In Silico Network Perturbation Reveals Hierarchical Roles of DNA Repair and Glycosylation Linking Exercise to Human Ageing Clocks

Ciara G. Juan[1], Lazaros Ntasis[2]

1. Faculty of Life and Health Sciences, Ulster University, Belfast, BT15 1ED
2. Department of Sport Organisation and Management, University of Peloponnese, Sparta, Greece 231 00

Abstract

Regular physical activity is associated with delayed biological ageing, yet the molecular hierarchy linking habitual exercise to long-term ageing signatures in humans remains poorly resolved. Acute exercise activates genome maintenance and stress response pathways, whereas biological ageing clocks capture stable, system-level molecular states that integrate cumulative change over time. How these processes are organised within molecular networks—and which exercise-responsive pathways predictively shape ageing-related molecular architecture—remain unclear. We introduce the first *in silico* perturbation framework applied to human exercise biology, integrating genetically anchored gene selection with predictive network modelling and human experimental validation. Genes causally associated with habitual vigorous physical activity (VPA) were prioritised using Mendelian randomisation (MR) across proteomic, epigenomic, glycomic, and single-cell transcriptomic layers, and represented using self-supervised graph learning. Targeted *in silico* perturbations were propagated using network diffusion to forecast downstream alignment with ageing clocks (Horvath, Hannum, PhenoAge, DunedinPACE, Proteomic Clock). Perturbation of glycosylation enzymes yielded diffusion neighbourhoods significantly enriched for clock-associated genes (empirical $p < 0.01$), whereas DNA repair and stress response genes—despite robust causal prioritisation and acute experimental activation—did not consistently align with ageing clock architecture. Acute exercise validation demonstrated rapid modulation of plasma glycosylation alongside activation of DNA repair programs, providing biological context for these distinct network behaviours. Together, these findings reveal a hierarchical organisation in which DNA repair pathways support acute physiological adaptation, while glycosylation networks predictively encode the stable molecular states captured by human ageing clocks, linking habitual VPA to long-term biological ageing trajectories.

Keywords: In silico perturbation, exercise, ageing, ageing clocks, biological age, DNA repair, glycosylation, human study

Introduction

Regular physical activity is one of the most robust non-pharmacological interventions associated with delayed biological ageing and reduced risk of age-related disease. At the molecular level, exercise engages conserved stress-response programmes—including redox signalling, DNA damage sensing, and metabolic reprogramming—that promote cellular resilience and preserve genome integrity. In parallel, biological ageing clocks derived from epigenetic, proteomic, and glycomic data capture stable system-level molecular states that track chronological age and predict morbidity and mortality. Despite strong evidence linking exercise to both acute molecular responses and long-term ageing trajectories, the organisational hierarchy connecting transient exercise-induced stress responses to durable molecular ageing architecture in humans remains poorly defined.

DNA repair pathways occupy a central position in the immediate cellular response to exercise-induced physiological stress. Vigorous physical activity (VPA) transiently elevates reactive oxygen species (ROS) through increased mitochondrial flux, mechanical strain, and immune activation. These short-lived redox signals act as physiological cues that engage DNA damage sensing and repair pathways, including Poly(ADP-ribose) polymerase (PARP) and sirtuin-dependent mechanisms that coordinate genome maintenance with metabolic adaptation. Such responses are rapid, lesion-specific, and reversible, enabling repeated exposure to physiological stress without cumulative genomic instability (Radak et al., 2011; Williamson et al., 2020).

In contrast, protein glycosylation represents a fundamentally different layer of molecular regulation. Glycosylation is a ubiquitous post-translational modification that governs protein folding, stability, trafficking, immune recognition, and inflammatory signalling. Unlike acute stress response pathways, glycan structures—particularly N-glycans on circulating plasma proteins and immunoglobulin G (IgG)—encode integrative, time-averaged information reflecting chronic immune activation, metabolic state, and inflammatory tone. Extensive glycomic evidence demonstrates that specific IgG N-glycan features, including galactosylation, sialylation, fucosylation, and bisecting GlcNAc, are tightly linked to inflammaging, cardiovascular disease, autoimmunity, and neurodegeneration, and shift predictably with biological ageing (Onigbinde et al., 2025). Mechanistically, glycosylation integrates upstream metabolic, inflammatory, and redox inputs via ROS-sensitive regulation of glycosyltransferase activity, Golgi and endoplasmic reticulum redox organisation, nucleotide sugar availability and transport, and cytokine-driven signalling pathways, enabling dynamic remodelling of glycan structures in response to oxidative and inflammatory stress (Kellokumpu, 2019; Khoder-Agha & Kietzmann, 2021). As a result, glycan patterns change slowly and persistently, in contrast to the rapid but transient activation of DNA repair pathways (Trzos et al., 2023). This distinction positions glycosylation as a plausible molecular interface through which repeated physiological stress—such as habitual exercise—may be translated into stable ageing-related molecular states, rather than acting as an acute stress sensor itself.

Genetic approaches provide an entry point into resolving this hierarchy. Mendelian randomisation (MR) enables causal inference between habitual physical activity and molecular traits, allowing prioritisation of exercise-associated genes beyond observational correlation. However, most MR studies of exercise have focused on isolated genes or traits and have not resolved how exercise-responsive molecular features are organised within interacting networks. We recently addressed this limitation by integrating MR with supervised graph neural networks to perform the first network-level prioritisation of exercise-associated genes across multiple molecular layers in humans (Juan & Ntasis, 2026). While this framework provided causal and mechanistic ranking, it remained inherently descriptive, identifying which genes are important without testing how perturbations propagate through molecular networks or whether such genes organise ageing-related molecular architecture.

In this study, we focus on two mechanistically distinct but interconnected systems: DNA repair networks and glycosylation pathways. Transient exercise-induced DNA lesions activate DNA damage response and repair pathways that are efficiently resolved, promoting adaptive remodelling while preventing the accumulation of persistent genomic damage over repeated training (Zhu et al., 2026). Glycosylation pathways, by contrast, regulate immune signalling, inflammation, and protein-protein interactions, and encode stable molecular features captured by biological ageing clocks. Importantly, plasma and IgG glycosylation represents one of the most scalable and cost-effective omic layers for human validation, enabling direct experimental grounding of network-level predictions in intervention studies.

To move beyond descriptive prioritisation, we apply network-based *in silico* perturbation. By simulating gene-level perturbations and propagating their effects through learned molecular interaction networks, this framework enables prediction of how exercise-associated genes differentially reorganise ageing-related molecular states. We quantify perturbation effects by measuring redistribution of network diffusion toward epigenetic ageing clock regions and proteomic ageing axes, providing a direct test of whether exercise-responsive pathways are positioned to shape stable ageing architecture. Together, this framework integrates causal inference, self-supervised graph representation learning, and *in silico* perturbation to dissect how exercise-responsive molecular programmes differentially contribute to acute stress buffering versus long-term biological ageing in humans.

Results

Causal multi-omic prioritisation of exercise-associated genes

We first leveraged genetically proxied habitual vigorous physical activity (VPA) to prioritise exercise-responsive molecular features using linkage disequilibrium (LD)-aware, overlap-aware Mendelian randomisation (MR) across four orthogonal omic layers: plasma proteomics, whole blood DNA methylation, plasma glycomics, and blood single-cell transcriptomics. This analysis builds on our previously described supervised MR-Graph Attention Network (GAT) framework for exercise gene prioritisation, in which graph neural networks were explicitly

trained to predict MR-derived targets, but is summarised here to provide context for subsequent label-free *in silico* perturbation analyses (Juan et al., 2026, preprint).
Across molecular layers, MR signals were dense but unevenly distributed. In the proteomic layer, 2,304 out of 2,991 VPA single nucleotide polymorphism (SNP)-matched proteins (77.0%) demonstrated nominal evidence of causal association ($p < 0.05$). Similarly strong signal density was observed for epigenomic and glycomic traits, with 151 of 184 CpGs (82.1%) and 101 of 117 plasma glycan traits (86.3%) showing evidence of causal association with habitual VPA. In contrast, MR signals in the single-cell transcriptomic layer were more cell-type specific, with 119 of 364 gene–cell type tests (32.7%) meeting nominal significance, most prominently in B cells, dendritic cells, and $CD8^+$ T cells.

To permit unified network analysis, all MR signals were harmonised to gene symbols, yielding 2,370 unique genes with causal support from at least one molecular layer. Of these, only 21 genes showed convergent evidence across two layers and a single gene across three layers, indicating limited cross-layer redundancy and motivating a network-level rather than gene-centric interrogation.

For downstream perturbation analyses, these genes were embedded within a human multi-omic interaction network and represented using self-supervised contrastive graph representation learning, followed by MR-guided refinement of the learned embeddings. Importantly, no ageing-related targets or clock annotations were used during representation learning or refinement, ensuring that subsequent perturbation analyses were not biased toward ageing clock definitions.

In silico perturbation reveals hierarchical network roles of exercise-responsive genes

We next performed targeted *in silico* perturbation of MR-prioritised genes by propagating localised perturbations through the gene network using personalised PageRank diffusion. This diffusion-based perturbation framework is unsupervised with respect to ageing outcomes, relying solely on network topology and learned gene representations. For each perturbed gene, downstream diffusion neighbourhoods were quantified and evaluated for enrichment of genes defining established biological ageing clocks—Horvath, Hannum, PhenoAge, DunedinPACE, and Proteomic Clock—using degree-matched null models.

Perturbation of seven experimentally validated glycosylation enzymes produced diffusion neighbourhoods significantly enriched for ageing clock-associated genes across multiple epigenetic clocks (empirical $p < 0.01$). These effects were robust to weighting by differential methylation magnitude and showed disproportionate alignment with the proteomic ageing axis, indicating that glycosylation networks occupy positions closely coupled to stable ageing signatures.

In contrast, perturbation of DNA repair and stress response genes—including those robustly prioritised by supervised MR-GAT modelling and experimentally validated following acute exercise in humans—did not yield consistent enrichment for ageing clock architecture ($p > 0.1$

across clocks). Diffusion from these genes was broader and less clock-specific, consistent with upstream roles in buffering transient physiological stress rather than encoding long-term molecular state.

Experimental validation of network-predicted exercise responses in humans

To anchor *in silico* perturbation findings to biological responses in vivo, we examined acute molecular adaptations to vigorous exercise in a controlled human intervention study. Participants completed a high-intensity exercise bout (average ≈80% VO_2max), with paired pre and post-exercise sampling enabling assessment of plasma N-glycosylation traits and targeted expression of canonical genome maintenance and DNA repair genes. Experimental procedures and analytical workflows have been reported previously (Juan et al., 2025, preprint). Importantly, the glycomic platform quantified the same plasma N-glycan traits represented in the multi-omic network, permitting direct comparison between experimental responses and network-level predictions.

Acute exercise elicited a coordinated plasma glycomic response characterised by reduced agalactosylation (G0) alongside increased mono and di-galactosylation (G1, G2), elevated total galactosylation (Gtotal), and increased total sialylation (Stotal) (all $p < 0.05$). In contrast, bisecting GlcNAc abundance (Btotal) did not change acutely post-exercise. These patterns reflect selective modulation of terminal glycan features rather than total remodelling of glycan architecture.

At the network level, these acute glycomic shifts aligned with pathways prioritised by MR and graph-based modelling. Among the highest-ranked exercise-associated genes identified by the MR-GAT, multiple enzymes governing galactosylation and sialylation were consistently represented, including *B4GALT1, ST6GAL1, ST3GAL1, ST3GAL4, MGAT3, MGAT5*, and *MGAT5B*.

In parallel, expression of key genome maintenance and DNA repair genes prioritised by MR-GAT—*SIRT1, PARP1,* and *RAD51*—increased significantly following acute exercise (all $p < 0.05$). These genes form a coordinated repair axis supporting high-fidelity homologous recombination during cellular stress, yet their activation occurred without corresponding acute changes in ageing clock-associated molecular features.

Together, these experimental results demonstrate concordance between network-prioritised pathways and acute human exercise responses at the level of pathway engagement rather than uniform gene-level effects. Acute exercise preferentially modulated terminal glycosylation features while activating DNA repair programs, providing biological context for the distinct network roles observed in subsequent *in silico* perturbation analyses.

Network hierarchy links acute exercise responses to long-term ageing architecture

To characterise the network roles of exercise-responsive genes, we quantified the topological consequences of targeted *in silico* perturbation using complementary diffusion metrics. Diffusion concentration, measured by the Gini coefficient, captured the degree of localisation of perturbation effects, whereas diffusion entropy quantified regulatory breadth and network-wide signal dispersion (Figure 1).

Perturbation of validated glycosylation enzymes consistently produced highly localised diffusion profiles, reflected by elevated Gini coefficients and comparatively lower entropy (Figure 1A–B). These patterns indicate that glycosylation enzymes operate as compact enzymatic modules, exerting focused influence over tightly connected network neighbourhoods. Notably, these neighbourhoods were significantly enriched for genes defining multiple epigenetic ageing clocks and showed disproportionate alignment with the proteomic ageing axis, consistent with a role in encoding stable, system-level molecular state.

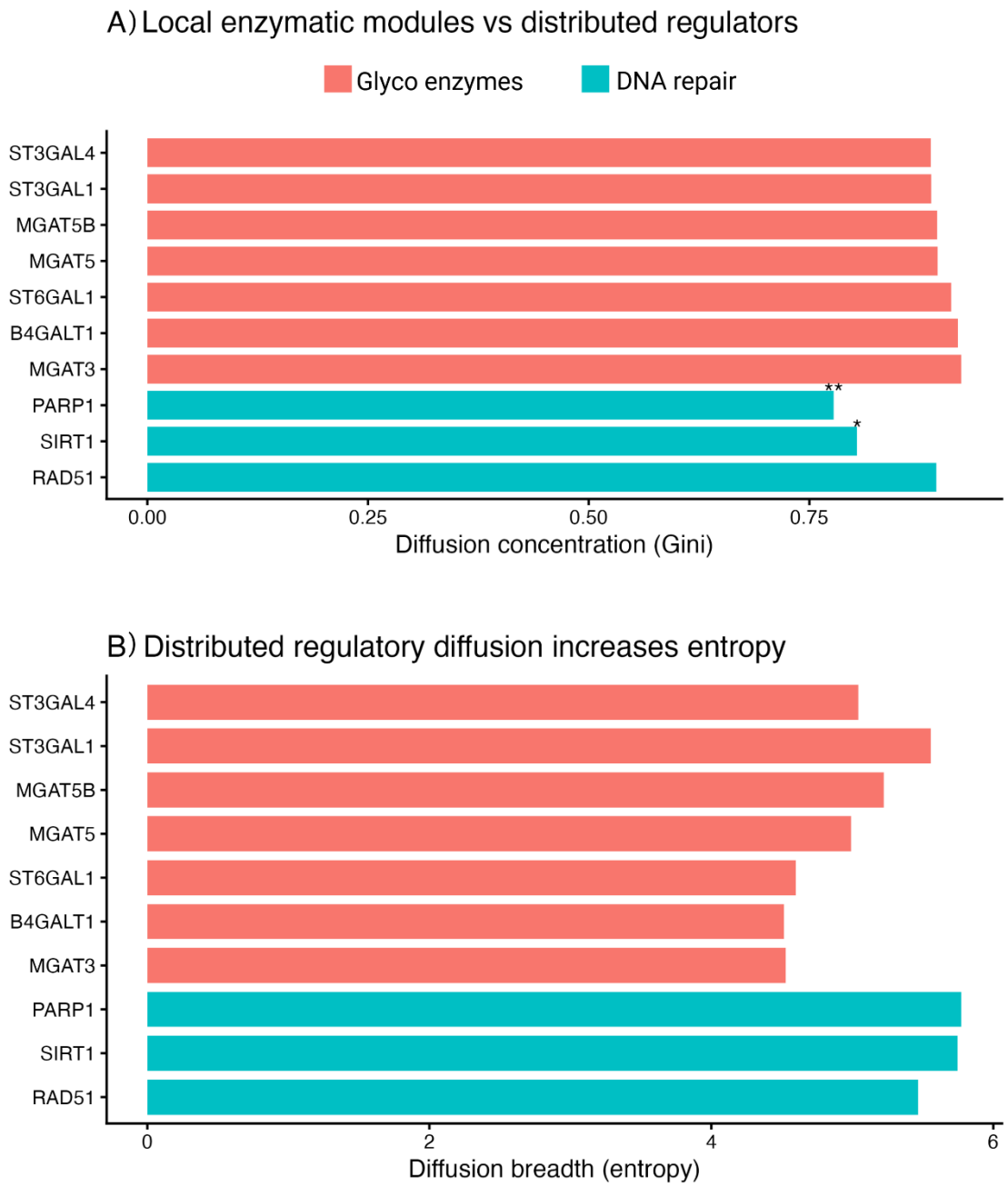

Figure 1. Distinct diffusion signatures distinguish local enzymatic modules from distributed regulatory genes. *A) Diffusion concentration (Gini coefficient) following targeted in silico perturbation of exercise-responsive genes. Glycosylation enzymes exhibit highly concentrated diffusion profiles, consistent with localised, modular network influence, whereas DNA repair and regulatory genes display lower concentration, indicating broader signal propagation. B) Diffusion breadth (entropy) is increased following perturbation of DNA repair and regulatory genes compared with glycosylation enzymes, reflecting more distributed network engagement. Statistical annotations indicate differences relative to glycosylation enzymes (*p < 0.05, **p < 0.01; empirical permutation tests). This pattern indicates that glycosylation enzymes operate as local enzymatic modules, exerting focused influence over tightly connected network neighbourhoods. Notably, these neighbourhoods were significantly*

enriched for genes defining multiple epigenetic ageing clocks and showed disproportionate alignment with the proteomic ageing axis, consistent with a role in encoding stable molecular state.

In contrast, perturbation of DNA repair and stress-response genes—including *PARP1, SIRT1,* and *RAD51*—produced broad, low-concentration diffusion patterns characterised by increased entropy (Figure 1B). Although these genes occupied central positions within the network, their influence was distributed across diverse functional modules rather than concentrated within ageing clock-enriched regions. This topology is consistent with upstream regulatory roles in coordinating acute stress responses rather than specifying long-term molecular state.

Network visualisation of diffusion neighbourhoods reinforced this distinction (Figures 2 and 3). Glycosylation enzymes were embedded within compact subnetworks densely populated by ageing clock-associated genes (Figure 2), whereas DNA repair genes resided within expansive, highly interconnected regions spanning transcriptional regulation, signalling, chromatin organisation, and stress-response pathways (Figure 3).

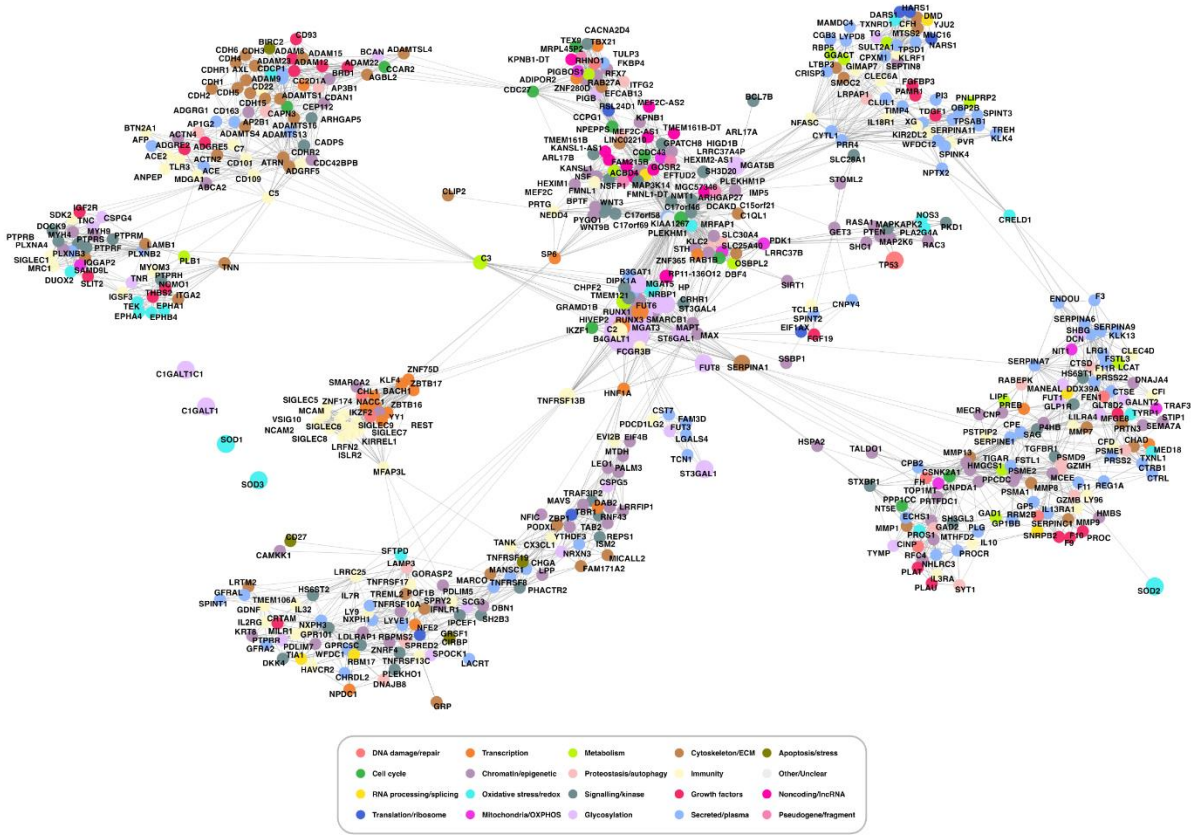

Figure 2. Glycosylation enzymes form locally concentrated network modules aligned with ageing clock architecture. *The top 500 genes with ageing clock support are shown in their network neighbourhood centred on the glycosylation enzyme B4GALT1, illustrating the topological organisation of glycosylation-related genes within the exercise-responsive molecular network. Nodes represent genes coloured by functional category, and edges indicate network connectivity. Glycosylation enzymes cluster within compact, highly interconnected local modules, consistent with concentrated diffusion and preferential enrichment of ageing clock-associated genes following perturbation. This localised architecture highlights glycosylation pathways as proximal encoders of stable molecular states captured by biological ageing clocks.*

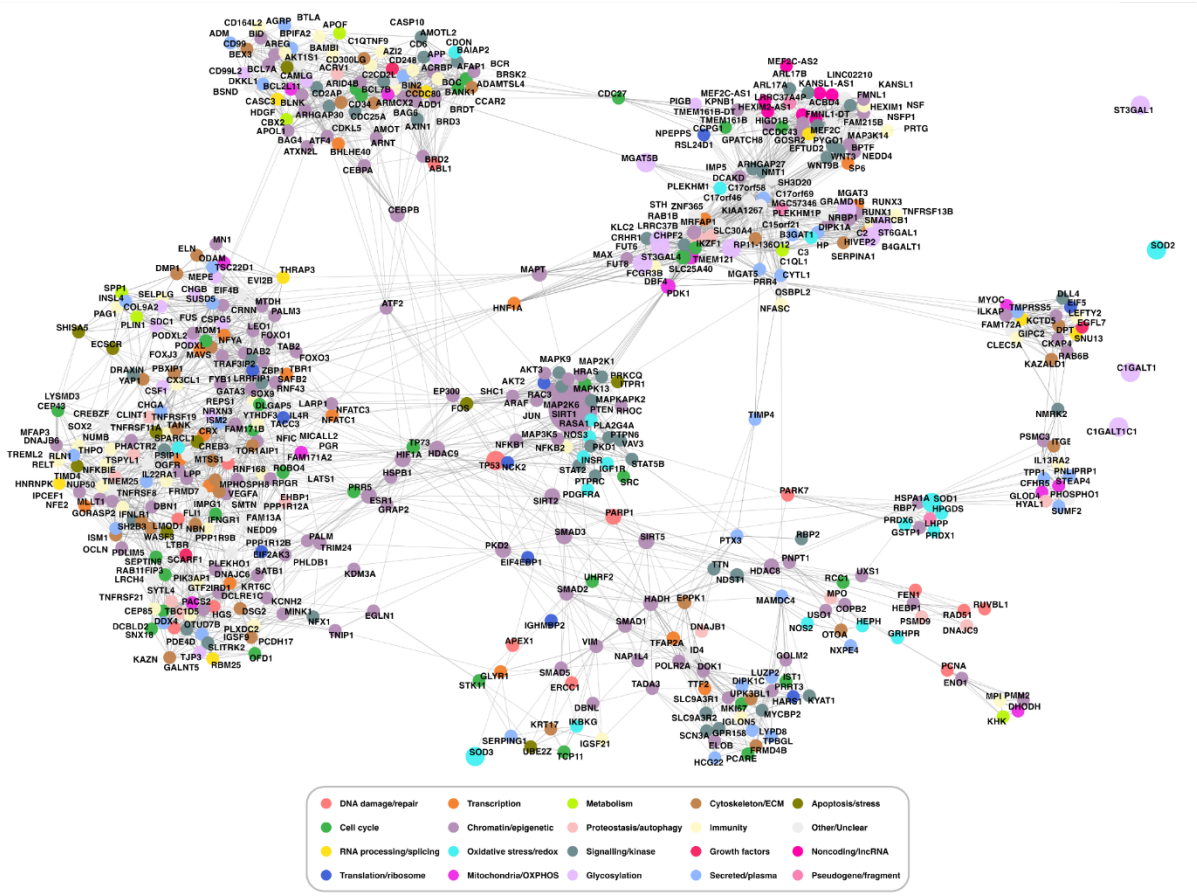

Figure 3. *DNA repair and stress response genes occupy distributed regulatory positions within the exercise-responsive network.* *The top 500 genes with ageing clock support are shown in their network neighbourhood centred on SIRT1 and DNA repair genes (PARP1, RAD51). Nodes represent genes coloured by functional category, and edges denote network connectivity. In contrast to the compact topology observed for glycosylation enzymes, DNA repair and stress response genes are embedded within distributed regulatory regions spanning transcriptional control, signalling, chromatin organisation, and metabolic processes. This*

diffuse topology is consistent with widespread diffusion, high entropy, and limited enrichment for ageing clock genes, reflecting upstream roles in buffering acute physiological stress rather than encoding long-term ageing-related molecular states.

Together, these analyses reveal a hierarchical organisation within exercise-responsive molecular networks. DNA repair pathways act as distributed regulators that preserve genome integrity during acute exercise-induced stress, whereas glycosylation networks function as localised effectors that encode the stable molecular patterns captured by biological ageing clocks. This hierarchy reconciles robust experimental evidence for exercise-induced DNA repair activation with the relative absence of DNA repair genes from ageing clock definitions, providing a mechanistic bridge linking habitual physical activity to long-term biological ageing trajectories without conflating acute repair responses with molecular ageing itself.

Impact of in silico perturbation on epigenetic and proteomic ageing clock architecture

To quantify how exercise-associated genes influence molecular ageing architecture, we assessed the downstream impact of *in silico* perturbation on genes defining established epigenetic and proteomic ageing clocks. For each perturbed gene, diffusion neighbourhoods generated by personalised PageRank were evaluated for enrichment of clock-associated differentially methylated region (DMR) genes across four epigenetic clocks (Horvath, Hannum, PhenoAge, DunedinPACE), as well as alignment with a proteomic ageing axis defined by first principal component (PC1) loadings of a Proteomic Clock derived from various cohorts. Significance was assessed using degree-matched null models with empirical permutation testing, ensuring that observed effects were not attributable to network connectivity alone.

Perturbation of validated glycosylation enzymes produced a consistent and significant impact on ageing clock architecture. Diffusion neighbourhoods originating from glycosylation genes captured a disproportionately large fraction of epigenetic clock DMR genes across multiple clocks, exceeding degree-matched expectations (empirical $p < 0.01$). These effects were robust across clocks that differ in biological emphasis, including mitotic history (Horvath), immune ageing (Hannum), and morbidity-linked phenotypes (PhenoAge, DunedinPACE). In parallel, glycosylation gene perturbations showed strong alignment with the proteomic ageing axis, reflected by elevated diffusion mass weighted by absolute PC1 loadings, indicating close coupling between glycosylation network structure and stable proteomic ageing signatures.

In contrast, perturbation of DNA repair and stress response genes—including *PARP1, SIRT1,* and *RAD51*—did not yield consistent enrichment of epigenetic clock DMR genes or significant alignment with the proteomic ageing axis (empirical $p > 0.1$ across clocks). Although these genes were centrally positioned within the exercise-responsive network and exhibited broad diffusion profiles, their downstream influence was distributed across diverse functional modules rather than concentrated within clock-associated regions. This pattern indicates that

DNA repair pathways, while essential for coordinating acute stress responses, do not directly organise the stable molecular features captured by ageing clocks.

Together, these findings demonstrate that *in silico* perturbation distinguishes exercise-associated genes that predictively shape ageing clock architecture from those that primarily mediate transient physiological responses. Glycosylation networks emerge as proximal organisers of long-term molecular ageing states, whereas DNA repair pathways function as upstream regulators that preserve genomic integrity without directly encoding ageing clock structure.

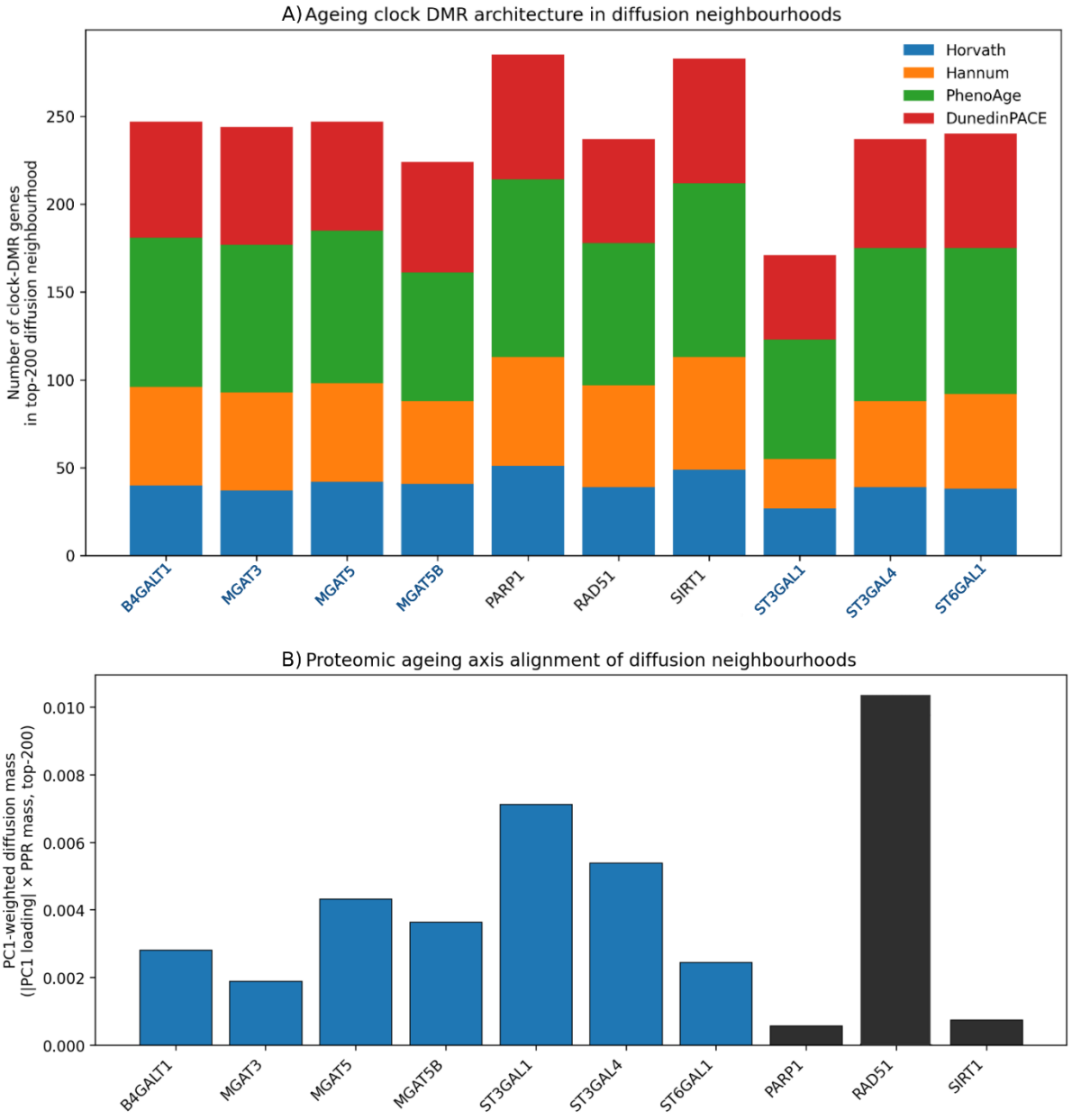

Figure 4. Diffusion-based impact of exercise-associated genes on epigenetic and proteomic ageing architecture. *A) Stacked bar plot showing the number of epigenetic ageing clock DMR-associated genes captured within the top 200 diffusion neighbourhood of each perturbed*

gene, stratified by clock (Horvath, Hannum, PhenoAge, DunedinPACE). Gene labels are colour-coded by functional class: glycosylation enzymes (blue) and DNA repair/stress-response genes (black). DNA repair genes exhibit broader diffusion neighbourhoods with high overall DMR coverage across clocks, consistent with central roles in coordinating stress and repair responses, whereas glycosylation enzymes capture a more selective and clock-structured subset of DMR targets. B) Bar plot showing alignment of diffusion neighbourhoods with a proteomic ageing axis, quantified as diffusion mass within the top 200 neighbourhood weighted by absolute consensus PC1 loadings. Glycosylation enzymes show consistently elevated PC1-weighted diffusion signals across multiple genes, whereas DNA repair/stress response genes display more heterogeneous behaviour, including a single dominant peak (RAD51). This pattern suggests that glycosylation networks are coherently coupled to stable proteomic ageing signatures, while DNA repair pathways influence the proteomic ageing axis in a more gene-specific and context-dependent manner.

Discussion

Exercise modulates biological ageing, yet the molecular hierarchy linking habitual physical activity to long-term ageing signatures in humans remains unclear. Addressing this gap requires moving beyond causal gene identification toward frameworks that can predict how exercise-responsive pathways reorganise molecular systems over time. In this study, we extend genetically anchored causal prioritisation with predictive, network-based *in silico* perturbation to forecast how exercise-associated genes propagate influence through molecular networks to shape ageing-relevant molecular architecture. By integrating MR, self-supervised graph representation learning, diffusion-based perturbation, and biological validation in a controlled human exercise intervention, we provide a unified framework that links acute exercise responses to stable ageing clock organisation without conflating transient stress signalling with biological ageing itself.

Consistent with prior work, genetically proxied habitual VPA was associated with widespread molecular effects across proteomic, epigenomic, glycomic, and single-cell transcriptomic layers. However, harmonisation of these signals to gene symbols revealed limited cross-layer convergence, with only a small subset of genes supported by multiple omic layers. This sparse overlap underscores a key limitation of gene-centric interpretations of exercise MR: causal signals are abundant but fragmented across biological scales. These findings motivated a network-based interrogation, in which exercise-responsive genes are interpreted in terms of their position, connectivity, and downstream influence, rather than isolated effect sizes.

Importantly, embedding of MR-prioritised genes within a human multi-omic interaction network was performed using self-supervised contrastive representation learning followed by MR-guided refinement, without incorporating any ageing-related labels. This design ensured

that subsequent associations with ageing clocks emerged from network topology and information flow, rather than from supervised optimisation toward ageing outcomes.

Targeted *in silico* perturbation provided a predictive lens through which to examine how exercise-responsive genes propagate influence across molecular systems. Unlike supervised ranking models, this diffusion-based framework explicitly simulates the downstream consequences of gene-level perturbation, allowing discrimination between genes that act as upstream regulators of acute stress responses and those that organise stable molecular states.

Across multiple epigenetic ageing clocks and a proteomic ageing axis, perturbation of validated glycosylation enzymes consistently produced diffusion neighbourhoods enriched for clock-defining genes, exceeding degree-matched expectations. These effects were robust across clocks with distinct biological emphases and were accompanied by strong alignment with proteomic ageing loadings. In contrast, perturbation of canonical DNA repair and stress response genes—including *PARP1, SIRT1,* and *RAD51*—did not yield consistent enrichment of ageing clock architecture, despite their central network positions and broad diffusion profiles.

Topological analyses clarified the basis of this divergence. Glycosylation enzymes occupied compact, highly localised network modules characterised by high diffusion concentration and low entropy, indicating focused influence over tightly connected neighbourhoods. DNA repair genes, by contrast, resided within distributed regulatory regions spanning transcriptional control, chromatin organisation, metabolic signalling, and stress-response pathways, consistent with broad but non-specific downstream propagation. These findings establish a hierarchical network organisation in which glycosylation pathways act as proximal encoders of ageing-related molecular state, while DNA repair pathways function as upstream buffers of acute physiological stress.

The controlled human high-intensity exercise intervention provides critical biological context for this hierarchy. Acute exercise elicited rapid modulation of terminal plasma N-glycosylation features, including reduced agalactosylation and increased galactosylation and sialylation, while core architectural features such as bisecting GlcNAc remained unchanged. These selective shifts are consistent with dynamic regulation of immune and inflammatory signalling rather than total glycan remodelling.

In parallel, expression of genome maintenance and DNA repair genes prioritised by the network—including *SIRT1, PARP1,* and *RAD51*—increased acutely following high-intensity exercise, confirming engagement of a coordinated repair programme *in vivo*. At the pathway level, experimental findings converged with network predictions. Enzymes governing galactosylation and sialylation were both acutely responsive and centrally positioned within ageing-aligned network modules, whereas enzymes controlling glycan architecture showed no acute change, consistent with their inferred role in shaping longer-term glycosylation capacity. Thus, concordance between MR, network topology, *in silico* perturbation, and

experimental validation emerged at the level of pathway engagement, rather than as uniform directional effects of individual genes.

Together, these results clarify how acute exercise-induced stress is translated into durable molecular resilience without being directly embedded within biological ageing clocks. Robust activation of DNA repair pathways following exercise reflects a protective, buffering response that preserves genome integrity across repeated physiological challenges, rather than a cumulative ageing signal. Such pathways are therefore essential enablers of long-term adaptation, but remain diffuse and non-specific at the level captured by ageing clocks. In contrast, glycosylation networks occupy a downstream, integrative position, encoding the time-averaged consequences of habitual physical activity on immune and inflammatory state. These stable, system-level adaptations accumulate across repeated exercise exposures and are consequently represented within epigenetic and proteomic ageing measures. Viewed in this framework, biological ageing clocks selectively capture molecular features that integrate and retain the imprint of repeated stress adaptation, while excluding transient repair responses whose role is to maintain system robustness rather than define ageing trajectory.

Importantly, this framework avoids causal overinterpretation. *In silico* perturbation models network influence rather than direct molecular causation, and ageing clocks remain abstractions of biological ageing. Nonetheless, the convergence of genetically anchored prioritisation, predictive network perturbation, and human experimental validation across independent data modalities provides strong support for the inferred hierarchy.

By introducing graph-based *in silico* perturbation to human exercise biology, this study demonstrates how predictive network modelling can complement causal inference and experimental validation. Rather than producing static gene lists, this approach enables forecasting of emergent molecular states, offering a principled route to identify pathways through which lifestyle exposures shape long-term ageing trajectories. More broadly, our framework illustrates how exercise science can move beyond mechanistic descriptions of acute responses toward systems-level models of adaptation and healthy ageing, with potential applications to intervention design, biomarker development, and precision exercise medicine.

Methods

Graph structure and node representations

All *in silico* perturbation analyses were performed on the same gene interaction graph and node index used in the supervised MR-GAT model, ensuring full topological consistency with the primary exercise-ageing network (Juan & Ntasis, 2026). Nodes represent protein-coding genes, and edges encode functional gene-gene associations derived from the k-nearest neighbors (KNN) similarity graph constructed from integrated multi-omic, structural, and functional features, as described previously. The graph topology was treated as fixed throughout all perturbation and diffusion analyses. Each gene node was associated with a

learned low-dimensional embedding vector. Base embeddings were obtained via self-supervised contrastive pretraining, followed by optional MR-informed refinement. These embeddings provide a continuous latent representation of gene context, integrating functional annotation, structural properties, and network neighbourhood information.

Self-supervised contrastive embedding pretraining

To obtain task-agnostic gene embeddings independent of explicit ageing or MR supervision, we performed self-supervised contrastive pretraining on the node feature matrix. Two stochastic views of each node were generated via feature-level dropout and encoded using a shared multilayer perceptron (MLP). Embeddings were trained using a SimCLR-style NT-Xent contrastive loss, which encourages embeddings of the same gene under stochastic feature dropout to be similar while maximising separation from all other genes in the batch. Training was performed for 500 epochs using the Adam optimiser (learning rate = 1×10^{-3}, weight decay = 1×10^{-5}). The resulting embeddings capture intrinsic graph-aware functional similarity without reference to ageing clocks or causal targets. Full loss definition and optimisation details are provided in the supplementary files.

MR-informed embedding refinement via masked multi-task regression

To align the latent embedding space with causal exercise biology while preserving global structure, embeddings were optionally refined using masked multi-task regression on standardised MR effect sizes across proteomic, CpG methylation, glycomic, and single-cell transcriptomic layers. A lightweight projection network mapped base embeddings **Z** to refined embeddings **Z'**, and task-specific regression heads predicted MR effect sizes where available. Missing targets were masked such that loss was computed only for genes with valid MR estimates. A Huber loss was used for robustness, and a stability regularisation term penalised large deviations between **Z** and **Z'**. All training objectives and masking logic are detailed in the supplementary files.

In silico gene perturbation in embedding space

In silico perturbation was implemented by explicitly modifying the embedding vector of a target gene while holding all other embeddings fixed. For a given gene ***g*** with embedding $\mathbf{z}_g$, perturbation was defined as:

$$\mathbf{z}'_g = (1 \pm s)\,\mathbf{z}_g$$

where ***s*** is the perturbation strength (default ***s*** = 1.0), with + corresponding to overexpression (OE) and − to knockdown (KD). The magnitude of the perturbation was quantified as:

$$\Delta_g = \left\|\mathbf{z}'_g - \mathbf{z}_g\right\|_2$$

This scalar ensured that downstream network effects scaled with perturbation size rather than direction alone.

Network diffusion of perturbation effects

To propagate local perturbations across the gene interaction network, we applied a personalised PageRank (PPR)-style diffusion process centred on the perturbed gene. Starting from a one-hot seed vector $\mathbf{e}_g$, diffusion was iteratively updated as:

$$\mathbf{p}^{(t+1)} = (1-\alpha)\,\mathbf{e}_g + \alpha\,\mathbf{D}^{-1}\mathbf{A}\,\mathbf{p}^{(t)}$$

where α is the restart probability (default α = 0.85), **A** is the adjacency matrix, and **D** is the diagonal degree matrix. Diffusion was run for 50 iterations, yielding a steady-state influence vector **p** over all genes. The final diffusion-weighted impact score for each gene *i* was defined as:

$$\text{impact}_i = p_i \cdot \Delta_g$$

Genes were ranked by this score to identify those most affected by the simulated perturbation.

Quantifying perturbation breadth and concentration

To characterise whether perturbations produced localised or global network effects, we quantified the breadth of diffusion using complementary metrics. Diffusion entropy was computed as:

$$H(\mathbf{p}) = -\sum_i q_i \log q_i, \quad q_i = \frac{p_i}{\sum_j p_j}$$

where higher entropy indicates broader network propagation. In parallel, the Gini coefficient of **p** was computed to quantify inequality of diffusion mass across nodes. Top-K mass concentration (K = 10, 50, 200) was also recorded.

Degree-matched null models and empirical significance testing

To assess whether observed diffusion patterns exceeded expectations given network topology, degree-matched null distributions were constructed. For each perturbed gene, null seed genes with similar degree were sampled, and diffusion was recomputed using identical parameters. Two-sided empirical p-values were computed for all diffusion metrics as:

$$p_{\text{emp}} = \frac{\#\left\{\,|x_{\text{null}} - \mu| \geq |x_{\text{obs}} - \mu|\,\right\} + 1}{n_{\text{null}} + 1}$$

where x_{obs} is the observed statistic, x_{null} are null values, and μ is the null mean. This non-parametric framework controls for degree-driven and topological confounding.

Quantifying impact on ageing-related molecular programmes

Diffusion footprints were further intersected with ageing-related gene sets, including epigenetic clock DMR genes (Horvath, Hannum, PhenoAge, DunedinPACE), proteomic ageing PC1 genes, and a validated glycosylation enzyme panel. For each annotation set ***S***, total and top-K captured diffusion mass were computed as:

$$\sum_{i \in S} p_i \quad \text{and} \quad \sum_{i \in S \cap T_K} p_i$$

Optionally, DMR-based analyses incorporated per-gene weights derived from the number of DMRs or mean absolute methylation change. All annotation-level metrics were evaluated against degree-matched null distributions using the empirical testing framework above.

External evaluation using independent ageing targets

Frozen embeddings were evaluated against independent ageing-related targets, including epigenetic clock DMR summaries and proteomic ageing principal components. Regularised linear models (Ridge or Elastic Net) were trained on predefined degree-stratified splits. Performance was assessed using test-set R^2, Spearman rank correlation, and permutation-based significance testing.

Data Availability

Supplementary files are available at https://github.com/cgjuan01

Code Availability

All scripts used for data processing, MR analyses, network construction, and GNN training are available at https://github.com/cgjuan01

References

Onigbinde, S., Adeniyi, M., Daramola, O., et al. (2025). Glycomics in human diseases and its emerging role in biomarker discovery. *Biomedicines, 13*(8), 2034. https://doi.org/10.3390/biomedicines13082034

Juan, C. G., & Ntasis, L. (2026). Multi-omic deep learning identifies exercise-responsive ageing pathways in humans [Preprint]. *medRxiv*. https://doi.org/10.64898/2025.12.26.25343061

Juan, C. G., Šimunić-Briški, N., Volpe-Zanutto, F., *et al.* (2025). Quercetin activates the SIRT6–Nrf2 axis during oxidative stress, modulating DNA repair and ageing-associated markers in healthy men [Preprint]. *medRxiv*. https://doi.org/10.1101/2025.10.31.25338366

Kellokumpu, S. (2019). Golgi pH, ion and redox homeostasis. *Frontiers in Cell and Developmental Biology, 7*, 93. https://doi.org/10.3389/fcell.2019.00093

Khoder-Agha, F., & Kietzmann, T. (2021). The glyco-redox interplay. *Redox Biology, 43*, 101962. https://doi.org/10.1016/j.redox.2021.101962

Radak, Z., Bori, Z., Koltai, E., et al. (2011). Age-dependent changes in 8-oxoguanine-DNA glycosylase activity are modulated by adaptive responses to physical exercise in human skeletal muscle. *Free Radical Biology and Medicine, 51*(2), 417–423. doi:10.1016/j.freeradbiomed.2011.04.018

Trzos, S., Link-Lenczowski, P., & Pocheć, E. (2023). The role of N-glycosylation in B-cell biology and IgG activity: The aspects of autoimmunity and anti-inflammatory therapy. *Frontiers in Immunology, 14*, 1188838. https://doi.org/10.3389/fimmu.2023.1188838

Williamson, J., Hughes, C. M., Burke, G., & Davison, G. W. (2020). A combined γ-H2AX and 53BP1 approach to determine the DNA damage-repair response to exercise in hypoxia. *Free Radical Biology and Medicine, 154*, 9–17. doi:10.1016/j.freeradbiomed.2020.04.026

Zhu, L., Su, Y., & Su, Z. (2026). Impact of exercise-induced DNA damage repair on age-related muscle weakness and sarcopenia. *Frontiers in Genetics, 16*, 1639224. https://doi.org/10.3389/fgene.2025.1639224