Robust disease module mining via enumeration of diverse prize-collecting Steiner trees

Judith Bernett, Dominik Krupke, Sepideh Sadegh, Jan Baumbach, Sándor P. Fekete, Tim Kacprowski, Markus List, David B. Blumenthal

Supplementary material

1 Supplementary methods

1.1 Formalizing the Steiner tree enumeration problem

As specified in Section 2.2 of the main document, our DMMM ROBUST is designed to provide a solution for the following problem specification: *Given a weighted network $G = (V,E,w)$ and a set of seed nodes $S \subseteq V$, compute an induced subgraph $G[M]$, where $M \subseteq V$ contains nodes that appear in many diverse near-optimal generalized Steiner trees for S.* To turn this informal specification into a well-defined definition, we need to formally define the qualifiers "diverse", "near-optimal", and "generalized". Below, we discuss possible formalizations. Note that we did not develop ROBUST as a solver for any specific formalization. Instead, ROBUST should be seen as a procedural approach designed with the informal problem specification in mind. The main purpose of the formalizations given below is that they may serve as points of departure for future algorithmic developments.

Generalized minimum-weight Steiner trees (MWSTs). Let $S \subseteq V$ be a seed set and $T = (V_T, E_T)$ be a generalized MWST for S. Recall that T is generalized in the sense that we do not strictly enforce $S \subseteq V_T$ but allow that some seeds are left uncovered. Intuitively, we would like to allow few seeds to be discarded if doing so dramatically decreases the tree's weight $w(T) \coloneqq \sum_{e \in E_T} w(e)$. In ROBUST, we achieve this algorithmically by using a prize-collecting Steiner tree (PCST) instead of a MWST model, where we assign every seed a high but not infinite value. However, the question remains how to formally define generalized MWSTs.

Let $OPT(S)$ denote the weight of an MWST for S and $\nu \in (0,1)$, $\rho \in (1,\infty)$ be constants. Then we can define a tree $T = (V_T, E_T)$ to be a (ν,ρ)-generalized MWST for S if and only if the following two conditions hold: (a) At least a fraction of ν of the seeds are covered, i. e., $|V_T \cap S| \geq \nu \cdot |S|$. (b) The weight per covered seed in T is at least ρ times lower than the weight per covered seed in an MWST for S, i. e., $\rho \cdot w(T)/|V_T \cap S| \leq OPT(S)/|S|$. In the sequel, we let $\mathbb{T}_{\nu,\rho}(S)$ denote the set of all such defined (ν,ρ)-generalized MWSTs for S and define $OPT_{\nu,\rho}(S) \coloneqq \min\{w(T) \mid T \in \mathbb{T}_{\nu,\rho}(S)\}$.

Near-optimal generalized MWSTs. Given some non-decreasing function $\alpha : \mathbb{N} \to (1,\infty)$, we can define the set of all α-near-optimal (ν,ρ)-generalized MWSTs as $\mathbb{T}_{\nu,\rho,\alpha}(S) \coloneqq \{T \in \mathbb{T}_{\nu,\rho}(S) \mid w(T) \leq \alpha(|S|) \cdot OPT_{\nu,\rho}(S)\}$. The function α should be seen as the admissible approximation ratio for the weights of the trees. In particular, choosing α as a constant means that a constant approximation guarantee is required.

Diverse near-optimal generalized MWSTs. To include diversity into the picture, we need to have access to a distance measure d for two α-near-optimal (ν,ρ)-generalized MWSTs T and T'. Since, in the context of disease module mining, we are interested in diverse node sets, a straightforward approach is to define $d(T,T') \coloneqq 1 - |V_T \cap V_{T'}|/|V_T \cup V_{T'}|$ as the Jaccard distance over the trees' node sets. Next, we need to define a diversity metric D for sets $\mathscr{T}$ of near-optimal generalized MWSTs. The natural choices here are to define $D(\mathscr{T})$ either as the mean distance $D(\mathscr{T}) \coloneqq \sum_{\{T,T'\} \in \binom{\mathscr{T}}{2}} d(T,T') / \binom{|\mathscr{T}|}{2}$ or as the minimum distance $D(\mathscr{T}) \coloneqq \min\{d(T,T') \mid \{T,T'\} \in \binom{\mathscr{T}}{2}\}$. Equipped with D and some constant n specifying the number of trees to be enumerated, we can then formalize our Steiner tree enumeration problem as the problem of maximizing D over all k-element subsets of the set $\mathbb{T}_{\nu,\rho,\alpha}(S)$ of all α-near-optimal (ν,ρ)-generalized MWSTs.

1.2 Selecting the values for the seeds

We do not use very high values for the seeds, because we cannot compute optimal prize-collecting Steiner trees but only approximations. The higher the values of the seeds, the lower the influence of the edge weights and other node values becomes but we are interested in those and not in the values of the seeds. The guaranteed approximation factor, thus, could be bloated and weakened. For instance, if the values of the seeds make up 50 % of the optimal objective value, the guaranteed approximation factor increases to 3. While the guaranteed approximation factor is only a worst-case upper bound and the actual solutions are often nearly optimal, it is still recommendable not to use unnecessarily high values.

1.3 Selecting the values for the non-seeds

Using a fraction of the minimal edge weight for the values of non-seeds works well if the edge weights are similar, best for uniform edge weights. If the minimal edge weight is (nearly) zero, this approach fails. By giving individual values based on the edge weights of incident edges, one could mitigate this problem. When choosing the initial values, it is important to make sure that they are not too high such that nodes become integrated just for their values but are actually bad for the Steiner tree. One also has to remember that the used algorithm is only an approximation algorithm and could, in worst-case, also choose bad nodes as long as they do not increase the objective by a factor greater than two. Therefore, a small safety margin is advised.

1.4 Enforcing full Steiner trees

In some cases, the proposed algorithm will not integrate all seeds into the returned trees. As discussed in the main document, this is the intended behaviour for potentially noisy seeds, but you may nonetheless find yourself in the situation where you want all seeds to be included. For this case, let us quickly mention two options to modify the algorithm:

- First, the used PCST-algorithm allows to specify a root node that has to be integrated. Using this option gives us a simple option to integrate a single missing seed node but is not sufficient for all cases.
- If a tree T returned by `pcst_apx` does not contain all seeds S, contract T in G and consider it as a seed node, replacing the seeds contained in T. Now repeatedly execute `pcst_apx` on the modified graph until only a single seed remains. At this point, we can expand the contractions again and return a connected tree. If in one step, `pcst_apx` returns a single node, simply use the cheapest path between any two seeds. This approach yields a proper Steiner tree in at most $O(|S|)$ iterations.

1.5 Structure of approximated solutions

Linear programming (LP) relaxations are not only the base for the best approximations of Steiner tree problems, but also for exact solvers, as shown by Fischetti *et al.* (2017) in the DIMACS competition on Steiner tree problems[1]. For many natural instances, the LP relaxation captures a lot of the structure of the optimal solution, allowing us to make good branching decisions to quickly reach the optimal solution among an exponential number of possibilities. The PCST algorithm used by us also utilizes the LP relaxation (and its dual), which enables it to find solutions close to the optimum for many natural instances. This is shown in the experiments by Hegde *et al.* (2014), in which the algorithm solved multiple instances with hundreds of nodes to (near) optimality. Thus, this algorithm not only limits the maximal deviation of the objective value from the optimum, but can actually compute (partially) optimal solutions if the instance is benign.

2 Supplementary findings for the case study in multiple sclerosis

In addition to the quantitative evaluation reported in the previous sections, we performed a case study in multiple sclerosis (MS) to showcase how to use ROBUST for hypothesis generation. First, we constructed a context-specific PPI network from IID by filtering for the interactions experimentally validated in brain tissue. Then, proteins associated with MS were obtained by merging DisGeNet and OMIM annotations. This yielded 42 seeds, 26 out of which were contained in the context-specific network.

Running ROBUST on these 26 seed nodes yielded a disease module with 90 additional proteins (see Supplementary Figure 3), including the galectin-1 (P09382), which was found in each of the 30 trees computed by ROBUST. It has

[1] http://dimacs11.zib.de/contest/results/results.html

been shown that galectin-1 plays an important regulatory role in MS patients (Lutomski *et al.*, 1997; Starossom *et al.*, 2012; Ramirez Hernandez *et al.*, 2020). We then took a closer look at the 2-hop neighborhood of galectin-1 within the computed diseases module (visualized in Supplementary Figure 4 and Supplementary Figure 5, 48 nodes, 3 seeds). In this submodule, we found i.a. thioredoxin (TXN), Peroxiredoxin-2 (PRDX2), the mitochondrial Thioredoxin-dependent peroxide reductase (PRDX3), and DJ-1.

Thioredoxin, Peroxiredoxin-2, PRDX3, and DJ-1 are antioxidant molecules related to oxidative stress, a sign of various neurological disorders (Krapfenbauer *et al.* (2003)), including MS (Liu *et al.* (2020)). Thioredoxin has been found to be significantly upregulated in MS patients compared to healthy controls (Pennisi *et al.* (2011); Mahmoudian *et al.* (2017)). Peroxiredoxin-2 and PRDX3 are enzymes which reduce H_2O_2 and hydroperoxides to water and alcohols using thioredoxin as substrate (Kamariah *et al.* (2016); Cao *et al.* (2007)). They are therefore involved in cell protection against oxidative stress. Peroxiredoxin-2 was shown to be upregulated in white matter MS lesions (Voigt *et al.* (2017)). While DJ-1 is not directly linked to Thioredoxin, the two molecules share downstream targets and it has been suggested that there is some cross-talk between these two systems (Raninga *et al.* (2014); Im *et al.* (2012)). Various studies have linked DJ-1 to MS (Hirotani *et al.* (2008); Lev *et al.* (2006); van Horssen *et al.* (2010)). These findings demonstrate how ROBUST can identify a submodule related to oxidative stress in MS whose participants share common pathways.

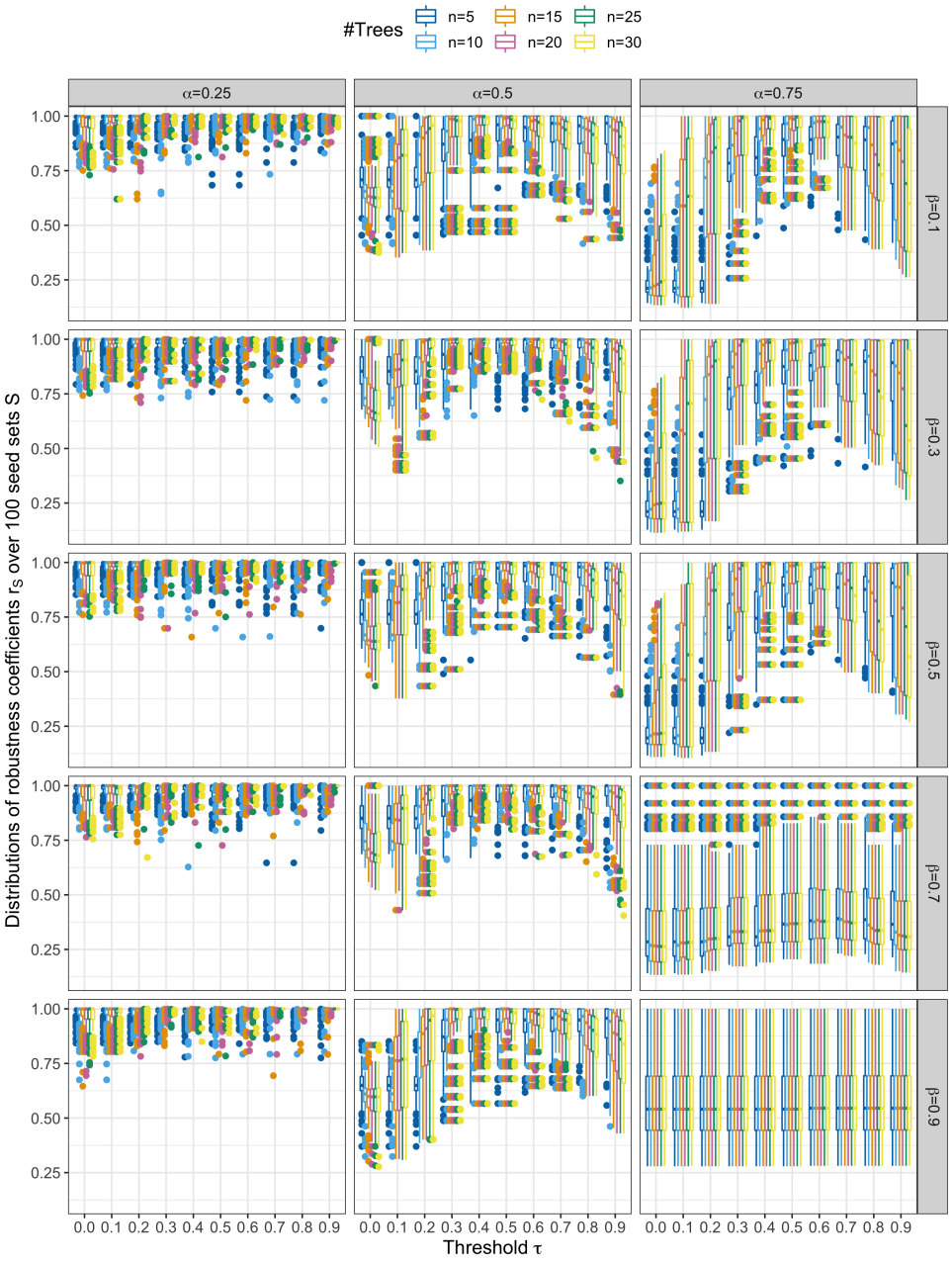
The submodule contained additional interesting proteins: poly [ADP-ribose] polymerase 1 (PARP-1), the high mobility group protein B1 (HMGB1), as well as 14-3-3 protein epsilon and 14-3-3 protein beta/alpha.

Poly [ADP-ribose] polymerase 1 is nuclear enzyme for DNA repair that has been shown to be activated in a primate model of MS (Kauppinen *et al.* (2005)) and inhibitors of PARP1 have been suggested as drug candidates (Veto *et al.* (2010); Cavone and Chiarugi (2012)).

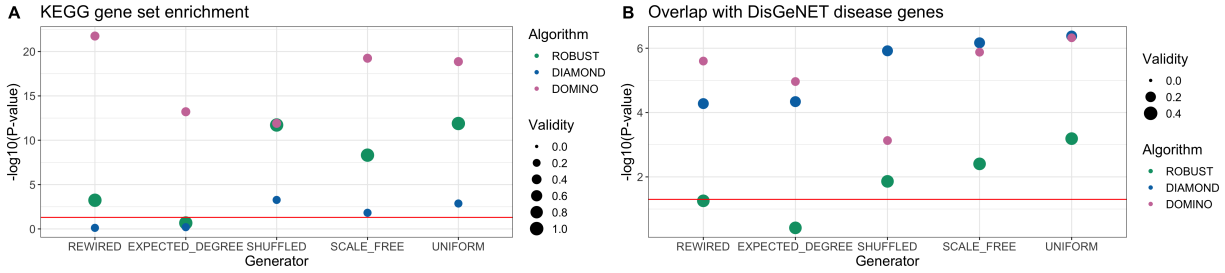
The 14-3-3 protein has seven isoforms, indicated by greek letters and is expressed abundantly in neurons in the central nervous system. It was shown to be a biomarker for MS severity (Colucci *et al.* (2004); Martínez-Yélamos *et al.* (2001); Satoh *et al.* (2004)).

High mobility group protein B1 is elevated in various neurological diseases like brain injury, neuroinflammation, Alzheimer's and Parkinson disease, as well as multiple sclerosis (Imitola (2019); Paudel *et al.* (2019, 2020)).

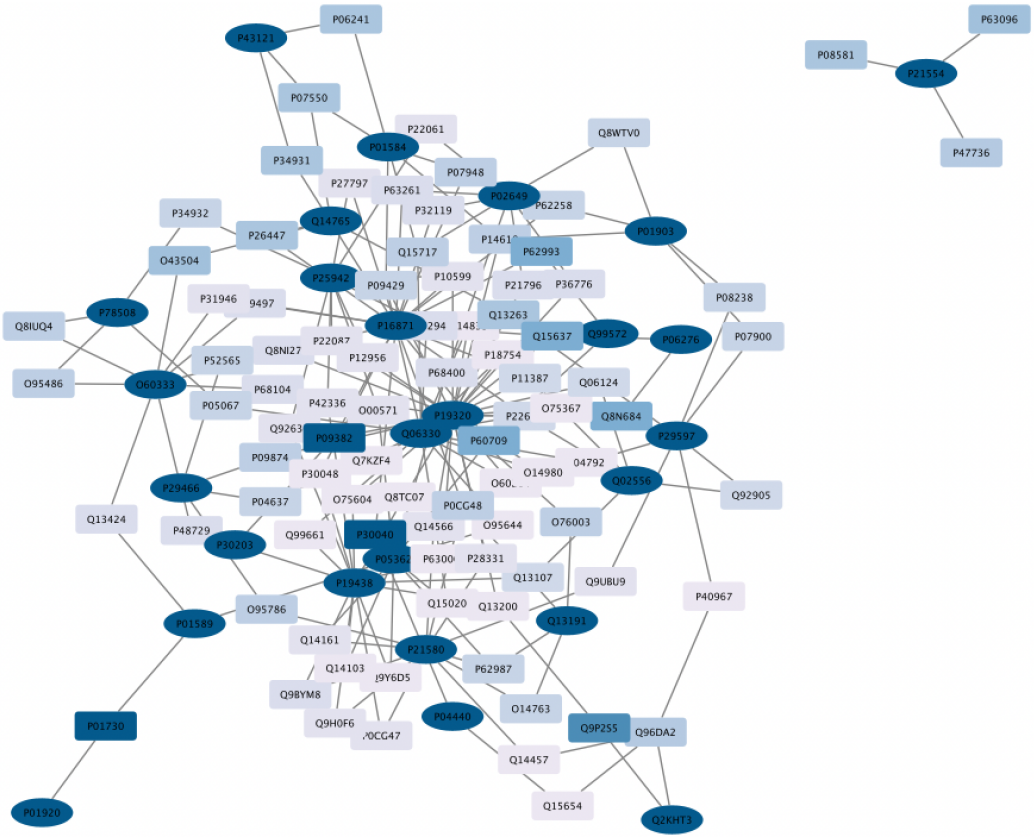
3 Supplementary figures



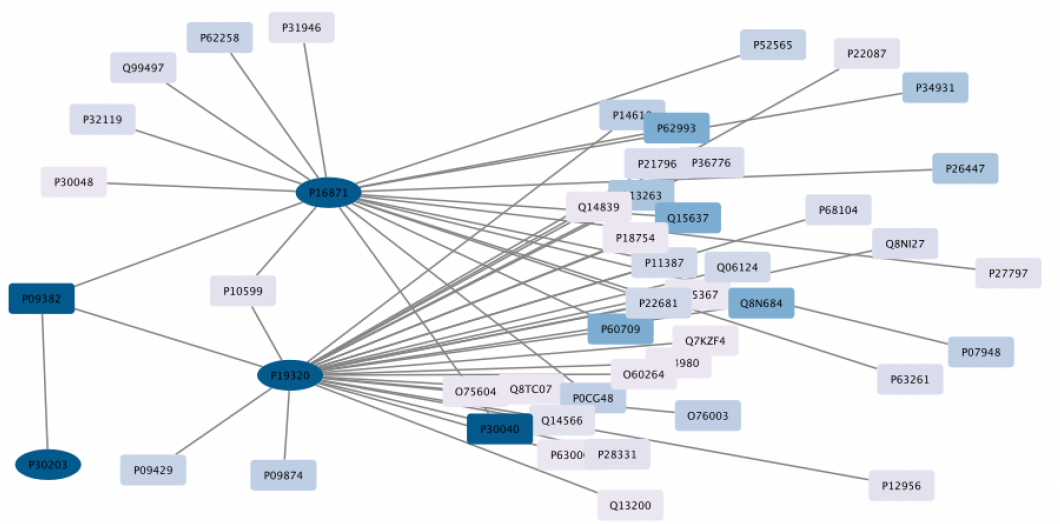
Supplementary Figure 1. Effect of hyper-parameters on the robustness of ROBUST.



Supplementary Figure 2. Results of network permutation tests implemented in the test suite introduced by Lazareva *et al.* (2021). The visualized negative log-transformed *P*-values are obtained by using the Mann-Whitney *U* test to compare the functional relevance scores obtained on real PPI networks to those obtained on randomized networks generated via five different network generators. The validity scores visualized via the sizes of the data points indicate to which extent the DMMM obtained meaningful functional relevance scores on the original PPI networks. For more details, we again refer to Lazareva *et al.* (2021). **A**: KEGG gene set enrichment *P*-values are used as functional relevance scores. **B**: Overlap coefficients with DisGeNET disease genes are used as functional relevance scores.



Supplementary Figure 3. MS disease module computed by ROBUST. Seeds are marked as ellipses, additional genes discovered by ROBUST as rectangles. The fraction ρ of occurrences in the enumerated Steiner trees is coded as different shades of blue (dark: $\rho = 1.0$, light: $\rho = 0.1$). Node labels are UniProt IDs.



Supplementary Figure 4. Submodule induced by 2-hop neighborhood of galectin-1 (P09382) within the MS disease module computed by ROBUST. The nodes of the left-hand side have been linked to multiple sclerosis.

References

Cao, Z. *et al.* (2007). Reconstitution of the mitochondrial prxiii antioxidant defence pathway: general properties and factors affecting prxiii activity and oligomeric state. *J. Mol. Biol.*, **372**(4), 1022–1033.

Cavone, L. and Chiarugi, A. (2012). Targeting poly (adp-ribose) polymerase-1 as a promising approach for immunomodulation in multiple sclerosis? *Trends Mol. Med.*, **18**(2), 92–100.

Colucci, M. *et al.* (2004). The 14-3-3 protein in multiple sclerosis: a marker of disease severity. *Mult. Scler. J.*, **10**(5), 477–481.

Fischetti, M. *et al.* (2017). Thinning out Steiner trees: a node-based model for uniform edge costs. *Math. Program. Comput.*, **9**(2), 203–229.

Hegde, C. *et al.* (2014). A fast, adaptive variant of the Goemans-Williamson scheme for the prize-collecting Steiner tree problem. In *11th DIMACS Implementation Challenge*.

Hirotani, M. *et al.* (2008). Correlation between DJ-1 levels in the cerebrospinal fluid and the progression of disabilities in multiple sclerosis patients. *Mult. Scler. J.*, **14**(8), 1056–1060.

Im, J.-Y. *et al.* (2012). DJ-1 induces thioredoxin 1 expression through the Nrf2 pathway. *Hum. Mol. Genet."*, **21**(13), 3013–3024.

Imitola, J. (2019). New age for progressive multiple sclerosis. *Proc. Natl. Acad. Sci. USA*, **116**(18), 8646–8648.

Kamariah, N. *et al.* (2016). Transition steps in peroxide reduction and a molecular switch for peroxide robustness of prokaryotic peroxiredoxins. *Sci. Rep.*, **6**(1), 1–15.

Kauppinen, T. M. *et al.* (2005). Poly (adp-ribose) polymerase-1 activation in a primate model of multiple sclerosis. *J. Neurosci. Res.*, **81**(2), 190–198.

Krapfenbauer, K. *et al.* (2003). Aberrant expression of peroxiredoxin subtypes in neurodegenerative disorders. *Brain Res.*, **967**(1-2), 152–160.

Lazareva, O. *et al.* (2021). On the limits of active module identification. *Brief. Bioinform.*, **22**(5), bbab066.

Lev, N. *et al.* (2006). Experimental encephalomyelitis induces changes in DJ-1: implications for oxidative stress in multiple sclerosis. *Antioxid. Redox Signal.*, **8**(11-12), 1987–1995.

Liu, J. *et al.* (2020). Effects of peroxiredoxin 2 in neurological disorders: a review of its molecular mechanisms. *Neurochem. Res.*, **45**(4), 720–730.

Lutomski, D. *et al.* (1997). Anti-galectin-1 autoantibodies in serum of patients with neurological diseases. *Clin. Chim. Acta*, **262**(1-2), 131–138.

Mahmoudian, E. *et al.* (2017). Thioredoxin-1, redox factor-1 and thioredoxin-interacting protein, mRNAs are differentially expressed in multiple sclerosis patients exposed and non-exposed to interferon and immunosuppressive treatments. *Gene*, **634**, 29–36.

Martínez-Yélamos, A. *et al.* (2001). 14-3-3 protein in the csf as prognostic marker in early multiple sclerosis. *Neurology*, **57**(4), 722–724.

Paudel, Y. N. *et al.* (2019). High mobility group box 1 (HMGB1) protein in multiple sclerosis (MS): Mechanisms and therapeutic potential. *Life Sci.*, **238**, 116924.

Paudel, Y. N. *et al.* (2020). Potential neuroprotective effect of the hmgb1 inhibitor glycyrrhizin in neurological disorders. *ACS Chem. Neurosci.*, **11**(4), 485–500.

Pennisi, G. *et al.* (2011). Redox regulation of cellular stress response in multiple sclerosis. *Biochem. Pharmacol.*, **82**(10), 1490–1499.

Ramirez Hernandez, E. *et al.* (2020). The therapeutic potential of galectin-1 and galectin-3 in the treatment of neurodegenerative diseases. *Expert Rev. Neurother.*, **20**(5), 439–448.

Raninga, P. V. *et al.* (2014). Cross talk between two antioxidant systems, thioredoxin and DJ-1: consequences for cancer. *Oncoscience*, **1**(1), 95.

Satoh, J.-i. *et al.* (2004). The 14-3-3 protein ε isoform expressed in reactive astrocytes in demyelinating lesions of multiple sclerosis binds to vimentin and glial fibrillary acidic protein in cultured human astrocytes. *Am. J. Pathol.*, **165**(2), 577–592.

Starossom, S. C. *et al.* (2012). Galectin-1 deactivates classically activated microglia and protects from inflammation-induced neurodegeneration. *Immunity*, **37**(2), 249–263.

van Horssen, J. *et al.* (2010). Nrf2 and DJ1 are consistently upregulated in inflammatory multiple sclerosis lesions. *Free Radic. Biol. Med.*, **49**(8), 1283–1289.

Veto, S. *et al.* (2010). Inhibiting poly (ADP-ribose) polymerase: a potential therapy against oligodendrocyte death. *Brain*, **133**(3), 822–834.

Voigt, D. *et al.* (2017). Expression of the antioxidative enzyme peroxiredoxin 2 in multiple sclerosis lesions in relation to inflammation. *Int. J. Mol. Sci.*, **18**(4), 760.