. coli* and Spearman ρ = 0.87, permutation test P-value = 7.2×10^{-21} for *D. radiodurans*), signifying that our model can predict relative propensity to carbonylation of different protein species. Due to prioritized sensitivity, our model tends to predict higher enrichment values than derived experimentally (1.9-fold on average for *E. coli* and 1.7-fold for *D. radiodurans*), but these predicted enrichment values are plausible given the fact that *in vivo* carbonylation events are undersampled experimentally (Fig EV1B).

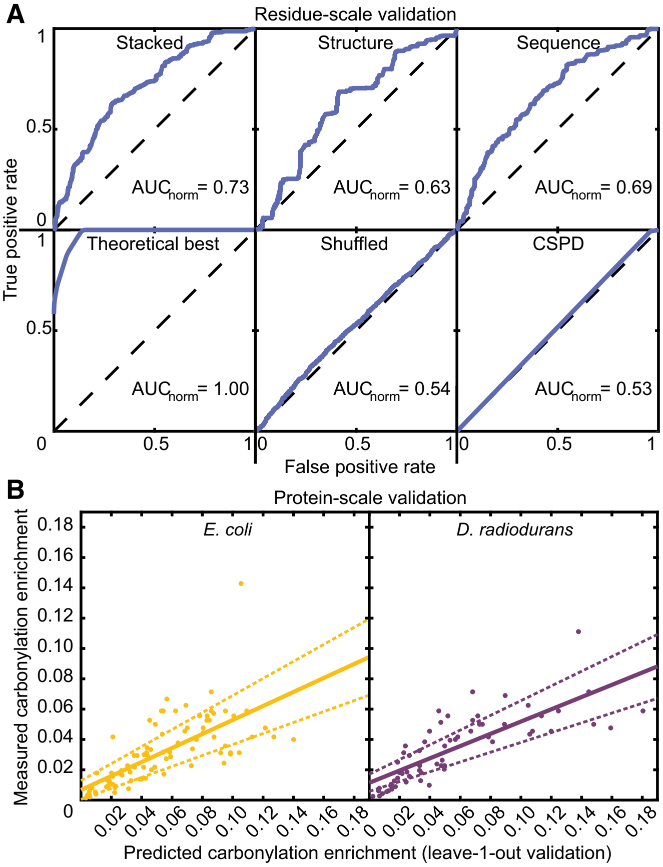

Figure 5. Multi-scale validation of protein carbonylation predictor.

A Residue-scale validation: Receiver operating characteristic (ROC) curves for CS predictors derived by leave-1-out validation. The dashed black line at y=x corresponds to performance expected by chance. Top left = final predictor trained by stacking structure- and sequence-based models. Top middle = predictor trained only on structure-based features. Top right = predictor trained only on sequence-based features. Bottom left = theoretical maximum predictive power for a probability estimator (AUC = 0.98). Bottom middle = same algorithm as used for final predictor but with all features shuffled beforehand. Bottom right = CSPD model developed using metal-catalyzed oxidation (MCO) site data from *E. coli*. See also Figs EV3 and EV4.

B Protein-scale validation: Comparison between predicted CS enrichment from leave-1-out validation to CS enrichment computed from all carbonylated peptides measured for *E. coli* (left) and *D. radiodurans* (right). Each point represents a different protein species. Predicted probability-weighted CS enrichment = (sum of carbonylation probabilities across training set sites)/(number of residues in corresponding peptides from experiments). Experimentally measured probability-weighted CS enrichment = (sum of empirical oxidation probabilities across training set sites)/(number of residues in corresponding peptides from experiments). The solid line is the fitted regression line, and dashed lines indicate the boundaries of the 95% confidence interval.

Molecular properties explain vulnerability to carbonylation

Although we included ~400 structure-based features in the modeling, only seven of the logistic regression coefficients were nonzero: relative reactivity with ROS (reactivity_res), codon diversity, whether the RKPT site was a threonine residue, molecular volume, local solvent accessible surface area, local positive charge, and local lysine residues. Codon diversity (AAindCodonDiv_res) itself is unlikely to be causal. Instead, this feature has the same rank order as carbonylation prevalence in *D. radiodurans* from our experiments (Fig 3A) and is therefore a fortuitous proxy for γ-specific reactivity. Threonine is by far the least frequently carbonylated of RKPT in both species (Fig 3A), and inclusion of this feature (Thr_res) in our model reflects this lower propensity to reaction with ROS.

Aside from the reactivity features differentiating RKPT, all other explanatory properties for ROS susceptibility derived from 3D structures (Fig 6). Accessibility to ROS promotes carbonylation (Fig 6A). The lower the molecular volume of a residue (AAindMolVol_res), the more likely it will react with ROS due to lower steric effects. Similarly, lower local surface area (areaSAS_5A_sum) surrounding a near-surface site indicates less likelihood of shielding by surrounding structure, such as the protrusion in Fig 6D. Local positive charges (posCharge_8A_sum) promote carbonylation by attracting negatively charged superoxide radicals (Fig 6B). Colocalization of highly reactive sites may cause progressive protein misfolding, exposing neighboring residues to ROS (Maisonneuve *et al*, 2009; Fig 6C). In our model, neighboring lysine residues (Lys_8A_sum) contribute to the probability of carbonylation, lysine being the most prevalently carbonylated RKPT under γ-irradiation in our data (Fig 3A). Polarity leading to solubility of lysine-rich regions could also contribute to this effect. Sites without neighboring lysines are less likely to be carbonylated (Fig 6D).

Our algorithm also extends to prediction of metal-catalyzed oxidation

We applied Carbonylated Site and Protein Detection (CSPD) developed by Maisonneuve *et al* (2009) to predict CS across our training set (Fig 5A). CSPD performance on our data was essentially random (AUC_{norm} = 0.53). It is important to note that CSPD was developed using metal-catalyzed oxidation (MCO) data from a set of only 23

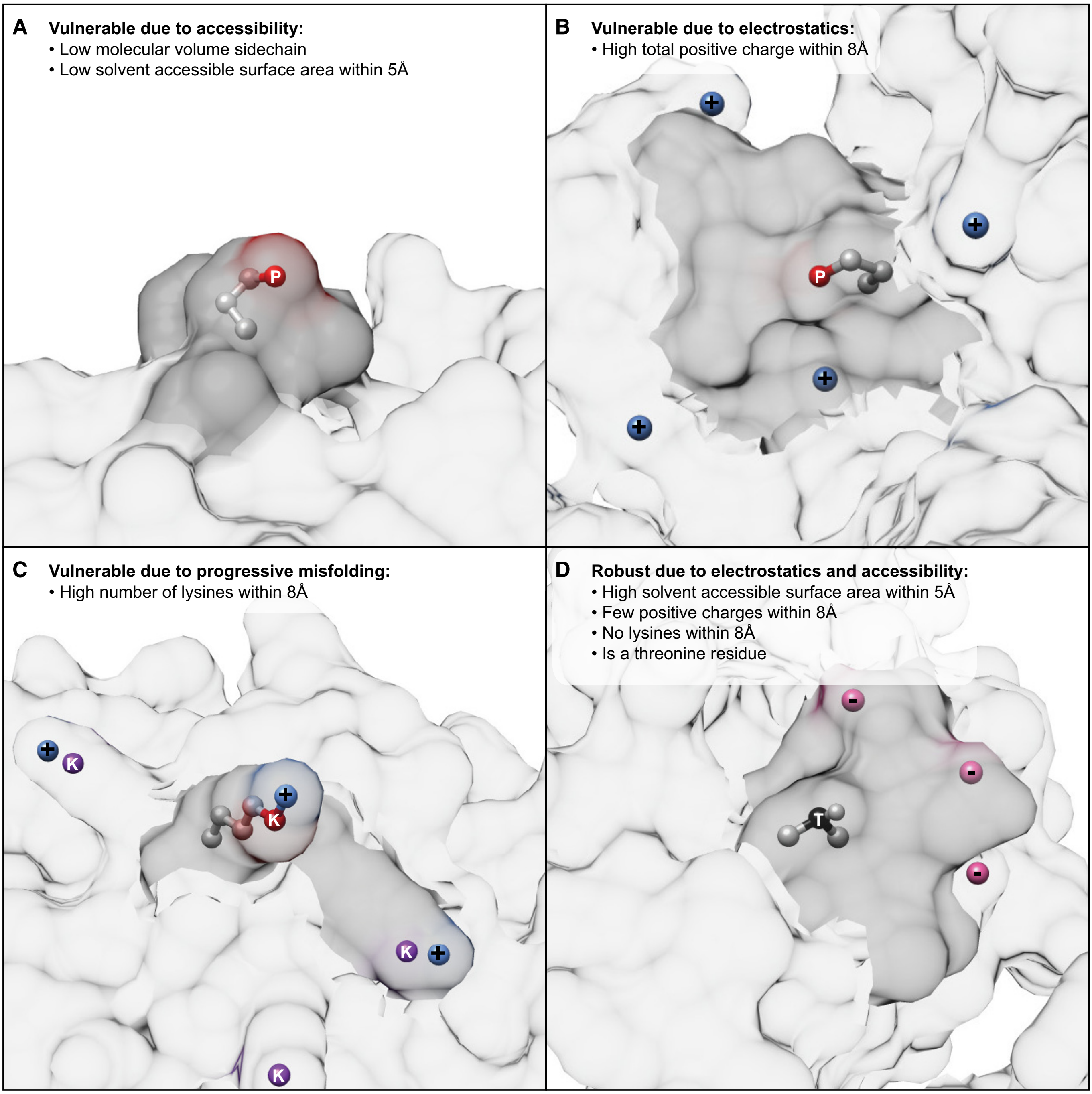
A
Vulnerable due to accessibility:
• Low molecular volume sidechain
• Low solvent accessible surface area within 5Å
P
B
Vulnerable due to electrostatics:
• High total positive charge within 8Å
P
+
C
Vulnerable due to progressive misfolding:
• High number of lysines within 8Å
K
+
D
Robust due to electrostatics and accessibility:
• High solvent accessible surface area within 5Å
• Few positive charges within 8Å
• No lysines within 8Å
• Is a threonine residue
T
-

Figure 6. Molecular properties predicting protein vulnerability to carbonylation.

A–D Example sites prone to carbonylation. (A) DRA0302_P252, (B) DR0099_P51, and (C) b0911_K411; and example robust site (D) b3313_P69.

Data information: All atoms of central RKPT side chains are shown, with carbonylatable atomic site in red (predicted and measured carbonylated) or black (predicted and measured not oxidized) and labeled with the 1-letter code of the containing amino acid. Positive (blue) and negative (pink) charges within 8 Å are labeled. Carbonylatable lysine sites (purple) within 8 Å are labeled. Molecular surfaces within 5 Å of the central CS are dark gray. See also Fig EV3.

carbonylated *E. coli* proteins derived from samples prepared under similar conditions to our negative controls. However, while we kept our samples on ice after harvesting the exponential phase cells, Maisonneuve *et al* did not report any similar temperature treatment for their samples. In this way, the samples of Maisonneuve *et al* being prepared at higher temperature allowed protein synthesis and turnover that would have led to fewer detectable carbonylated proteins than we measured. Furthermore, Maisonneuve *et al* performed 2D SDS–PAGE and excised only visible spots labeled for carbonylation, which could have further limited the number of distinct carbonylated proteins identified from their samples. In all, we identified 82 carbonylated proteins in our *E. coli*-negative controls, including 10 in common with the Maisonneuve *et al* data. The inability of CSPD to generalize to carbonylation from γ-irradiation may be due in part to the experimental differences noted above in addition to a difference in effects of each specific source of ROS. Therefore, to more directly compare algorithmic performance we also used our algorithm to train a model predicting MCO using the same redox proteomic data used to develop CSPD (Fig EV4). CSPD showed modest positive performance on this dataset (AUC_{norm} = 0.58), the discrepancy in previously reported performance owing to our inclusion of all carbonylated peptides with carbonylated and non-carbonylated residues defined as described above. We conclude that CSPD was overfitted to the MCO data and depends on the assumption of deterministic protein carbonylation and on less-strict standards for defining non-carbonylated residues in proteomic data.

Furthermore, our stacked model for MCO prediction performed better (AUC_{norm} = 0.75) than our γ-induced oxidation model with better synergy in stacking the structure- (AUC_{norm} = 0.72) and sequence-based (AUC_{norm} = 0.67) models. This performance difference was likely due to the relatively less diverse products of MCO than γ-induced oxidation. ROS production in MCO is more localized because it depends on the presence of Fe or Cu cations to drive the Fenton reaction and therefore affects a smaller number of proteins than γ-induced oxidation. Indeed, data from γ-irradiation experiments include not only CS caused by ROS from water radiolysis but also basal cellular oxidation due to native ROS sources, including MCO and cellular respiration. Thus, oxidation from γ-irradiation is more diverse and complex than MCO products and more challenging for learning structure and sequence signatures.

Intra- and interspecies differences in protein vulnerability to carbonylation

D. radiodurans proteome maintenance is protected from carbonylation

Orthologs and isozymes mapped between *E. coli* and *D. radiodurans* (Appendix Fig S3) were compared by their unweighted carbonylation enrichment (Fig 7 and Table EV7) as computed from proteome-wide CS prediction in *E. coli* and *D. radiodurans* to reveal functional classes and individual proteins differing in susceptibility between and within these proteomes. Functional classes known to be involved in resistance and recovery from oxidative stress include the following: ribosomal, ribosomal assembly, translation, protein chaperone, protease and peptidase, amino acid and peptide transport, DNA repair, DNA damage response and regulation of repair, native ROS production, ROS detoxification, ROS response, metal transport, terpenoid synthesis, and polyamine accumulation.

Pairwise orthologs were compared based on protein-intrinsic and extrinsic factors contributing to their propensity to carbonylation (Fig 7). Perpendicular distance to the y = x diagonal represents the relative degree to which one ortholog is intrinsically more or less sensitive given the same ROS dosage on the basis of carbonylation enrichment alone. Protein-extrinsic factors, such as the Mn-dependent scavenging system in *D. radiodurans* (Daly, 2012) and the antioxidant carotenoid deinoxanthin (Tian *et al*, 2009), also contribute to interspecies differences in protein oxidation. Such protein-extrinsic factors act broadly by reducing the effective cellular dosage of ROS. An acute gamma dosage of 7 kGy, approximately the same as in this study, yielded about 3.78-fold more protein carbonyls in *E. coli* lysate than in *D. radiodurans* (Materials and Methods) due to small molecules removable by dialysis (Krisko & Radman, 2010). Assuming such factors act globally without favoring protection of specific proteins, the degree to which these extrinsic factors differentiate vulnerability to carbonylation between orthologs can be modeled in combination with protein-intrinsic factors simply by computing perpendicular distance to the y = x/3.78 diagonal (Fig 7). By this model, especially susceptible proteins benefit more from an effectively lower dosage of ROS in *D. radiodurans*.

Relative vulnerability to ROS differed between *E. coli* and *D. radiodurans* within particular functional classes (Fig 7). We predicted the intrinsic susceptibility of *E. coli* ribosomal proteins to be more than 2.4-fold greater than across all-orthologs (unpaired *t*-test *P*-value = 0.01). Accounting for extrinsic ROS protection predicted ribosomal proteins to be the most favored functional class in *D. radiodurans* over *E. coli* (1.5-fold, unpaired *t*-test *P*-value = 1.2×10^{-26}), in agreement with *D. radiodurans* ribosomal proteins being enriched among those with relative abundance increases after irradiation. Protein chaperones in *E. coli* were predicted on average 1.13-fold more intrinsically vulnerable than in *D. radiodurans* (unpaired *t*-test *P*-value = 0.02), a difference further distinguished due to being more than 4.5-fold greater than the difference across all-orthologs (unpaired *t*-test *P*-value = 0.003) and 1.14-fold greater when accounting for extrinsic protection as well (unpaired *t*-test *P*-value = 0.02). *E. coli* proteins involved in polyamine synthesis and uptake are predicted to be more than 3.7-fold intrinsically vulnerable than across all-orthologs (unpaired *t*-test *P*-value = 0.04). Revisiting the observation that methionine usage featured prominently in *D. radiodurans* proteins retained after irradiation, we predicted that methionine sulfoxide reductases acting on protein-incorporated methionine MsrB and MsrP are both 1.4-fold more intrinsically sensitive to carbonylation in *E. coli*. MsrP was also in the 94th percentile of proteins benefiting from extrinsic protection in *D. radiodurans*.

Comparison of interspecies outliers reveals proteins involved in oxidative stress resistance

Many proteins involved in coping with oxidative stress were significant outliers in predicted intrinsic vulnerability to carbonylation (Fig 7). There were 111 orthologous pairs greater than 3 standard deviations of distance from the mean of the distribution or greater than 3 standard deviations away from the mean perpendicular distance from the y = x diagonal. We grouped these outliers according to three properties: (i) intrinsic sensitivity or robustness compared to the rest of the proteome, (ii) comparative intrinsic vulnerability between *D. radiodurans* and *E. coli*, and (iii) relative effect of ROS detoxification in *D. radiodurans* over *E. coli* (Fig EV5).

Proteins predicted as significantly more intrinsically or extrinsically protected from ROS in *D. radiodurans* relative to *E. coli* fall into three groups based on the three properties described above. Group 1 proteins were predicted highly carbonylation-prone but more protected intrinsically and extrinsically in *D. radiodurans* than in *E. coli*. On average, these 12 proteins were 1.4-fold more CS-enriched in *E. coli* and above the 99th percentile of extrinsic protection in *D. radiodurans*. Of 10 proteins detected in both organisms by proteomics, eight had more negative γ-induced relative abundance

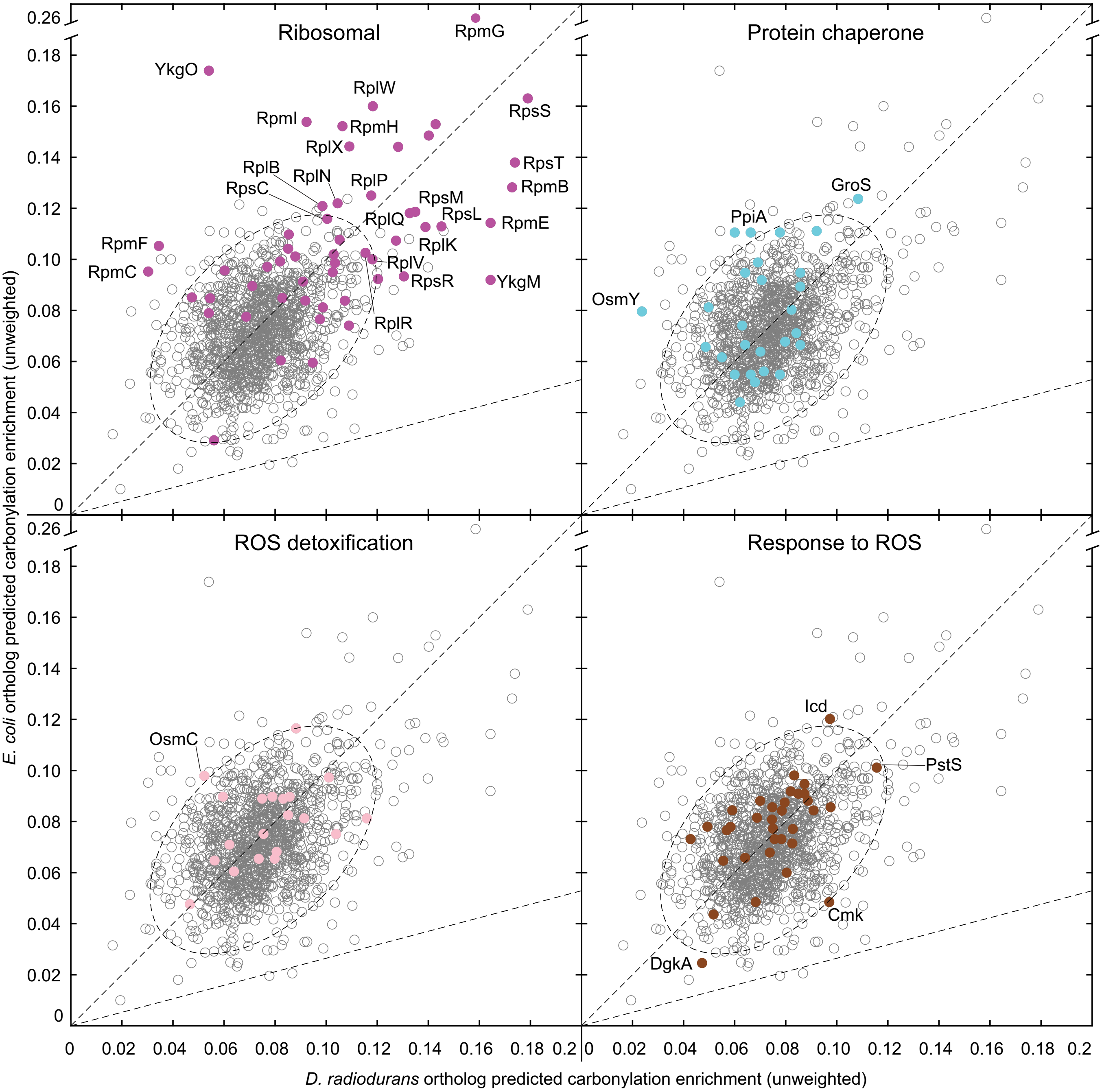

Figure 7. Interspecies comparison of predicted protein vulnerability to carbonylation.

Each circle represents a distinct protein pair (ortholog or isozyme) between species. Each plot shows identical point values but highlights a different functional class of proteins with relevance to oxidative stress. The y = x diagonal line is a reference to compare intrinsic vulnerability to carbonylation between orthologs. The y = x/3.78 diagonal line is a reference to compare combined intrinsic and extrinsic carbonylation properties between orthologs. Elliptical dotted line encircles the points falling within 3 standard deviations of the mean coordinates and 3 standard deviations of the distance from reference line y=x, encompassing ~91% of all data points. This reference region distinguishes outlier points that are distant from the main population. Outliers with associated experimental evidence related to hypersensitivity to oxidative stress are labeled with their protein names. See also Appendix Fig S3 and Fig EV5.

changes in *E. coli* than *D. radiodurans*, with a median *E. coli*-to-*D. radiodurans* ratio of 0.47. Ribosomal subunits comprised 11 of these proteins, eight of which are essential in *E. coli*. *E. coli* knockouts of *rpmI* (Nakayashiki & Mori, 2013) are hypersensitive to oxidative stress. Overexpression of *rpmG* increases resistance to oxidative stress from mitomycin C (Bolt *et al*, 2015), and GroS overexpression decreases protein carbonyl accumulation (Fredriksson *et al*, 2005). Seven of these proteins exhibit oxidative stress-induced expression in *D. radiodurans* (Liu *et al*, 2003; Slade & Radman, 2011). Group 2 proteins were predicted as similarly intrinsically carbonylation-prone in both species but significantly extrinsically protected in *D. radiodurans*. On average, these 22 proteins are above the 86th percentile of extrinsic protection in *D. radiodurans*. Of 13 proteins detected in both organisms by proteomics, 11 showed substantially more positive γ-induced relative abundance changes in *D. radiodurans*. In this group, 13 proteins are ribosomal subunits. In *E. coli*, *pstS* knockouts are hypersensitive to oxidative stress (Sargentini *et al*, 2016), *rpsL* mutants have been shown to affect oxidative stress tolerance (Ballesteros *et al*, 2001; Miskinyte & Gordo, 2013), and 13 others are essential genes (Baba *et al*, 2006; Bubunenko *et al*, 2007). In *D. radiodurans, rpsS* and *hupA* knockouts are hypersensitive to oxidative stress (Dulermo *et al*, 2015), and overexpression of *rpsS*, *rpsT*, *rplQ*, *rpsM*, *rpmB*, *rplK*, *rpsL*, *thpR*, *rpmE*, *nrdH*, *rplR*, *rplV*, and *rpsR* occurs during oxidative stress (Liu *et al*, 2003; Slade & Radman, 2011). Group 3 proteins were predicted significantly more susceptible to carbonylation in *E. coli* than in *D. radiodurans*. On average, these 27 proteins were 1.9-fold more CS-enriched in *E. coli* and above the 95th percentile of extrinsic protection in *D. radiodurans*. In *E. coli*, *rpmF* (Nakayashiki & Mori, 2013; Sargentini *et al*, 2016) and *icd* (Krisko *et al*, 2014) knockouts are hypersensitive to oxidative stress, and *osmY* (Basak & Jiang, 2012) is also involved in oxidative stress resistance. In *D. radiodurans, xseB* knockouts are hypersensitive to oxidative stress (Dulermo *et al*, 2015), and *adk*, *icd*, *malE*, *osmC*, *ppiA*, *rplB*, *rpmC*, *rpsC*, and *yceI* are highly expressed under oxidative stress (Liu *et al*, 2003; Slade & Radman, 2011; Basu & Apte, 2012). Higher resistance to carbonylation of proteins from these groups sets *D. radiodurans* apart from *E. coli* and delineates transgenes that could serve to increase stress tolerance in *E. coli*.

Interspecies outliers not predicted as significantly more protected from ROS in *D. radiodurans* fall into two groups. Group 4 proteins were predicted as highly intrinsically robust to carbonylation in both species and therefore not to benefit substantially from extrinsic protection in *D. radiodurans*. Of these five proteins, three were more intrinsically vulnerable in *E. coli*, including *secE*, which is essential in *E. coli* (Baba *et al*, 2006), and *fdx*, which is highly expressed under oxidative stress in *D. radiodurans* (Liu *et al*, 2003). Group 5 proteins were predicted as significantly more intrinsically vulnerable to carbonylation in *D. radiodurans* than in *E. coli*. These 14 functionally diverse proteins include three known oxidative stress-hypersensitive knockout mutants in *D. radiodurans* (Dulermo *et al*, 2015); however, all but 2 still lie above $y = x/3.78$ in Fig 7, suggesting that extrinsic protection could still compensate for intrinsic vulnerability differences between these species.

Discussion

In this study, we successfully developed a highly integrative systems biology approach that predicts protein targets of oxidative stress and offers mechanistic explanations for cellular phenotypes at multiple biological scales spanning amino acid residues, protein molecules, and protein functional classes. This constitutes a major advancement in development of predictive techniques for protein oxidation, a recognized need in the field (Krisko & Radman, 2019). We have provided extensive evidence substantiating the theory that intrinsic properties of proteins lead to differential rates of protein oxidation (Stadtman, 1986) and affect vulnerability to oxidative stress through function of key proteins involved in response phenotypes in bacteria. Multiple lines of evidence support that the susceptibility of ribosomal proteins to ROS is strongly differentiated from the rest of the proteome and plays a key role in the radioresistance—and by analogy also desiccation tolerance—of *D. radiodurans*. Furthermore, our results not only identified oxidized proteins but also characterized explanatory molecular properties of precise sites of carbonylation, providing a much finer-resolution analysis of molecular properties leading to differential protein oxidation than in previous studies (Vidovic *et al*, 2014).

A recent study reported protein oxidative products upon ionizing radiation of *E. coli* and *D. radiodurans* (Bruckbauer *et al*, 2020). Bruckbauer *et al* performed quintuplicate experiments identifying slightly more but similar numbers of peptides and proteins compared to our reported results. Like ours, their result identified a relatively small fraction of peptides bearing oxidative products and mostly low magnitude changes in abundance in ~12% of peptides. Although there were some similar findings, our study differs from that of Bruckbauer *et al* in several important ways. In our irradiation treatment, we dosed with nearly 7-fold greater radiation than their 1 kGy by electron beam linear accelerator, presumably leading to a higher ROS production through water radiolysis in the present study. Importantly, Bruckbauer *et al* neither included carbonylation of lysine, proline, or threonine residues in their search database nor performed any stabilizing derivatization of these labile oxidative products. As a result, it is highly likely that their data underrepresent actual carbonylation events. Finally, Bruckbauer *et al* used relative absolute mass (RAM) to broadly analyze the specificity of proteins targeted by ROS and concluded that target theory does not fully explain this specificity (i.e., proteins differ in ROS susceptibility due to more than just their relative sizes and abundances). One of the major goals of our study was to identify molecular properties that determine the protein and site specificity of carbonylation by ROS using our structure-informed machine learning model.

Our model suggests that a combination of protein-intrinsic properties and global ROS detoxification implicates vital proteins for resisting oxidative stress in *D. radiodurans* and explains targeted patterns in relative abundance changes following irradiation that are not observable in *E. coli*. The evolution of desiccation tolerance in an aerobe like *D. radiodurans*, by way of multiple mechanisms to protect against ROS, sharply contrasts with a facultative anaerobe like *E. coli*. The contrast is likely even more extreme in comparison to anaerobic ancestral bacteria that evolved in oxygen-poor environments and therefore theoretically would not have faced selective pressure from high ROS conditions to have evolved such mechanisms of protection. Thus, we suspect that adaptations like intrinsic protection of proteins against ROS are relatively rare and likely less prevalent among anaerobes.

Those *D. radiodurans* proteins predicted to rely primarily on intrinsic properties to avoid carbonylation are candidate transgenes to confer resistance to more sensitive species. Amino acid usage differentiating *D. radiodurans* from *E. coli* and molecular properties predictive of protein carbonylation comprise a set of design principles that may be used to control ROS tolerance in synthetic protein engineering efforts. The analytical strategy that we have established could be applied not only to the study of oxidative stress in other systems (e.g., human disease, aging, and manned space exploration) but also to other forms of post-translational modification and more broadly to molecular properties of proteins, extending and enriching proteomic analysis.

Materials and Methods

Redox proteomics

Additional details can be found in the Appendix Supplementary Material and Methods for the redox proteomics experiment including reagent preparation, bacterial culture, protein extraction and derivatization, protein concentration (Appendix Table S1) measurement, SDS–PAGE (Appendix Fig S2), trypsinization, and drying.

Gamma-irradiator culture tube rack

In order to irradiate bacterial cultures in the GC-220E ^{60}Co γ-irradiator, a custom culture tube rack was designed and fabricated by 3D printing. The design requirements for the rack were that it:

- Hold six 120-ml samples in leak-proof culture tubes, triplicates of two conditions
- Be made of γ-resistant but non-shielding materials
- Hold enough ice and provide insulation to keep samples near 0°C for ≥ 2 h
- Fit inside the GC-220E sample chamber (I.D. = 15.2 cm × H = 20 cm)
- Provide for radially symmetric sample distribution for even dosing

Six 170-ml pyrex culture tubes (O.D. = 38 mm × L = 200 mm) with screw caps (Corning® 9825-38) were shortened so that the height of the loaded rack would remain below 20 cm. A diamond cutter was used to shorten the pyrex tubes at the open end, and a wet orbital sander was used to shorten the screw caps at the open end and smooth the cut end of the pyrex tubes to fit the rack. The pyrex and resin caps are resistant to repeated acute doses of γ-radiation (http://www.sterigenics.com/services/medical_sterilization/contract_sterilization/material_consideration__irradiation_processing.pdf), although the pyrex slightly discolors upon initial irradiation.

The rack was printed using a compound primarily composed of acrylic and polyacrylate material (Stratasys® Objet RGD515). Acrylic and polyacrylate exhibit high radiation stability, up to 100 kGy with repeated exposures (http://www.nordion.com/wp-content/uploads/2014/10/GT_Gamma_Compatible_Materials.pdf). The culture tubes are held in place by hemispherical recesses in the base. The rack has space to add ice through a capped opening at the top using a funnel, sufficient to keep the samples near 0°C for up to 5 h at room temperature and at least 2 h in the GC-220E with a dose rate of 60 Gy/min. The top of the rack is removable for easy cleaning and has a key slot for the alignment of holes in the lid with hemispherical recesses in the base. STL files containing the design data required for 3D printing are available in Dataset EV1 and EV2.

Gamma-irradiation experiment

Samples in the γ-irradiator culture tube rack were placed in the sample chamber of a GC-220E ^{60}Co γ-irradiator (dose rate = 55.18 Gy/min) and irradiated for 2 h, receiving a total dose of 6.7 kGy. After irradiation, the γ-irradiator culture tube rack was refilled with fresh ice and transported (30 min) before protein extraction. Maintaining cultures near 0°C for 30 min before, throughout 2-h irradiation, and 30 min after ensured that proteomic changes were primarily due to oxidation-induced damage and not regulation of *de novo* gene expression. This is because of the deleterious effects of irradiation and cold on transcription and translation combined with the short timescale of our experiment.

Protein carbonylation leads to loss of protein function and lack of detection by shotgun proteomics due to aggregation, self-cleavage, and proteolysis (Nystrom, 2005). Aggregation accounts for ~95% of carbonylated proteins in *E. coli* (Maisonneuve *et al*, 2008), affecting solubility and detection by proteomics (Pallares & Ventura, 2016). Proline carbonylation can lead to non-enzymatic self-cleavage at the peptide backbone via the proline oxidation pathway (Uchida *et al*, 1990). Low temperature slows all enzymatic activity, but proteases targeting carbonylated proteins retain at least some rate at low temperature, such as the *E. coli* Lon protease functioning at 16°C (Sakr *et al*, 2010).

Regulation of transcription and translation in *D. radiodurans* in response to irradiation does not substantially occur until return to permissive growth conditions. Previous irradiation experiments without low temperature show that radiation-responsive promoters in *D. radiodurans* do not activate transcription until after irradiation stops, and peak transcriptional rates occur 1–2 h into recovery in fresh medium at optimal temperature (Anaganti *et al*, 2016); similar trends hold transcriptome-wide (Liu *et al*, 2003). RNA samples taken mid-irradiation, without post-irradiation recovery, result in nearly 50% fewer reads originating from mRNA relative to pre-irradiation or during recovery, and nearly all individual transcripts decrease (Luan *et al*, 2014). Similar to transcription, *D. radiodurans* proteins expressed in response to irradiation without cold temperature reach their peak translation rate 0.5–1 h after transfer to fresh media under optimal temperature (Basu & Apte, 2012).

Transcription and translation both substantially slow or halt under cold temperatures in *E. coli* and *D. radiodurans*. *D. radiodurans* and *E. coli* RNA polymerases slow 30-fold transitioning from 37°C to 0°C (Kulbachinskiy *et al*, 2004). *E. coli* translational elongation is slowed 3-fold transitioning from 37°C to 25°C (Zhu *et al*, 2016) and > 10-fold from 37°C to 10°C (Farewell & Neidhardt, 1998). *E. coli* protein synthesis slows to a halt after 30 min at 0°C because translation cannot initiate (Broeze *et al*, 1978). In *D. radiodurans*, cold shock without irradiation only influences ~5% of expressed proteins after 3 h at 20°C (Airo *et al*, 2004), and translation occurs only very little at 0°C (Lipton *et al*, 2002). Taken together, proteomic changes observed in our experiments must be due to oxidation-induced damage leading to a combination of aggregation and degradation as opposed to *de novo* protein synthesis.

Small volumes (110 µl) of each irradiated and unirradiated replicate were reserved in 1.5-ml microfuge tubes and held on ice in dark, while protein extraction was initiated (see below). Serial dilutions of each irradiated and unirradiated replicate were made: undiluted, 1:1,000, and 1:10,000 dilutions in LB for *E. coli*; and undiluted, 1:10,000, and 1:100,000 dilutions in TGY for *D. radiodurans*. We chose different dilution factors for each species to facilitate colony counting in anticipation of drastically different irradiation survival rates. One hundred microlitre of each dilution was spread on prewarmed solid LB medium or TGY medium plates for *E. coli* and *D. radiodurans*, respectively. *E. coli* plates were grown overnight at 37°C in dark, and *D. radiodurans* plates were grown for 3 days at 30°C in dark. Colony-forming units (CFUs) were counted at the end of each growth period for each plate, and the irradiation survival rate (Fig EV1A) was computed for each replicate as the ratio of CFUs on plates with irradiated samples to CFUs on plates with unirradiated samples. Due to our irradiation treatment, no *E. coli* colonies formed even from undiluted irradiated cultures, and 55–70% survival of CFUs was observed for *D. radiodurans* (Fig EV1A).

Liquid chromatography–tandem mass spectrometry (LC-MS/MS)

Protein identification and quantification was performed essentially as in (Matallana-Surget *et al*, 2013) using a label-free strategy on a UHPLC-HRMS platform composed of an eksigent 2D liquid chromatograph and an AB SCIEX TripleTOF™ 5600. Peptides were separated on a 25 cm C18 column (Acclaim pepmap100, 3 µm, Dionex) by a linear acetonitrile (ACN) gradient (5–35% (v/v), in 15 or 120 min for short and long runs, respectively) in water containing 0.1% (v/v) formic acid at a flow rate of 300 nl/min. In order to reach high retention stability, which is a requirement for label-free quantification, the column was equilibrated with a 10× volume of 5% ACN before each injection. Eluant was sprayed using the Nanospray Source into the TripleTOF™ 5600. Mass spectra (MS) were acquired across 400–1,500 *m/z* in high resolution mode (resolution > 35,000) with 500 ms accumulation time. The instrument was operated in DDA (data dependent acquisition) mode, and MS/MS were acquired across 100–1,800 *m/z*. For short runs, precursor selection parameters were as follows: intensity threshold 400 cps, 20 precursors maximum per cycle, 100 ms accumulation time, and 10 s exclusion after one spectrum. For long runs, precursor selection parameters were as follows: intensity threshold 200 cps, 50 precursors maximum per cycle, 50 ms accumulation time, and 30 s exclusion after one spectrum. A long run procedure was used to acquire quantitative data, and a duty cycle of 3 s per cycle was used to ensure that high quality extracted ion chromatograms (XIC) could be obtained.

Modified peptide identification

Protein searches for mass spectra obtained on the Triple-TOF 5600 LC-MS/MS were performed against a local copy of the *D. radiodurans* R1 (UP000002524) and *E. coli* K-12 MG1655 (UP000000625) database (retrieved from UniProt on June 26, 2017; 3085 and 4307 proteins, respectively) using ProteinPilot™ 5.0.1 Revision 4895 (Paragon™ Algorithm 5.0.1.0, 4874; Seymour & Hunter, 2017). One missed internal tryptic cleavage site per peptide was accounted for in the search parameters. Mass tolerance was set to 15 ppm in MS and 0.05 Da in MS/MS. In addition to the standard biological modifications set including a differential amino acid mass shift for oxidized methionine (+15.99 Da), custom modifications accounting for DNPH derivatization adducts (Appendix Fig S1), both with and without a typical neutral loss (noNL), were added to the data dictionary and parameter translation files. These custom modifications included:

- Arginine → DNP-glutamic-5-semialdehyde (+136.97 Da)
- Lysine → DNP-allysine (+179.00 Da)
- Proline → DNP-glutamic-5-semialdehyde (DNP-Glu5A) (+196.02 Da)
- Proline → DNP-pyroglutamic acid (DNP-PCA) (+194.01 Da)
- Threonine → DNP-2-amino-3-oxobutanoic acid (DNP-AKB) (+178.01 Da)

The Paragon™ Algorithm was run to identify peptides from each replicate sample with the following parameter settings:

- Sample Type: Identification
- Cys Alkylation: Iodoacetamide
- Digestion: Trypsin
- Instrument: TripleTOF 5600
- Special Factors: Urea denaturation; DNPH derivatization; DNPH derivatization noNL
- Species: *Deinococcus radiodurans* (OR *Escherichia coli*)
- ID Focus: Biological modifications
- Database: Deinococcus_radiodurans.fasta (OR Escherichia_coli.fasta)
- Search Effort: Thorough ID
- Detected Protein Threshold (Unused ProtScore (Conf)) > : 0.05 (10.0%)
- Run False Discovery Rate Analysis: (yes)

High-confidence peptides were selected using a 99% confidence threshold for proteins identified by at least two peptides. After identifying peptides from all sample replicates, carbonyl modifications were taken as those residues exhibiting one of the DNPH modifications described above or any of the following default modifications included in the data dictionary file that represent the direct products of RKPT side chain carbonylation by ROS:

- Arginine → Glutamic 5-semialdehyde (−43.05 Da)
- Lysine → Allysine (−1.03 Da)
- Lysine → 2-aminoadipic acid (double oxidation of lysine) (+14.96 Da)
- Proline → Glutamic 5-semialdehyde (+15.99 Da)
- Proline → Pyroglutamic acid (+13.98 Da)
- Threonine → 2-amino-3-oxobutanoic acid (−2.02 Da)

The false discovery rate (FDR) was calculated at the peptide level for all experimental runs using ProteinPilot; this rate was estimated to be lower than 1% for both *D. radiodurans* and *E. coli*.

Estimation of total CS in samples by saturation curve fitting

Saturation curve fitting was used to estimate the total number of CS in shotgun redox proteomic samples, a technique that has been demonstrated for phospho-proteomics when pooling data across many studies (Vlastaridis *et al*, 2017). To this end, we identified the number of redundant and non-redundant CS within each replicate experiment for each organism and computed the cumulative totals with the addition of each successive replicate (Fig EV1B). Exponential saturation functions were fit by minimization of the sum of squared errors with experimentally determined data points. The fit exponential functions asymptotically approach theoretical maxima, which represent the estimated total CS in our samples. The estimated percentage of sampling coverage is then taken as the ratio of total measured non-redundant CS to total estimated non-redundant CS for each species. Fit curves that remain very linear suggest very low sampling coverage, and curves that saturate suggest near-complete sampling coverage of unique sites. The more replicate experiments available, the higher the confidence in such estimates.

Protein quantification

Protein quantification was performed essentially as in Matallana-Surget *et al* (2013). The quant application of PeakView was used to calculate XIC for all peptides identified with a confidence > 0.99 using ProteinPilot™ (Seymour & Hunter, 2017). A retention time window of 2 min and a mass tolerance of 0.015 *m/z* were used. The area under the curve was exported in MarkerView™, in which they were normalized based on the summed area of the entire run. MarkerView™ enabled an average intensity for irradiated and unirradiated conditions to be calculated, as well as the significance of the difference between conditions based on a paired *t*-test. Quantified proteins were kept with a *P*-value < 0.1 and with at least one peptide quantified with a *P*-value < 0.1. In order to be considered as significantly changed in relative abundance, proteins had to meet two criteria: (i) mean relative abundance fold change within a cutoff of > 2-fold (for increased) or < 0.5-fold (for decreased) in the irradiated samples relative to the controls, and (ii) the paired *t*-test *P*-value had to be < 0.05.

Amino acid prevalence analysis

For RKPT carbonylation prevalence, only RKPT on peptides bearing at ≥1 CS were included. For pre-irradiation prevalence and change in prevalence after irradiation, all amino acid residues in peptides identified as described above were included. Occurrence of each amino acid was counted within the respective set of included peptides, and relative proportion was computed as the percentage of the total number of residues. Statistical significance of differences in relative proportions was determined using a two-tailed z-test for two independent proportions with the following equation:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

In this equation, Z is the z-score test statistic, $\hat{p}$ is the overall sample proportion, $\hat{p}_1$ is the proportion of sample 1, $\hat{p}_2$ is the proportion of sample 2, n_1 is the size of sample 1, and n_2 is the size of sample 2. *P*-values were estimated by *z*-score lookup table. The significance threshold was taken as results with *P*-value < 0.01.

Functional classification analysis of proteomics data

The aGOtool web server (Scholz *et al*, 2015) was used to perform Gene Ontology (GO; The Gene Ontology Consortium, 2017) biological process term enrichment analysis among groups of proteins of interest. These groups included proteins for which carbonyl sites were identified or for which relative abundance changes after irradiation were > 2-fold or < 0.5-fold in *E. coli* or *D. radiodurans*. To correct for protein abundances, the background mean normalized abundance pre-irradiation was used for all detected proteins. A *P*-value cutoff of < 0.05 was used for reporting over- and underrepresented GO terms.

Selection of positive and negative experimentally identified CS for computational analysis

From all high-confidence modified peptide calls, taken as described above, all residues bearing RKPT side chain carbonyls or derivatized labels for carbonyls were assigned as positive CS. Unmodified RKPT on this same set of peptides were counted as negative CS. All positive and negative CS and the peptide sequences with modifications used to assign them are listed in Table EV3.

Protein structure derivation

Target proteins

The target proteins consisted of all genomically encoded *D. radiodurans* R1 proteins. Amino acid sequences for target proteins (3184 proteins) were taken from the translated *D. radiodurans* R1 genome (White *et al*, 1999) as annotated by the RAST server (Aziz *et al*, 2008) as of May 4, 2016, and reconciled against their respective UniProt entries when available (The UniProt Consortium, 2018). An index of all target proteins and statistics for their structures is listed in Table EV4.

Curation of experimentally solved structures

Experimentally solved structures were curated from the PDB (Berman *et al*, 2000), when available, for all *D. radiodurans* target proteins. The best PDB structure from alternative available structures for each protein was chosen based on the best resolution structure with highest coverage of the full-length protein sequence. The chosen structures were edited to remove all ligands and all but the single best representative chain for each protein. This resulted in 71 *D. radiodurans* proteins covered by crystal structures.

Structure modeling

Structures were modeled for all *D. radiodurans* target proteins using five methods, as permitted by the technical constraints of each method. These methods included I-TASSER (Yang *et al*, 2015), QUARK (Xu & Zhang, 2013b), modeler and scwrl algorithms as implemented on the ProtMod server (http://protmod.godziklab.org/protmod-cgi/protModHome.pl; Sali & Blundell, 1993; Canutescu *et al*, 2003), and EVfold (Marks *et al*, 2012). Developer-recommended protein length limits and high-quality homologous template availability, as summarized in Appendix Table S2, determined which methods could be applied for each target protein. For targets modeled by I-TASSER, the LOMETS (Wu & Zhang, 2007) software was used to determine the number of high-quality homologous templates available and categorize each target as easy (>10 templates), medium (1–10 templates), or hard (0 templates) to model. For targets modeled using the ProtMod server, the FFAS-3D (Xu *et al*, 2014) software was used to select the single most appropriate homologous template. The top ranking of alternative models from each method was selected for further analysis as determined by scoring metrics particular to each method. Each method was run using default parameters as noted in their respective publications or as included as default settings for the web server implementations of each method as available.

Other than for training set proteins, which were curated from the PDB and modeled comprehensively as described above, *E. coli* target protein structures were taken as modeled previously (Xu &

Zhang, 2013a) using I-TASSER and QUARK across the whole proteome. Given that either I-TASSER or QUARK was the top performing modeling method for the majority of proteins in the *D. radiodurans* proteome and given the high coverage of the *E. coli* proteome by experimentally solved structures, comprehensive modeling of *E. coli* proteins using modeler, scwrl, and EVfold was seen as unlikely to yield higher quality structures for *E. coli* proteins.

Quality evaluations and selection of best representative structures

The best PDB structure and model from each method were all evaluated for quality with respect to 8 metrics (Appendix Table S3): target sequence coverage, computable surface using UCSF Chimera (Pettersen *et al*, 2004), three separate energy scores (Jaroszewski *et al*, 1998; Yang & Zhou, 2008a,b), estimated TM-score (Yang *et al*, 2016), percentage of residues in favored positions with respect to a Ramachandran plot (Laskowski *et al*, 1993), and an overall conformation score (Laskowski *et al*, 1993). Two criteria were absolutely required to be included in downstream analyses: >90% target sequence coverage and computable surface. This is because high structural coverage for each protein was seen as critical for comprehensive evaluation of possible CS and because several molecular properties relied on protein surface computation. The criteria used for favorable quality evaluation with respect to the other metrics were derived from publications of respective methods as noted in Appendix Table S3. The best representative for each protein was chosen based on which structure satisfied not only the two absolutely required criteria but also the highest total number of satisfied criteria. Ties for total highest number of satisfied criteria were broken by comparing the means of maximum-normalized metrics for the tying structures. Only the best representative structure for each protein was used in subsequent structural analysis and feature computation for machine learning.

Feature engineering

Selection of molecular properties

The published literature on protein oxidation and modification was reviewed in search of protein molecular properties previously hypothesized or tested for contribution to susceptibility or robustness to damage by ROS. Rate constants for reaction with hydroxyl radical were included with respect to each type of amino acid (Buxton *et al*, 1988). As a proxy for reaction rate constants with all ROS that might lead to carbonyl formation on RKPT, the relative reactivity of each of these four amino acids was derived based on the relative proportion of experimentally measured CS distributed across the four amino acids and relative proportion of non-oxidized occurrence distributed across the four amino acids, an approached described previously (Rao & Moller, 2011).

The more exposed a residue is, the higher probability of contact with any molecules in the solvent (Gao *et al*, 2013; Dou *et al*, 2014; Jia *et al*, 2015). Therefore, solvent accessible surface area and depth (i.e., minimum distance from the protein surface) were computed using the surface computation and distance measurement tools in UCSF Chimera (Pettersen *et al*, 2004). Similarly, residues that form secondary structure often have less exposed side chains; the UCSF Chimera implementation of the DSSP algorithm (Kabsch & Sander, 1983) was used to assign secondary structure. Molecular density can also be expected to sterically interfere with small molecule interaction; circular variance (Laine & Carbone, 2015) was computed as a proxy for molecular density. Number of contacts, another measure of density, was computed based on interatomic distances using UCSF Chimera. Protein–protein interactions can also be expected to interfere with small molecule interaction, and DISOPRED (Ward *et al*, 2004) and SPPIDER (Porollo & Meller, 2007) were used to assess residues likely to take part in protein binding sites.

Correlation of dehydration tolerance with both disorder and hydrophobicity has been demonstrated previously (Krisko *et al*, 2010); disorder was computed using DISOPRED (Ward *et al*, 2004) and hydrophobicity using the Kyte-Doolittle scale (Kyte & Doolittle, 1982).

Electrostatic charge is expected to impact protein interaction with charged molecules (Mahalingam *et al*, 2014), such as hydroxyl and superoxide radicals; therefore, canonical formal charges for amino acids were included.

Metal binding sites for Fe and Cu cations can cause residues nearby to undergo metal-catalyzed oxidation (Stadtman & Levine, 2003), but metalloproteins coordinating Mn cations are involved in sequestering ROS (Daly *et al*, 2010). Experimentally characterized metal binding sites were curated from UniProt (The UniProt Consortium, 2018) and mapped to protein structures included in this study, and to expand the very sparse available experimental data, FINDSITE-metal (Brylinski & Skolnick, 2011) was used to predict metal binding sites from structural homology with homologous templates found using LOMETS (Wu & Zhang, 2007).

Surface methionines and cysteines can serve to protect other nearby residues from oxidative damage through their own reversible oxidation (Stadtman & Levine, 2003); these amino acids were included as found on protein surfaces defined using UCSF Chimera.

To further broaden the search space of potentially useful molecular features, a database of hundreds of amino acid properties (Kawashima *et al*, 2008), as summarized by five factors called AAMetrics (Atchley *et al*, 2005), was included in this study. These factors can be conceptually summarized as corresponding to electrostatic charge, propensity to form secondary structure, molecular volume, codon diversity, and polarity.

Features from parameterization of molecular properties

In the conversion of molecular properties to features, several parameters for computation were varied:

Scale: In addition to computing each molecular property with respect to the atomic location of possible carbonyl formation on each RKPT (i.e., residue scale), each property was also computed with respect to all residues within a defined radius (5Å, 8Å, 10Å, and 12Å) and with respect to the entire protein to account for the global scale. The radii used for local-scale feature computation were selected based on effective protein contact distances previously determined in published studies (Huang *et al*, 1995; Kim & Kihara, 2014; Mahalingam *et al*, 2014; Laine & Carbone, 2015). The residue scale is not applicable to surface cysteine and methionines, contacts, RKPT contacts, or negative formal charge because of the type of amino acids under consideration or due to logical exclusions.

Summary statistic: As a summary statistic of each molecular property, the sum and mean of component values were computed. At the residue scale, the sum and mean are equivalent because there is only one component value. The mean summary statistic is not applicable to metal binding site features, contacts, or circular variance.

Grouping: When applicable, molecular properties were converted to features as a composite and as separate features. Applicable properties included surface cysteines and methionines, canonical charge, RKPT contacts, secondary structure, and metal binding sites. For example, separate features were computed with respect to binding sites for each elemental type of metal cation, and also, features were computed with respect to all metal cation binding sites irrespective of elemental type.

Data type: Some molecular properties may be treated as either continuous or binary variables; each option was used to yield separate features with respect to such properties. Applicable properties included protein disorder and protein binding sites as output by DISOPRED, which outputs these properties both as a continuous probability score for each residue and as a binary classification of disordered or not.

Features from sequence alignment

In addition to structure-based features, local amino acid sequence homology was used to generate features to represent positive and negative CS. We defined a local neighborhood of 21 residues centered on each RKPT covered by carbonyl-bearing peptides in our proteomic data. All-by-all semi-global alignments of the non-redundant subset of these sequences was performed, setting the central RKPT of each 21-mer as anchor points to seed each alignment. The resultant alignment score matrix had the same dimensionality as the number of non-redundant RKPT sites in our CS-positive and -negative data (979); therefore, feature reduction was performed during the machine learning phase (see below) by principal component analysis (PCA).

Machine learning

Training data

The training set consisted of 397 molecular property-derived features computed using best representative protein structures and 979 sequence alignment-derived features corresponding to experimentally measured positive and negative CS from *E. coli* and *D. radiodurans* proteomes. This dataset includes 869 non-unique positives and 2777 non-unique negative CS in 153 unique proteins. The full training dataset is included in Table EV3. All features in the training set were standardized by mean centering at 0 and scaling to unit variance using the scikit-learn StandardScaler function. The same scaling factors from the training data were used for standardizing the test data.

Logistic regression algorithm

We used 64-bit Python version 3.6.2 with Scikit-learn version 0.19.1 for all machine learning in this study. To model the probability of RKPT carbonylation, we used a logistic regression estimator with stochastic gradient descent training, a "log" loss function, and elastic net regularization that linearly combines L1 and L2 penalties. The elastic net mixing parameter for L1 and L2 penalties was set to 0.5 and the learning rate to 0.1 to balance utilization of informative features with elimination of extraneous features. Fitting the logistic regression model with these parameters will drop many coefficients of uninformative features to zero. Balanced class weighting was used to account for the class imbalance in training data such that weights for each class i (positive or negative) are inversely proportional to the class size n for data sample size N ($\text{Weight}_i = N/(2 \times n_i)$).

Logistic regression models were fit separately to structure-based and sequence-based features. While the regularization strategy described above was sufficient to reduce the number of features in the structure-based model, the great number of sequence-based features and their relatively low individual predictive power required an additional feature reduction step. To this end, we performed PCA on the sequence alignment score matrix, using just the top principal components collectively accounting for 95% of data variability as features for the sequence-based model.

We combined the sequence-based and structure-based logistic regression models into a final stacked model. Input to the stacked model consists of features with non-zero coefficients from the structure-based model and the PCA-derived features from the sequence-based model along with two meta features representing the individual predicted probabilities from each model. The stacked model was fitted with L2 regularization (but not L1) to attain a stable solution that does not exclude any of the informative features from the initial two models. The magnitude and sign of the fit model coefficients quantify the relative contribution of each feature to overall probability estimation. This implies that molecular properties underlying contributing features are predictive of carbonylation probability but does not necessarily imply that molecular properties represented by features with zero coefficients cannot influence carbonylation by ROS *in vivo*, simply that our data do not support the predictive power of those features.

Leave-1-out validation strategy

To assess generalization of predictive performance of our predictive framework, comprehensive leave-1-out cross-validation was performed. Because the training data contain duplicate data points due to coverage of particular RKPT by multiple measured peptides, we implemented a variant of standard leave-1-out validation in which all duplicates of a data point (positive or negative) are treated as one data unit, i.e., validation is performed on them together in one leave-1-out iteration. Also, because our full set of sequence-based features contains information derived from local sequence surrounding every RKPT in the training data, we excluded sequence features corresponding to alignment scores against the held-out data for each leave-1-out iteration before performing PCA as described above.

Multi-scale model validation

Residue-scale validation

We used receiver operating characteristic (ROC) curve analysis to validate the overall performance of our machine learning framework for protein carbonylation site probability estimation. Area under the ROC curve (AUC) is a common, robust performance metric for machine learning predictors and is not affected by data imbalance, which is especially important for predicting protein CS because redox proteomic techniques yield a relatively much smaller number of positives than negatives in peptide calls. The calculation of the area under the ROC curve was weighted by class sizes using the same weighting factor described above.

Furthermore, because of the stochastic carbonylation of RKPT observed in the training data (i.e., the same site can appear carbonylated on one peptide but not oxidized on another), an AUC of 1.0 (a perfect predictor score) is not guaranteed to be achievable. Therefore, we computed the theoretical maximum ROC curve and corresponding AUC for our data (Fig 5A lower left) using empirically derived carbonylation probabilities. All reported AUCs were normalized to this theoretical maximum AUC.

We also performed a randomization test in which the initial structure-based and sequence-based features were first shuffled before running the entire predictive framework. This randomization test gives us a sense of the non-random significance of predictive features in our structure-based, sequence-based, and stacked models. We also compared our model performance to that of the Carbonylated Site and Protein Detection (CSPD) model (Maisonneuve *et al*, 2009). To test CSPD performance on our data, we used the CSPD web server and input all full-length protein sequences from our experimental dataset. CSPD output is a binary classification of carbonylation sites, rather than a probability estimator. ROC analysis was performed on these predictions as well.

Protein-scale validation by carbonyl site enrichment correlation

To validate our protein carbonyl prediction framework at the whole-protein scale, we computed a predicted enrichment score for each protein as the sum of the predicted probabilities for sites included in the carbonylated peptide data normalized by the length of the protein regions covered by those peptides. Similarly, an enrichment score was computed based solely on the carbonylation state of RKPT residues in the experimentally measured carbonylated peptides. These scores appear in the *x*- and *y*-axes of Fig 5B. Predicted carbonylation enrichment was validated against all 90 *E. coli* and 63 *D. radiodurans* proteins detected with at least one CS by redox proteomics and coverage by our 3D protein structures. Spearman rank correlation between predicted and experimental enrichment scores was computed for proteins from each species, and 95% confidence intervals were determined for a fitted linear regression (Fig 5B). *P*-values for rank correlation were determined by permutation test as implemented in the *corr* function in Matlab®.

Interspecies proteome-wide prediction

Datasets for prediction

After using our framework to train a model on the entire training set, we applied our predictor prospectively to compare protein carbonylation propensity across the full proteomes of *E. coli* (4,057 proteins) and *D. radiodurans* (3,031 proteins), containing 227,326 and 184,149 total RKPT, respectively. The full proteome datasets for *E. coli* and *D. radiodurans* are included in Tables EV5 and EV6, respectively.

Ortholog and isozyme mapping

In order to directly compare individual proteins between *E. coli* and *D. radiodurans*, pairs of proteins were mapped by shared function between these species. This was done in three steps. First, functional annotation mapping between species was performed using the RAST server (Overbeek *et al*, 2014) for annotation and the ModelSEED server (Henry *et al*, 2010) for mapping. This approach has the ability to map orthologs between species as well as non-homologous isozymes that only share function. Second, likely orthologous pairs were mapped between species by bi-directional BLAST (Altschul *et al*, 1990) in which mutual top hits between species are identified, and for cases where proteins do not have mutual top hits by bi-directional BLAST, simply the top hit in the other species is identified. Tying top hits with equal E-values and alignment scores were retained at this step. Finally, manual curation of the mapped interspecies pairs was performed, reconciling functional annotation of *E. coli* proteins in EcoCyc (Keseler *et al*, 2017) and *D. radiodurans* protein annotations in BioCyc (Caspi *et al*, 2016). During the curation process, proteins from one species mapping to multiple proteins in the other species were reduced to mapping to as few proteins from the other species as possible based on a combination of functional annotation specificity and any fine differences in sequence alignments. As a result, 1,170 *D. radiodurans* proteins were mapped to 1,110 *E. coli* proteins for a total of 1,300 interspecies pairs that also have structural coverage in our data (Appendix Fig S3).

Pairwise comparison of predicted vulnerability to oxidative damage

To compare protein-scale predicted carbonylation susceptibility between proteins from each species, we computed an enrichment score for each protein equal to the number of RKPT with probability >0.5 for carbonylation normalized by the protein length. These scores appear in the *x*- and *y*-axes of Fig 7.

To quantify the relative protein-intrinsic vulnerability to carbonylation between interspecies pairs, we take the perpendicular distance of the point representing the pair to the y = x diagonal reference line. Proteins less intrinsically vulnerable to carbonylation in one organism lie a greater distance to one side of y = x. To quantify the combined protein-intrinsic and extrinsic differences between interspecies pairs, we take the perpendicular distance of the point representing the pair to the y = x/3.78 reference line. The 3.78 factor represents the ratio of protein carbonyls generated *in vivo* in *E. coli* compared to *D. radiodurans* after exposure to 7 kGy γ-radiation (ratio = 3.86), minus the contribution from protein-intrinsic factors alone. This latter contribution from protein-intrinsic factors comes from the ratio of protein carbonyls measured in *E. coli* lysate compared to *D. radiodurans* lysate after dialysis to remove ions (ratio = 1.08). Proteins less susceptible in *D. radiodurans* than in *E. coli* when accounting for intrinsic and extrinsic factors lie a greater distance above y=x/3.78.

Extreme outliers of the interspecies comparison are expected to be involved in cellular resistance to oxidative stress, if indeed *D. radiodurans* proteins have evolved intrinsic properties to protect them from carbonylation. We define a boundary to highlight extreme outliers of the all-pair distribution based on a minimum-fitted ellipse that encompasses all points within 3 standard deviations of the mean coordinates and within 3 standard deviation of the mean distance from y = x. Outliers to this boundary account for 8% of all interspecies mapped protein pairs.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) and PRIDE via the iProX partner repository (Ma *et al*, 2019) with the dataset identifier PXD020058 (http://www.ebi.ac.uk/pride/archive/projects/PXD020058). Code for feature computation from structures is available in GitHub (https://github.com/julianstanley/ProteinFeatures.git). Best representative structures and models for *D. radiodurans* proteins are available from the authors upon request. All other data are available in the main text or the supplementary materials.

Expanded View for this article is available online.

Acknowledgements
Roger Chang especially thanks Dr. Pamela Silver (Harvard) who as his Postdoctoral Advisor allowed and supported this unique, independently conceived effort. As such, Dr. Silver provided laboratory resources, feedback on the work, comments on the manuscript and full support of this independent publication. We also thank Dr. Debora Marks (Harvard) for advice on machine learning, data analysis and feedback on the manuscript; Dr. James Collins (MIT) for facilitating γ-irradiator access; Dr. James Weaver (Wyss Institute) for fabricating the irradiator rack; and Dr. Fatma Elzahraa Eid (Broad Institute) for proofreading the manuscript. Computational modeling was enabled by the HMS O2 high performance compute cluster. Funding: This work was funded by the Gordon and Betty Moore Foundation GBMF 2550.04 Life Sciences Research Foundation post-doctoral fellowship and the Wyss Institute. The UMons proteomic facility was supported by the Belgian National Funds for Scientific Research (FNRS).

Author contributions
RLC and SM-S conceived, supervised, developed experimental and analytical methods for, performed experiments, and acquired funding for this study. RLC, JAS, MCR, JWS, and ARO developed software and performed computations. RLC, SM-S and RW performed mass spectrometry analysis. RLC, MCR, and SM-S performed statistical analysis and validations. ZL, YAC, RW, AG, and SM-S provided study materials and computing resources. RLC, MCR, JWS, ZL, ARO, and SM-S performed data curation. RLC, JAS, MCR, JWS, ZL, YAC, ARO, and SM-S wrote the manuscript.

Conflict of interest
The authors declare that they have no conflict of interest.

References

Airo A, Chan SL, Martinez Z, Platt MO, Trent JD (2004) Heat shock and cold shock in *Deinococcus radiodurans*. *Cell Biochem Biophys* 40: 277 – 288

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403 – 410

Anaganti N, Basu B, Apte SK (2016) *In situ* real-time evaluation of radiation-responsive promoters in the extremely radioresistant microbe *Deinococcus radiodurans*. *J Biosci* 41: 193 – 203

Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci USA* 102: 6395 – 6400

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M *et al* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genom* 9: 75

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2: 2006 0008

Ballesteros M, Fredriksson A, Henriksson J, Nystrom T (2001) Bacterial senescence: protein oxidation in non-proliferating cells is dictated by the accuracy of the ribosomes. *EMBO J* 20: 5280 – 5289

Basak S, Jiang R (2012) Enhancing *E. coli* tolerance towards oxidative stress via engineering its global regulator cAMP receptor protein (CRP). *PLoS ONE* 7: e51179

Basu B, Apte SK (2012) Gamma radiation-induced proteome of *Deinococcus radiodurans* primarily targets DNA repair and oxidative stress alleviation. *Mol Cell Proteomics* 11: M111 011734

Belenky P, Ye JD, Porter CB, Cohen NR, Lobritz MA, Ferrante T, Jain S, Korry BJ, Schwarz EG, Walker GC *et al* (2015) Bactericidal antibiotics induce toxic metabolic perturbations that lead to cellular damage. *Cell Rep* 13: 968 – 980

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28: 235 – 242

Bolt EL, Jenkins T, Russo VM, Ahmed S, Cavey J, Cass SD (2015) Identification of *Escherichia coli* ygaQ and rpmG as novel mitomycin C resistance factors implicated in DNA repair. *Biosci Rep* 36: e00290

Brbic M, Warnecke T, Krisko A, Supek F (2015) Global shifts in genome and proteome composition are very tightly coupled. *Genome Biol Evol* 7: 1519 – 1532

Broeze RJ, Solomon CJ, Pope DH (1978) Effects of low temperature on *in vivo* and *in vitro* protein synthesis in *Escherichia coli* and *Pseudomonas fluorescens*. *J Bacteriol* 134: 861 – 874

Bruckbauer ST, Minkoff BB, Yu D, Cryns VL, Cox MM, Sussman MR (2020) Ionizing radiation-induced proteomic oxidation in *Escherichia coli*. *Mol Cell Proteomics* 19: 1375 – 1395

Brylinski M, Skolnick J (2011) FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins* 79: 735 – 751

Bubunenko M, Baker T, Court DL (2007) Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in *Escherichia coli*. *J Bacteriol* 189: 2844 – 2853

Buxton GV, Greenstock CL, Helman WP, Ross AB (1988) Critical Review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals (·OH/·O− in Aqueous Solution. *J Phys Chem Ref Data* 17: 513 – 886

Cabiscol E, Tamarit J, Ros J (2000) Oxidative stress in bacteria and protein damage by reactive oxygen species. *Int Microbiol* 3: 3 – 8

Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12: 2001 – 2014

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA *et al* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 44: D471 – D480

Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BO (2013a) Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. *Science* 340: 1220 – 1223

Chang RL, Xie L, Bourne PE, Palsson BO (2013b) Antibacterial mechanisms identified through structural systems pharmacology. *BMC Syst Biol* 7: 102

Chin WC, Lin KH, Liu CC, Tsuge K, Huang CC (2017) Improved n-butanol production via co-expression of membrane-targeted tilapia metallothionein and the clostridial metabolic pathway in *Escherichia coli*. *BMC Biotechnol* 17: 36

Daly MJ, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, Venkateswaran A, Hess M, Omelchenko MV, Kostandarithes HM, Makarova KS *et al* (2004) Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates gamma-radiation resistance. *Science* 306: 1025 – 1028

Daly MJ (2006) Modulating radiation resistance: Insights based on defenses against reactive oxygen species in the radioresistant bacterium *Deinococcus radiodurans*. *Clin Lab Med* 26: 491 – 504, x

Daly MJ, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, Leapman RD, Lai B, Ravel B, Li SM, Kemner KM *et al* (2007) Protein oxidation implicated as the primary determinant of bacterial radioresistance. *PLoS Biol* 5: e92

Daly MJ (2009) A new perspective on radiation resistance based on *Deinococcus radiodurans*. *Nat Rev Microbiol* 7: 237 – 245

Daly MJ, Gaidamakova EK, Matrosova VY, Kiang JG, Fukumoto R, Lee DY, Wehr NB, Viteri GA, Berlett BS, Levine RL (2010) Small-molecule antioxidant proteome-shields in *Deinococcus radiodurans*. *PLoS ONE* 5: e12570

Daly MJ (2012) Death by protein damage in irradiated cells. *DNA Repair (Amst)* 11: 12 – 21

Dou Y, Yao B, Zhang C (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* 46: 1459 – 1469

Dukan S, Nystrom T (1998) Bacterial senescence: stasis results in increased and differential oxidation of cytoplasmic proteins leading to developmental induction of the heat shock regulon. *Genes Dev* 12: 3431 – 3441

Dulermo R, Onodera T, Coste G, Passot F, Dutertre M, Porteron M, Confalonieri F, Sommer S, Pasternak C (2015) Identification of new genes contributing to the extreme radioresistance of *Deinococcus radiodurans* using a Tn5-based transposon mutant library. *PLoS ONE* 10: e0124358

Farewell A, Neidhardt FC (1998) Effect of temperature on *in vivo* protein synthetic capacity in *Escherichia coli*. *J Bacteriol* 180: 4704 – 4710

Fredriksson A, Ballesteros M, Dukan S, Nystrom T (2005) Defense against protein carbonylation by DnaK/DnaJ and proteases of the heat shock regulon. *J Bacteriol* 187: 4207 – 4213

Ganini D, Leinisch F, Kumar A, Jiang J, Tokar EJ, Malone CC, Petrovich RM, Mason RP (2017) Fluorescent proteins such as eGFP lead to catalytic oxidative stress in cells. *Redox Biol* 12: 462 – 468

Gao YF, Li BQ, Cai YD, Feng KY, Li ZD, Jiang Y (2013) Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. *Mol BioSyst* 9: 61 – 69

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 28: 977 – 982

Hohn A, Weber D, Jung T, Ott C, Hugo M, Kochlik B, Kehm R, Konig J, Grune T, Castro JP (2017) Happily (n)ever after: aging in the context of oxidative stress, proteostasis loss and cellular senescence. *Redox Biol* 11: 482 – 501

Huang ES, Subbiah S, Levitt M (1995) Recognizing native folds by the arrangement of hydrophobic and polar residues. *J Mol Biol* 252: 709 – 720

Jaroszewski L, Pawlowski K, Godzik A (1998) Multiple model approach: exploring the limits of comparative modeling. *Mol Model Annual* 4: 294 – 309

Jia L, Yarlagadda R, Reed CC (2015) Structure based thermostability prediction models for protein single point mutations with machine learning tools. *PLoS ONE* 10: e0138022

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

Kaleta C, Schauble S, Rinas U, Schuster S (2013) Metabolic costs of amino acid and protein production in *Escherichia coli*. *Biotechnol J* 8: 1105 – 1114

Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 110: 15674 – 15679

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36: D202 – D205

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martinez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M *et al* (2017) The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res* 45: D543 – D550

Kim H, Kihara D (2014) Detecting local residue environment similarity for recognizing near-native structure models. *Proteins* 82: 3255 – 3272

Krisko A, Radman M (2010) Protein damage and death by radiation in *Escherichia coli* and *Deinococcus radiodurans*. *Proc Natl Acad Sci USA* 107: 14373 – 14377

Krisko A, Smole Z, Debret G, Nikolic N, Radman M (2010) Unstructured hydrophilic sequences in prokaryotic proteomes correlate with dehydration tolerance and host association. *J Mol Biol* 402: 775 – 782

Krisko A, Copic T, Gabaldon T, Lehner B, Supek F (2014) Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol* 15: R44

Krisko A, Radman M (2019) Protein damage, ageing and age-related diseases. *Open Biol* 9: 180249

Kulbachinskiy A, Bass I, Bogdanova E, Goldfarb A, Nikiforov V (2004) Cold sensitivity of thermophilic and mesophilic RNA polymerases. *J Bacteriol* 186: 7818 – 7820

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157: 105 – 132

Laine E, Carbone A (2015) Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLoS Comput Biol* 11: e1004580

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26: 283 – 291

Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, Daly MJ, Fredrickson J, Hixson KK, Kostandarithes H *et al* (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci USA* 99: 11049 – 11054

Liu Y, Zhou J, Omelchenko MV, Beliaev AS, Venkateswaran A, Stair J, Wu L, Thompson DK, Xu D, Rogozin IB *et al* (2003) Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc Natl Acad Sci USA* 100: 4191 – 4196

Lott BB, Wang Y, Nakazato T (2013) A comparative study of ribosomal proteins: linkage between amino acid distribution and ribosomal assembly. *BMC Biophys* 6: 13

Luan H, Meng N, Fu J, Chen X, Xu X, Feng Q, Jiang H, Dai J, Yuan X, Lu Y *et al* (2014) Genome-wide transcriptome and antioxidant analyses on gamma-irradiated phases of deinococcus radiodurans R1. *PLoS ONE* 9: e85649

Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D (2014) CarSPred: a computational tool for predicting carbonylation sites of human proteins. *PLoS ONE* 9: e111478

Ma J, Chen T, Wu S, Yang C, Bai M, Shu K, Li K, Zhang G, Jin Z, He F *et al* (2019) iProX: an integrated proteome resource. *Nucleic Acids Res* 47: D1211 – D1217

Mahalingam R, Peng HP, Yang AS (2014) Prediction of fatty acid-binding residues on protein surfaces with three-dimensional probability distributions of interacting atoms. *Biophys Chem* 192: 10 – 19

Maisonneuve E, Fraysse L, Lignon S, Capron L, Dukan S (2008) Carbonylated proteins are detectable only in a degradation-resistant aggregate state in *Escherichia coli*. *J Bacteriol* 190: 6609 – 6614

Maisonneuve E, Ducret A, Khoueiry P, Lignon S, Longhi S, Talla E, Dukan S (2009) Rules governing selective protein carbonylation. *PLoS ONE* 4: e7269

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30: 1072 – 1080

Matallana-Surget S, Cavicchioli R, Fauconnier C, Wattiez R, Leroy B, Joux F, Raftery MJ, Lebaron P (2013) Shotgun redox proteomics: identification and quantitation of carbonylated proteins in the UVB-resistant marine bacterium, *Photobacterium angustum* S14. *PLoS ONE* 8: e68112

Miskinyte M, Gordo I (2013) Increased survival of antibiotic-resistant *Escherichia coli* inside macrophages. *Antimicrob Agents Chemother* 57: 189 – 195

Moosmann B, Behl C (2000) Cytoprotective antioxidant function of tyrosine and tryptophan residues in transmembrane proteins. *Eur J Biochem* 267: 5687 – 5692

Nakayashiki T, Mori H (2013) Genome-wide screening with hydroxyurea reveals a link between nonessential ribosomal proteins and reactive oxygen species production. *J Bacteriol* 195: 1226 – 1235

Nystrom T (2005) Role of oxidative carbonylation in protein quality control and senescence. *EMBO J* 24: 1311 – 1317

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M *et al* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42: D206 – D214

Pallares I, Ventura S (2016) Understanding and predicting protein misfolding and aggregation: Insights from proteomics. *Proteomics* 16: 2570 – 2581

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605 – 1612

Porollo A, Meller J (2007) Prediction-based fingerprints of protein-protein interactions. *Proteins* 66: 630 – 645

Rao RS, Moller IM (2011) Pattern of occurrence and occupancy of carbonylation sites in proteins. *Proteomics* 11: 4166 – 4173

Ruenwai R, Neiss A, Laoteng K, Vongsangnak W, Dalfard AB, Cheevadhanarak S, Petranovic D, Nielsen J (2011) Heterologous production of polyunsaturated fatty acids in *Saccharomyces cerevisiae* causes a global transcriptional response resulting in reduced proteasomal activity and increased oxidative stress. *Biotechnol J* 6: 343 – 356

Sakr S, Cirinesi AM, Ullers RS, Schwager F, Georgopoulos C, Genevaux P (2010) Lon protease quality control of presecretory proteins in *Escherichia coli* and its dependence on the SecB and DnaJ (Hsp40) chaperones. *J Biol Chem* 285: 23506 – 23514

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234: 779 – 815

Sargentini NJ, Gularte NP, Hudman DA (2016) Screen for genes involved in radiation survival of *Escherichia coli* and construction of a reference database. *Mutat Res* 793–794: 1 – 14

Scholz C, Lyon D, Refsgaard JC, Jensen LJ, Choudhary C, Weinert BT (2015) Avoiding abundance bias in the functional annotation of post-translationally modified proteins. *Nat Methods* 12: 1003 – 1004

Seymour SL, Hunter C (2017) ProteinPilot™ Report for ProteinPilot™ Software - Detailed Analysis of Protein Identification/Quantitation Results Automatically. SCIEX technical note

Slade D, Radman M (2011) Oxidative stress resistance in *Deinococcus radiodurans*. *Microbiol Mol Biol Rev* 75: 133 – 191

Stadtman ER (1986) Oxidation of proteins by mixed-function oxidation systems: implication in protein turnover, ageing and neutrophil function. *Trends Biochem Sci* 11: 11 – 12

Stadtman ER, Levine RL (2003) Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. *Amino Acids* 25: 207 – 218

Sun XM, Ren LJ, Zhao QY, Ji XJ, Huang H (2018) Microalgae for the production of lipid and carotenoids: a review with focus on stress regulation and adaptation. *Biotechnol Biofuels* 11: 272

The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 45: D331 – D338

The UniProt Consortium (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46: 2699

Tian B, Sun Z, Shen S, Wang H, Jiao J, Wang L, Hu Y, Hua Y (2009) Effects of carotenoids from *Deinococcus radiodurans* on protein oxidation. *Lett Appl Microbiol* 49: 689 – 694

Uchida K, Kato Y, Kawakishi S (1990) A novel mechanism for oxidative cleavage of prolyl peptides induced by the hydroxyl radical. *Biochem Biophys Res Commun* 169: 265 – 271

Vidovic A, Supek F, Nikolic A, Krisko A (2014) Signatures of conformational stability and oxidation resistance in proteomes of pathogenic bacteria. *Cell Rep* 7: 1393 – 1400

Vlastaridis P, Kyriakidou P, Chaliotis A, Van de Peer Y, Oliver SG, Amoutzias GD (2017) Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *Gigascience* 6: 1 – 11

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20: 2138 – 2139

Weng SL, Huang KY, Kaunang FJ, Huang CH, Kao HJ, Chang TH, Wang HY, Lu JJ, Lee TY (2017) Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features. *BMC Bioinformatics* 18: 66

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL *et al* (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* 286: 1571 – 1577

Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* 35: 3375 – 3382

Xu D, Zhang Y (2013a) *Ab Initio* structure prediction for *Escherichia coli*: towards genome-wide protein structure modeling and fold assignment. *Sci Rep* 3: 1895

Xu D, Zhang Y (2013b) Toward optimal fragment generations for *ab initio* protein structure assembly. *Proteins* 81: 229 – 239

Xu D, Jaroszewski L, Li Z, Godzik A (2014) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* 30: 660 – 667

Yamagishi A, Kawaguchi Y, Hashimoto H, Yano H, Imai E, Kodaira S, Uchihori Y, Nakagawa K (2018) Environmental data and survival data of *Deinococcus aetherius* from the exposure facility of the Japan experimental module of the International Space Station obtained by the Tanpopo mission. *Astrobiology* 18: 1369 – 1374

Yampolsky LY, Stoltzfus A (2005) The exchangeability of amino acids in proteins. *Genetics* 170: 1459 – 1472

Yang Y, Zhou Y (2008a) *Ab initio* folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci* 17: 1212 – 1219

Yang Y, Zhou Y (2008b) Specific interactions for *ab initio* folding of protein terminal regions with secondary structures. *Proteins* 72: 793 – 803

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 12: 7 – 8

Yang J, Wang Y, Zhang Y (2016) ResQ: an approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *J Mol Biol* 428: 693 – 701

Zhou Y, Shen P, Lan Q, Deng C, Zhang Y, Li Y, Wei W, Wang Y, Su N, He F *et al* (2017) High-coverage proteomics reveals methionine auxotrophy in *Deinococcus radiodurans*. *Proteomics* 17: 13 – 14

Zhu M, Dai X, Wang YP (2016) Real time determination of bacterial *in vivo* ribosome translation elongation speed based on LacZalpha complementation system. *Nucleic Acids Res* 44: e155