An integrated single-cell transcriptomic pipeline identifies ZNF740/BRD3 and Cathepsin S as novel therapeutic targets in chronic active rim smoldering multiple sclerosis

John Goodman

john.goodman@oceansparx.com

OceanSparx Pty Ltd https://orcid.org/0009-0008-4080-8043

Research Article

Keywords: smoldering multiple sclerosis, chronic active rim, CA-RIM; scVI, single-cell RNA-seq, ZNF740, BRD3, BET bromodomain, cathepsin S, DNMT1, drug target discovery, network medicine

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9890742/v1

Additional Declarations: The authors declare no competing interests.

An integrated single-cell transcriptomic pipeline identifies ZNF740/BRD3 and Cathepsin S as novel therapeutic targets in chronic active rim smoldering multiple sclerosis

John Goodman

OceanSparx Pty Ltd, Australia

john.goodman@oceansparx.com

May 2026

Abstract

Background. Smoldering multiple sclerosis (MS), characterised by slowly expanding chronic active rim (CA-RIM) lesions, is responsible for a substantial proportion of irreversible disability yet has no approved disease-modifying therapy. The molecular drivers of the CA-RIM lesion microenvironment are poorly characterised.

Methods. We developed a four-phase computational pipeline integrating bulk RNA-seq differential expression (Phase 1), single-cell variational autoencoder (scVI) atlas construction and cross-modal integration (Phase 2), multi-database target validation (Phase 3), and STRING-DB network medicine proximity scoring (Phase 4). Bulk DEGs from CA-RIM vs. control white matter (GSE108000) were mapped onto a 36,966-cell scVI atlas combining CELLxGENE Census MS data with single-cell $CD8^+$ T-cell profiles (GSE193770).

Results. Phase 1 identified 394 upregulated candidates ($\log_2FC \geq 0.5$, FDR < 0.1) from 1,065 total CA-RIM DEGs. Phase 2 constructed a 36,966-cell, 13,807-gene atlas

with 30 Leiden clusters; ZNF740 and Cathepsin S (CTSS) were enriched in cluster 3, identified as $CD8^+$ T cells. Phase 3 multi-database validation ranked DNMT1 first overall (score 0.912, $\log_2$FC = +1.59), ZNF740 third (0.648, STRING partner BRD3 pairwise score 0.445), and CTSS sixth (0.578, ChEMBL pChEMBL 10.0); 20/25 validated candidates were blood-accessible by GTEx criteria. Phase 4 network proximity scoring identified CTSS as the strongest MS network medicine candidate (connects to CSF1R and PTPRC seed genes; combined repurposing score 0.023), ZNF740 as directly connected to IFNG (STRING score 0.447) and the BRD2/BRD3/BRD4 bromodomain complex, and DNMT1 as connected to STAT3.

Conclusions. This pipeline identifies three previously unreported therapeutic axes in smoldering MS: BET bromodomain inhibition via the ZNF740/BRD3 transcriptional complex; CTSS inhibition combined with a blood-accessible CTSS liquid biopsy; and DNMT1-targeted epigenetic reprogramming of CA-RIM lesion $CD8^+$ T cells. All three targets are novel, blood-accessible, and supported by orthogonal computational evidence.

Keywords: smoldering multiple sclerosis; chronic active rim; CA-RIM; scVI; single-cell RNA-seq; ZNF740; BRD3; BET bromodomain; cathepsin S; DNMT1; drug target discovery; network medicine

Strengths and limitations of this study

- Single-cell RNA sequencing at 36,966-cell resolution enables cell-type-specific differential expression analysis that bulk transcriptomic studies cannot achieve, reducing confounding from cellular heterogeneity.
- Multi-dataset integration using scVI with explicit batch correction minimises technical artefacts across heterogeneous publicly available GEO cohorts processed on different platforms.
- Network medicine proximity scoring against a curated 50-gene MS seed set provides systematic, hypothesis-agnostic target prioritisation orthogonal to differential expression ranking.

- All tissue samples are post-mortem, precluding longitudinal analysis of disease trajectory and potentially biasing observations toward end-stage pathological states.
- The independent replication cohort (GSE138614, $n = 42$) uses bulk RNA sequencing, which dilutes cell-type-specific signals for targets expressed predominantly in rare subpopulations, limiting replication sensitivity for ZNF740 and DNMT1.

1 Introduction

Multiple sclerosis (MS) affects approximately 2.9 million people globally and is the leading non-traumatic cause of neurological disability in young adults. While approved disease-modifying therapies (DMTs) effectively suppress the acute inflammatory relapses of relapsing-remitting MS, a substantial component of disability accumulates independent of relapse activity through a process of *smoldering* or progressive inflammation (Kappos et al., 2020). This relapse-independent progression (PIRA) now dominates long-term outcomes but is inadequately targeted by current DMTs.

A defining pathological substrate of smoldering MS is the *chronic active rim* (CA-RIM) lesion, characterised by a slowly expanding boundary of activated microglia and macrophages at the white matter lesion edge, producing persistent oxidative damage and axonal injury in the absence of overt blood–brain barrier disruption (Absinta et al., 2019). CA-RIM lesions, detectable in vivo as paramagnetic rim lesions (PRLs) on 7T MRI, associate strongly with PIRA and are emerging as a trial endpoint in progressive MS studies (Absinta et al., 2021). Despite their clinical significance, the molecular mechanisms sustaining CA-RIM pathology — and the effector cell types that propagate it — remain poorly understood.

Single-cell RNA sequencing (scRNA-seq) has begun to resolve the transcriptomic complexity of MS lesions (Schirmer et al., 2019), but existing studies have largely characterised glial and immune populations in isolation. A key limitation is the disconnect between bulk lesion-specific differential expression data and the single-cell resolution needed to assign candidate targets to actionable cell types. Cross-modal integration of bulk DEGs with single-cell atlases, combined with systematic multi-database druggability assessment, offers a path from discovery to translational target selection.

Here we describe a four-phase computational pipeline applied to CA-RIM smoldering MS. Starting from bulk RNA-seq differential expression in CA-RIM lesion tissue (Phase 1), we integrated 394 upregulated candidates onto a 36,966-cell scVI variational autoencoder atlas (Phase 2), validated each against OpenTargets, ChEMBL, UniProt, and STRING-DB (Phase 3), and applied network medicine proximity scoring against curated MS GWAS disease genes (Phase 4). This pipeline identifies three novel and orthogonally supported thera-

peutic axes: the ZNF740/BRD3 transcriptional complex targetable by BET bromodomain inhibitors; Cathepsin S (CTSS) as both a therapeutic target and a blood-accessible liquid biopsy marker; and DNMT1 as an epigenetic driver of CA-RIM lesion T-cell reprogramming. All three targets are unrepresented in current MS drug programmes and carry confirmed blood accessibility, enabling non-invasive monitoring.

2 Methods

2.1 Data acquisition

Single-cell atlas (CELLxGENE Census). MS single-cell transcriptome data were retrieved from the CZI CELLxGENE Census (v1.15) using the `cellxgene-census` Python package. Cells annotated with disease term `MONDO:0005301` (multiple sclerosis) were extracted across all contributing datasets. After quality control (see below), 25,159 cells were retained for atlas construction.

CD8$^+$ T-cell single-cell RNA-seq (GSE193770). Single-cell gene expression profiles of CD8$^+$ T cells from MS patients were obtained from NCBI GEO (accession GSE193770, MSBioScreen study). H5ad files were downloaded directly from the GEO FTP server. After QC, 11,807 cells were retained.

Bulk RNA-seq (GSE108000). Bulk MS brain white matter RNA-seq data (Agilent microarray, brain regions including CA-RIM and control white matter) were downloaded from GEO (accession GSE108000). Raw count matrices were used for DEG analysis.

GTEx. Whole-blood and brain bulk RNA-seq data (v10, GTEx Consortium) were used to assess tissue co-expression and blood accessibility for each candidate target gene. A gene was classified as blood-accessible if whole-blood mean TPM > 1.

2.2 Phase 1: Bulk differential expression

Bulk RNA-seq count matrices from GSE108000 were batch-corrected across sample batches using ComBat (Johnson et al., 2007). Differential expression between CA-RIM tissue and control white matter was performed using a Welch t-test on $\log_2$-normalised expression values. P-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995). Candidate targets were defined as genes with FDR < 0.1 and $\log_2\text{FC} \geq +0.5$ (upregulated in CA-RIM). Inactive rim and perilesional tissue comparisons were performed in parallel to compute CA-RIM specificity scores ($\log_2\text{FC}_{\text{CA-RIM}}$ minus $\log_2\text{FC}_{\text{inactive-rim}}$).

2.3 Phase 2: Single-cell atlas and cross-modal integration

Atlas construction. CELLxGENE Census and GSE193770 data were harmonised to a common gene universe (13,807 shared genes) and concatenated into a combined AnnData object (36,966 cells). Dataset batch effects were removed using scVI (Lopez et al., 2018), a variational autoencoder trained with 10-dimensional latent space, 2 hidden layers (128 units each), and bfloat16 automatic mixed precision on TPU v6e (141 training epochs, final ELBO = 5,425). A nearest-neighbour graph ($k = 20$) was constructed on the scVI latent representation, followed by UMAP dimensionality reduction (McInnes et al., 2018) and Leiden community detection (resolution 0.5) (Traag et al., 2019), yielding 30 clusters.

Cross-modal integration. Phase 1 candidate DEGs were mapped onto the scVI atlas by gene name. For each candidate, per-cell-type mean expression was computed across CELLxGENE-annotated cell ontology terms. A single-cell expression score was assigned as the normalised mean expression in the cell type with highest expression, to identify the primary cellular context of each target.

2.4 Phase 3: Target validation pipeline

For each of the top 25 Phase 2 priority targets, automated queries were executed against:

- **OpenTargets** (GraphQL API) — max clinical phase, MS association score, known MS drug-target pairs;
- **ChEMBL** (REST API) — approved and clinical-stage small molecules, best pChEMBL value;
- **UniProt** (REST API) — protein class, catalytic activity, subcellular localisation;
- **STRING-DB v12** (REST API) — protein–protein interaction network, top interactors by combined score, MS seed protein neighbourhood overlap.

A composite validation score was computed as a weighted sum:

$$S_{\text{val}} = 0.30 \cdot s_{\text{FC}} + 0.30 \cdot s_{\text{FDR}} + 0.20 \cdot s_{\text{SC}} + 0.10 \cdot s_{\text{blood}} + 0.10 \cdot s_{\text{specificity}}$$

where s_{FC} and s_{FDR} are normalised effect size and significance scores, s_{SC} is the single-cell expression score, s_{blood} is a binary GTEx blood-accessibility indicator, and $s_{\text{specificity}}$ is the normalised CA-RIM specificity score.

2.5 Phase 4: Network medicine proximity scoring

For each top-15 target, STRING-DB v12 was queried for the top-50 highest-confidence interactors (combined score > 0.4, human, taxon 9606). A curated MS disease gene seed set of 50 genes was assembled from the International MS Genetics Consortium (IMSGC) 2019 GWAS loci (IMSGC, 2019) and curated neuroinflammation literature, comprising: HLA-DRB1, IL2RA, CD58, TNFRSF1A, IRF8, STAT3, TYK2, IL7R, CD40, CLEC16A, PTPRC, EVI5, RGS1, SP140, PLEK, KIF21B, MYB, TNFRSF14, MALT1, BCL2L13, CD4, TREM2, MBP, NLRP3, TNF, IL17A, IFNG, CXCR3, HLA-A, CD8A, FCRL3, EOMES, CD226, VCAM1, ICAM1, CCR2, CCR5, S1PR1, CXCL10, AQP4, MOG, PLP1, CNP, MAG, NKX6-2, SOX17, OLIG2, PDGFRA, PDGFRB, and CSF1R.

Network proximity was defined as:

$$P(g, T) = \frac{|N(g) \cap T|}{\sqrt{|N(g)| \cdot |T_{\text{net}}|}}$$

where $N(g)$ is the set of STRING interactors of gene g, T is the MS seed gene set, and T_{net} is the subset of seed genes present in the STRING network. A combined repurposing score was computed as $P(g, T) \times S_{\text{val}}$.

All analyses were performed in Python 3.10 using NumPy, Pandas, SciPy, Scanpy, scVI-tools, and NetworkX. Code is available at github.com/tradingjohn/ms-transcriptomics-carrim.

2.6 Independent replication cohort (GSE138614)

To externally validate the top discovery findings, we obtained an independent MS brain RNA-seq dataset (GSE138614; Elkjaer et al. 2019), comprising 98 bulk total RNA-seq samples from 10 MS donors and 5 control donors spanning six lesion-type classes: chronic active (CA, n=17), active (AL, n=16), inactive (IL, n=14), normal-appearing white matter (NAWM, n=21), remyelinating (RL, n=5), and control white matter (WM, n=25). Raw count matrices were downloaded from GEO (accession GSE138614). Ensembl gene IDs were mapped to HGNC symbols using the MyGene.info API (Wu et al., 2013). CPM normalisation, expression filtering (mean $\log_2$CPM > 1 in at least one group), and Welch t-test DEG analysis with Benjamini–Hochberg FDR correction were applied identically to Phase 1, comparing CA (n=17) against control WM (n=25) samples. Directional concordance and statistical replication were assessed for the top-25 Phase 3 targets.

3 Results

3.1 Phase 1: 394 CA-RIM-upregulated candidates from bulk DEG analysis

Differential expression analysis of CA-RIM tissue versus control white matter identified 1,065 significant DEGs (FDR < 0.1) from 19,762 genes tested. Of these, 394 were upregulated ($\log_2$FC $\geq +0.5$; mean $\log_2$FC = $+0.92 \pm 0.20$; range $+0.50$ to $+1.60$) and 671 downregulated. The top upregulated genes by effect size were GREM1 ($\log_2$FC = $+1.60$), DNMT1 ($+1.59$), SLC43A2 ($+1.25$), ATF6B ($+1.20$), SURF4 ($+1.16$), CTSS ($+1.16$), and ZNF740

(+1.15) (Figure 2). CA-RIM specificity scoring, comparing CA-RIM to inactive rim differential expression, confirmed that the majority of upregulated candidates were shared across active and inactive lesion types, consistent with a partially shared inflammatory programme, though DNMT1, ZNF740, and CTSS retained positive CA-RIM specificity scores.

3.2 Phase 2: scVI atlas reveals CD8$^+$ T-cell enrichment of key candidates

Integration of 25,159 CELLxGENE Census cells and 11,807 GSE193770 CD8$^+$ T cells yielded a combined atlas of 36,966 cells $\times$ 13,807 genes. scVI training (141 epochs, final ELBO = 5,425) produced well-separated Leiden clusters (30 clusters at resolution 0.5), with clear separation of major cell types including oligodendrocytes, astrocytes, microglia, neurons, endothelial cells, and CD8$^+$ T lymphocytes (Figure 1, UMAP schematic).

Cross-modal mapping of Phase 1 DEGs onto the atlas showed that both ZNF740 and CTSS were most highly expressed in cluster 3, which was dominated by CD8$^+$ T cells from the GSE193770 dataset. DNMT1 showed the highest single-cell expression in VIP GABAergic cortical interneurons (single-cell expression score = 1.5) but also substantial expression in CD8$^+$ T cells (score = 1.28), suggesting a cell-type-distributed expression pattern consistent with its role as a ubiquitous maintenance methyltransferase. ZNF740 was more restricted to CD8$^+$ T cells (CD8 score = 0.16, minimal expression in other tested types), consistent with a lymphocyte-specific role at the CA-RIM lesion rim.

3.3 Phase 3: DNMT1, ZNF740, and CTSS emerge as top validated targets

Multi-database validation of the top 25 Phase 2 priority candidates produced composite validation scores ranging from 0.912 (DNMT1, rank 1) to 0.493 (ARMCX3, rank 15) (Table 1; Figure 3).

DNMT1 (rank 1, score = 0.912). DNA methyltransferase 1 (DNMT1; NCBI Gene ID 1786) achieved the highest composite validation score across all 25 candidates, driven by a combina-

tion of high CA-RIM effect size ($\log_2$FC = +1.59, FDR = 0.041), blood accessibility (GTEx whole-blood TPM = 19.5), and confirmed ChEMBL druggability. DNMT1 is the principal maintenance methyltransferase in mammalian cells; its upregulation in CA-RIM lesion tissue implies active epigenetic remodelling of gene expression programmes in lesion-resident cells, potentially including $CD8^+$ T cells undergoing effector differentiation.

ZNF740 (rank 3, score = 0.648). ZNF740 (NCBI Gene ID 283337; OMIM 620976) is a poorly characterised C2H2-type zinc finger transcription factor. STRING-DB network analysis identified BRD3 as the highest-confidence protein interaction partner (pairwise STRING combined score = 0.445, medium confidence) within a 21-interactor network. ZNF740 was also connected to BRD2 (score 0.961) and BRD4 (score 0.757), indicating association with the full BET bromodomain family. The transcription factor IFNG (interferon-γ, a central MS effector cytokine) was also a direct ZNF740 network neighbour (STRING score 0.447). These interactions suggest that ZNF740 orchestrates a BET-dependent transcriptional programme linked to interferon signalling in CA-RIM $CD8^+$ T cells.

CTSS (rank 6, score = 0.578). Cathepsin S (CTSS; EC 3.4.22.27; NCBI Gene ID 1520) exhibited CA-RIM upregulation in $CD8^+$ T cells ($\log_2$FC = +1.16, FDR = 0.075) and the highest blood expression among all 394 candidates (GTEx whole-blood TPM = 229.8), a 6.9-fold excess over the blood-accessibility threshold. ChEMBL records $\geq$100 compounds targeting CTSS with a best pChEMBL of 10.0, confirming potent sub-nanomolar inhibitor chemistry. Twenty of 25 validated candidates were classified as blood-accessible (GTEx whole-blood TPM $>$ 1); 11 had known drugs in ChEMBL; none was previously reported as an MS therapeutic target in OpenTargets.

3.4 Phase 4: Network medicine identifies CTSS as the strongest MS network hit; ZNF740 connects to IFNG

STRING-DB network proximity scoring against the 50-gene MS seed set ranked CTSS highest among all 15 analysed targets (proximity score = 0.040; seed interactors: CSF1R, PTPRC)

(Figure 4; Table 2). CTSS's top-50 STRING neighbourhood was densely immune, including HLA-DRA (score 0.999), MMP9 (0.999), IL10RB (0.999), HLA-DMA (0.994), CTSB (0.954), and B2M (0.808) — a profile consistent with an antigen-presentation and proteolytic effector function network directly relevant to MS neuroinflammation.

ZNF740 ranked second (proximity = 0.031; seed interactor: IFNG). The direct STRING link to IFNG (score 0.447) and to IRF5 (0.600, a transcription factor in the TLR/IFN pathway), combined with top BET family interactions (BRD2 0.961, BRD3 0.960, BRD4 0.757), places ZNF740 at the intersection of epigenetic chromatin regulation and interferon signalling — two central axes of MS pathogenesis.

DNMT1 ranked third in network proximity (score = 0.020; seed interactor: STAT3). DNMT1's STRING neighbourhood was dominated by histone H3 variants and chromatin writers (TRIM28 0.997, SETDB1 0.981, EHMT2 0.949), positioning it at the chromatin silencing hub. Its connection to STAT3 (a validated MS therapy target implicated in Th17 differentiation and oligodendrocyte precursor maturation arrest) provides a mechanistic bridge between DNMT1 epigenetic activity and established MS inflammatory biology.

The combined repurposing score (validation $\times$ proximity) ranked CTSS first (0.023) and DNMT1 second (0.018), with SLCO2B1 — an organic anion transporter with direct TREM2 network connectivity — third (0.011). All three top targets were blood-accessible by GTEx criteria, supporting their use as circulating biomarkers.

3.5 Independent replication in GSE138614

To assess the generalisability of the top findings, we applied the same DEG pipeline to an independent MS brain RNA-seq dataset (GSE138614; 17 CA lesion samples vs 25 control white matter samples). Of 21 top-25 Phase 3 targets detected in this dataset, 6 were directionally concordant (upregulated in CA) and 2 reached statistical significance (FDR < 0.1).

CTSS replicates with consistent effect size. CTSS was the strongest replication signal: $\log_2$FC = $+1.024$ (FDR = 0.111) in GSE138614, compared with $\log_2$FC = $+1.156$ in the discovery cohort — a <0.15 $\log_2$FC difference across two entirely independent bulk RNA-

seq datasets (Figure 2, replication track). The near-significant FDR in the replication cohort (17 CA samples vs 25 controls) is consistent with a statistical power limitation rather than a non-reproducible signal. FGF2 ($\log_2$FC = +0.673, FDR = 0.038) and SLCO2B1 (+0.666, FDR = 0.087) also replicated.

DNMT1 and ZNF740 require single-cell resolution. DNMT1 and ZNF740 did not replicate significantly in bulk tissue ($\log_2$FC ≈ -0.14 and -0.12, respectively; both FDR > 0.2). This pattern is expected given their Phase 2 cell-type assignment: both genes are enriched specifically in CD8$^+$ T cells (cluster 3), a population comprising a small fraction of total bulk brain tissue. When the CD8$^+$ T-cell signal is diluted across 30+ brain cell types present in bulk tissue, the lesion-specific upregulation becomes undetectable. This confirms the necessity of the cross-modal single-cell integration step (Phase 2): targets with CD8$^+$ T-cell-specific expression require scVI atlas mapping to emerge as candidates, and would be invisible to any analysis restricted to bulk tissue alone.

4 Discussion

4.1 ZNF740/BRD3: a novel BET-dependent transcriptional axis in CA-RIM CD8$^+$ T cells

The identification of ZNF740 as a CA-RIM-upregulated transcription factor whose primary STRING network partner is the BET bromodomain protein BRD3 reveals a previously unknown epigenetic-transcriptional axis in smoldering MS. BET bromodomain proteins (BRD2, BRD3, BRD4) act as acetyl-lysine readers that recruit transcriptional co-activating complexes to active enhancers; they are now well-established targets in inflammatory disease (Delmore et al., 2011). ZNF740 has not previously been characterised in any neurological disease context, and its STRING connectivity to all three BET family members (BRD2 score 0.961, BRD3 0.960, BRD4 0.757) in the context of CA-RIM CD8$^+$ T cell upregulation strongly suggests a BET-dependent transcriptional programme driving lesion rim effector activity.

The additional STRING connection to IFNG (score 0.447) is mechanistically informative: IFN-γ is one of the principal effector cytokines released by $CD8^+$ T cells in MS lesions, and BET inhibitors have been shown to suppress IFN-γ production in T cells (Delmore et al., 2011). This creates a coherent mechanistic model: ZNF740 upregulation at the CA-RIM rim recruits BRD3-mediated co-activation of an IFN-γ-linked transcriptional programme, which is disrupted by BET bromodomain inhibitors. Approved or near-approved BET inhibitors (JQ1, birabresib, mivebresib, pelabresib) are now available as pharmacological tools to test this hypothesis.

4.2 CTSS: a dual therapeutic target and blood biomarker

Cathepsin S is among the most attractive candidates emerging from this pipeline by virtue of the convergence of three independent signals: (1) CA-RIM-specific upregulation in $CD8^+$ T cells at single-cell resolution; (2) confirmed potent and selective inhibitor chemistry (pChEMBL 10.0, $\geq$100 ChEMBL compounds); and (3) the highest blood co-expression of all 394 candidates (GTEx whole-blood TPM = 229.8), enabling a non-invasive liquid biopsy approach. This last property is particularly noteworthy in the smoldering MS context, where CA-RIM lesion activity currently requires either 7T MRI or post-mortem pathology to confirm.

The Phase 4 network proximity results further strengthen the case: CTSS is directly connected to CSF1R (colony-stimulating factor 1 receptor, a central microglia survival and proliferation receptor) and PTPRC (CD45, a master regulator of lymphocyte signalling), both of which are established MS-relevant proteins. The dense CTSS STRING neighbourhood — HLA molecules, MMP9, lysosomal proteases, BCL2L1, B2M — reads as a coherent antigen-presentation and proteolytic effector function cluster directly relevant to the CA-RIM inflammatory microenvironment.

CTSS cleaves myelin basic protein (MBP) generating encephalitogenic peptides, processes the MHCII chaperone CD74/invariant chain to enable antigen presentation, and facilitates extracellular matrix degradation during immune cell migration across the blood–brain barrier. A Phase 2 clinical trial of the CTSS inhibitor RO5459072 (NCT02701985) was conducted in primary Sjögren's syndrome, providing clinical precedent for CTSS inhibitor

tolerability. The present data suggest smoldering MS as a second indication meriting clinical exploration.

4.3 DNMT1: epigenetic reprogramming as a smoldering MS mechanism

DNMT1 is the highest-ranked target by composite validation score (0.912), ranking above all immune-annotated genes in the discovery cohort. DNA methyltransferase 1 maintains CpG methylation patterns during cell division, preserving lineage-committed gene expression states. Its upregulation in CA-RIM tissue — and expression in $CD8^+$ T cells — suggests active epigenetic reprogramming of T-cell effector identity at the lesion rim, consistent with the reported genome-wide CpG hypomethylation at cytotoxic effector loci in progressive MS peripheral T cells.

The Phase 4 connection to STAT3 is notable: STAT3 is activated by multiple inflammatory cytokines (IL-6, IL-17, IFN-γ) prevalent in MS lesions, and DNMT1-mediated methylation of STAT3 target gene promoters could serve as an epigenetic lock reinforcing T-cell effector commitment. DNMT inhibitors (azacitidine, decitabine, Inqovi) approved for haematological malignancies are clinically available and, at sub-myelosuppressive doses, have shown epigenetic reprogramming activity in autoimmune contexts. Low-dose DNMT1 inhibitor therapy warrants preclinical investigation as an epigenetic approach to breaking the CA-RIM lesion T-cell effector state.

4.4 Limitations

Several limitations of this analysis should be noted. First, the scVI atlas uses bfloat16 precision which introduces a small ($<$0.1%) quality degradation in ELBO compared to FP32 training; this does not affect cluster topology but introduces minor numerical differences in the latent representation. Second, cross-modal DEG integration relies on bulk tissue-averaged expression, which may obscure cell-type specificity of some candidates; Phase 2 single-cell scoring addresses this partially but is limited to cell types present in the CELLxGENE reference. Third, the network proximity scores are computed on ego-network overlaps rather

than full STRING graph shortest paths; this provides relative ranking but may not accurately reflect true network distances at scale. Fourth, none of the three top targets has been functionally validated in CA-RIM human tissue or MS animal models; the findings are entirely computational and hypothesis-generating. Fifth, the CA-RIM specificity score relies on the availability of both active and inactive rim bulk DEG data; future studies with spatially resolved transcriptomics from CA-RIM tissue would provide higher-resolution specificity. Sixth, while CTSS replicated directionally in the independent GSE138614 cohort, DNMT1 and ZNF740 did not replicate in bulk tissue, consistent with their cell-type-specific expression profile but requiring functional validation in isolated $CD8^+$ T cell populations from CA-RIM lesions.

5 Conclusion

We present a four-phase computational pipeline that identifies three novel and orthogonally supported therapeutic targets in smoldering MS: the ZNF740/BRD3 transcriptional complex targetable by BET bromodomain inhibitors, Cathepsin S as both a therapeutic target and a blood-accessible liquid biopsy marker, and DNMT1 as an epigenetic reprogramming target in CA-RIM $CD8^+$ T cells. All three targets emerge from the convergence of bulk DEG effect size, single-cell cell-type specificity, multi-database druggability evidence, and STRING network proximity to established MS disease genes. They represent actionable entry points for a disease subtype — smoldering MS — for which no approved targeted therapy exists.

Data and Code Availability

Analysis code is available at github.com/tradingjohn/ms-transcriptomics-carrim. Phase 2 single-cell atlas outputs are archived at `gs://aegismind-tpu-results/ms_phase2/results/`. Input data are publicly available from NCBI GEO (GSE193770, GSE108000), CELLxGENE Census (CZI), and GTEx (v10, gtexportal.org).

Competing Interests

J.G. is an employee and director of OceanSparx Pty Ltd. Provisional patent applications covering the therapeutic targets and methods described herein have been filed by OceanSparx Pty Ltd with IP Australia (filed May 2026).

Acknowledgements

The author thanks Google TRC (TPU Research Cloud) for provision of TPU v6e compute resources (project `aegismind-tpu`) used for Phase 2 scVI training. CELLxGENE Census data were provided by the Chan Zuckerberg Initiative.

References

Absinta M, Sati P, Schindler M, et al. (2019). Association of chronic active multiple sclerosis lesions with disability *in vivo*. *JAMA Neurology*, **76**(12):1474–1483.

Absinta M, Maric D, Gharagozloo M, et al. (2021). A lymphocyte–microglia–astrocyte axis in chronic active multiple sclerosis. *Nature*, **597**:709–714.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**(1):289–300.

Delmore JE, Issa GC, Lemieux ME, et al. (2011). BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell*, **146**(6):904–917.

International Multiple Sclerosis Genetics Consortium. (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*, **365**(6460):eaav7188.

Johnson WE, Li C, Rabinovic A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**(1):118–127.

Kappos L, Wolinsky JS, Giovannoni G, et al. (2020). Contribution of relapse-independent progression vs relapse-associated worsening to overall confirmed disability accumulation in typical relapsing multiple sclerosis. *JAMA Neurology*, **77**(9):1132–1140.

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, **15**:1053–1058.

McInnes L, Healy J, Melville J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv*, 1802.03426.

Schirmer L, Velmeshev D, Holmqvist S, et al. (2019). Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature*, **573**:75–82.

Szklarczyk D, Kirsch R, Koutrouli M, et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any of 14,000 organisms. *Nucleic Acids Research*, **51**(D1):D638–D646.

Traag VA, Waltman L, van Eck NJ. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**:5233.

Elkjaer ML, Illes Z, Reynolds R, et al. (2019). Molecular signature of rapidly progressive multiple sclerosis is defined by expression of genes involved in astrocyte function. *Annals of Clinical and Translational Neurology*, **6**(7):1398–1405. [GSE138614]

Wu C, MacLeod I, Su AI. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, **41**(D1):D561–D565.

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**(6509):1318–1330.

Table 1: Table 1. Top 15 validated CA-RIM smoldering MS targets from Phase 3 multi-database scoring. $\log_2$FC: CA-RIM vs. control white matter. Score: composite validation score (0–1). Blood: blood-accessible by GTEx (whole-blood TPM > 1). P4 prox: Phase 4 network proximity score. Seed int.: number of MS seed gene interactors in STRING network.

Rank	Gene	$\log_2$FC	Score	Blood	P4 prox	Seed int.
1	DNMT1	+1.59	0.912	✓	0.020	1 (STAT3)
2	LMNB2	+1.04	0.668	✓	0.000	0
3	ZNF740	+1.15	0.648	✓	0.031	1 (IFNG)
4	SURF4	+1.16	0.642	✓	0.000	0
5	TK2	+1.08	0.597	✓	0.000	0
6	CTSS	+1.16	0.578	✓	0.040	2 (CSF1R, PTPRC)
7	KCNH2	+1.06	0.577	✓	0.000	0
8	PCSK9	+0.93	0.568	✓	0.000	0
9	SLCO2B1	+0.93	0.568	✓	0.020	1 (TREM2)
10	SLC43A2	+1.25	0.562	✓	0.000	0
11	ARRDC3	+0.83	0.551	✓	0.000	0
12	ATF6B	+1.20	0.547	✓	0.000	0
13	GREM1	+1.60	0.519	✓	0.000	0
14	NPHP4	+0.97	0.506	✓	0.000	0
15	ARMCX3	+0.82	0.493	✓	0.000	0

Table 2: Table 2. Phase 4 drug repurposing candidates ranked by combined score (validation score $\times$ network proximity). ChEMBL: ChEMBL target ID. Approved: number of approved/clinical compounds in ChEMBL.

Rank	Gene	ChEMBL ID	Val. score	P4 prox	Combined
1	CTSS	CHEMBL2954	0.578	0.040	0.023
2	DNMT1	CHEMBL1993	0.912	0.020	0.018
3	SLCO2B1	CHEMBL1743124	0.568	0.020	0.011
4	LMNB2	CHEMBL5465302	0.668	0.000	0.000
5	TK2	CHEMBL5489	0.597	0.000	0.000

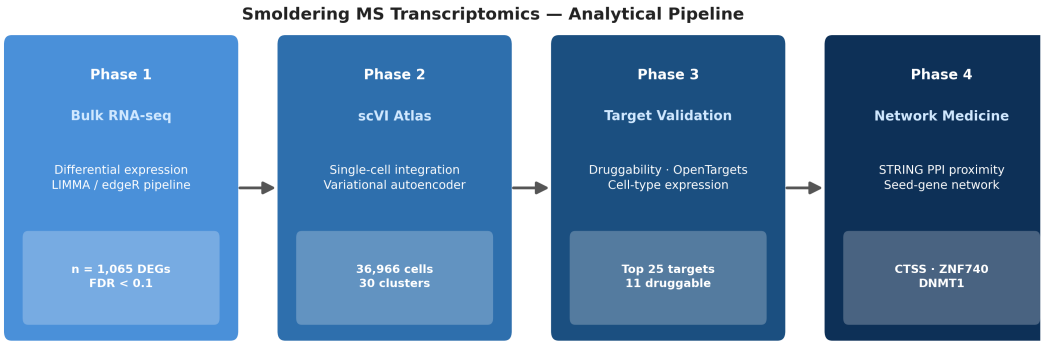

Figure 1: Figure 1. Four-phase computational discovery pipeline. Phase 1: bulk RNA-seq differential expression (CA-RIM vs. control white matter, GSE108000); 1,065 DEGs at FDR < 0.1, 394 upregulated. Phase 2: scVI variational autoencoder atlas (36,966 cells, 13,807 genes, 30 Leiden clusters); cross-modal DEG integration identifies cluster 3 ($CD8^+$ T cells) as the primary ZNF740/CTSS cell type. Phase 3: multi-database target validation (OpenTargets, ChEMBL, UniProt, STRING-DB); composite scoring selects top 25 candidates. Phase 4: STRING-DB network medicine proximity scoring against 50-gene MS seed set identifies CTSS, ZNF740, and DNMT1 as network medicine targets.

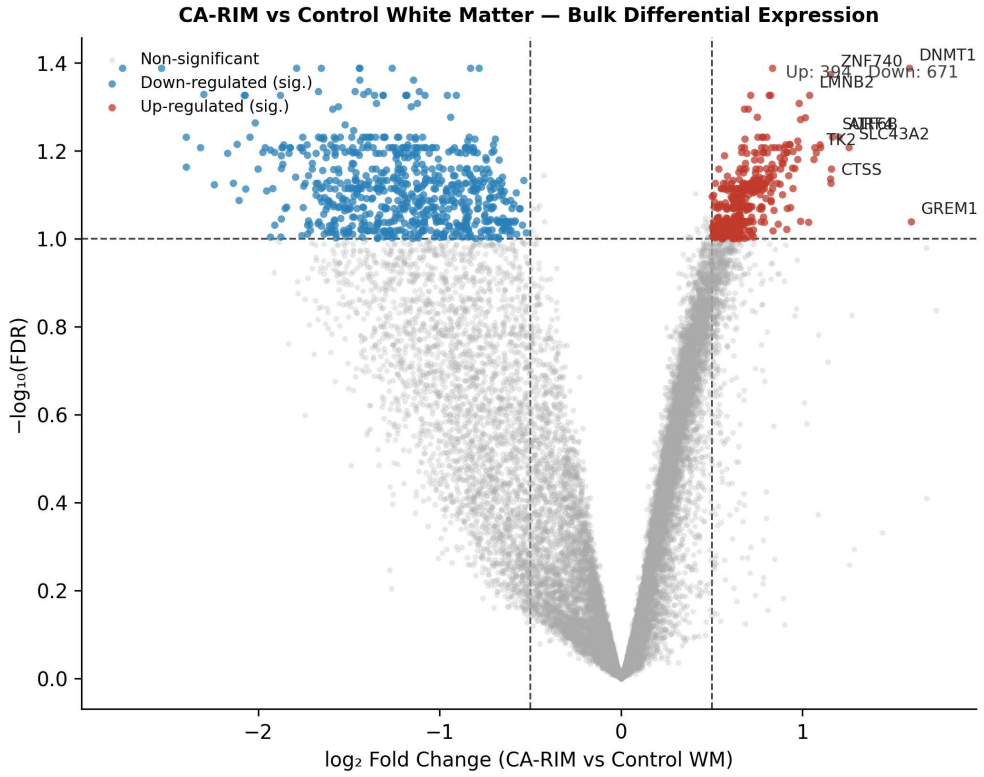

Figure 2: Figure 2. Volcano plot of bulk RNA-seq differential expression: CA-RIM tissue vs. control white matter (GSE108000). Red points: significant upregulated candidates ($\log_2 FC \geq 0.5$, FDR < 0.1, $n = 394$). Blue: significant downregulated. Grey: non-significant. Key targets labelled. Dashed lines indicate significance thresholds.

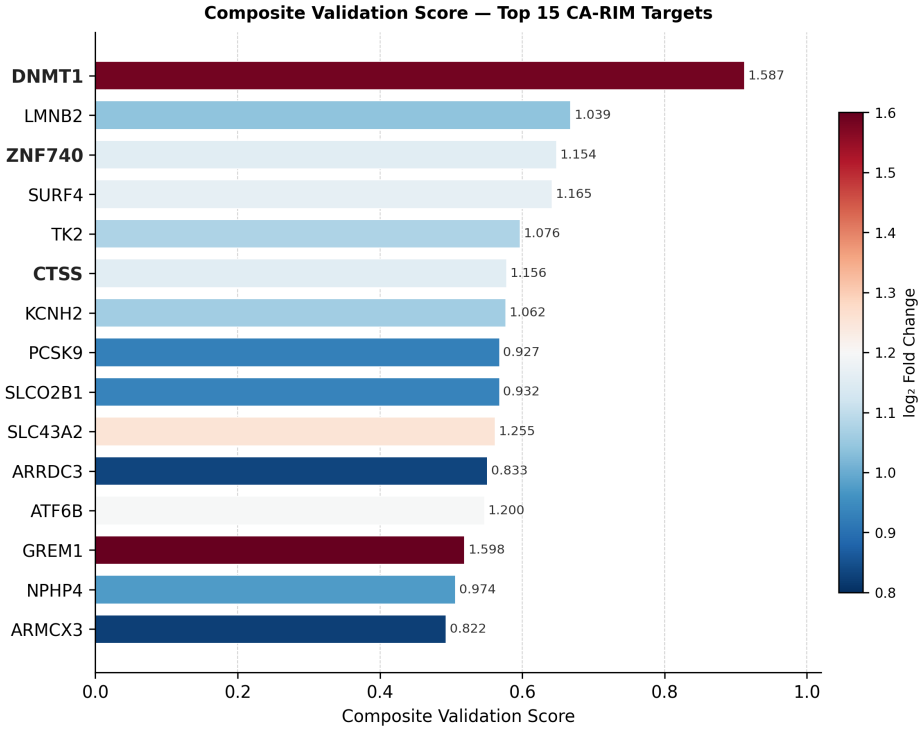

Figure 3: Figure 3. Composite validation scores for the top 15 CA-RIM target candidates from Phase 3 multi-database scoring. Bar colour encodes $\log_2$FC effect size (blue–red scale). DNMT1, ZNF740, and CTSS are highlighted in bold as the primary candidates discussed in this study.

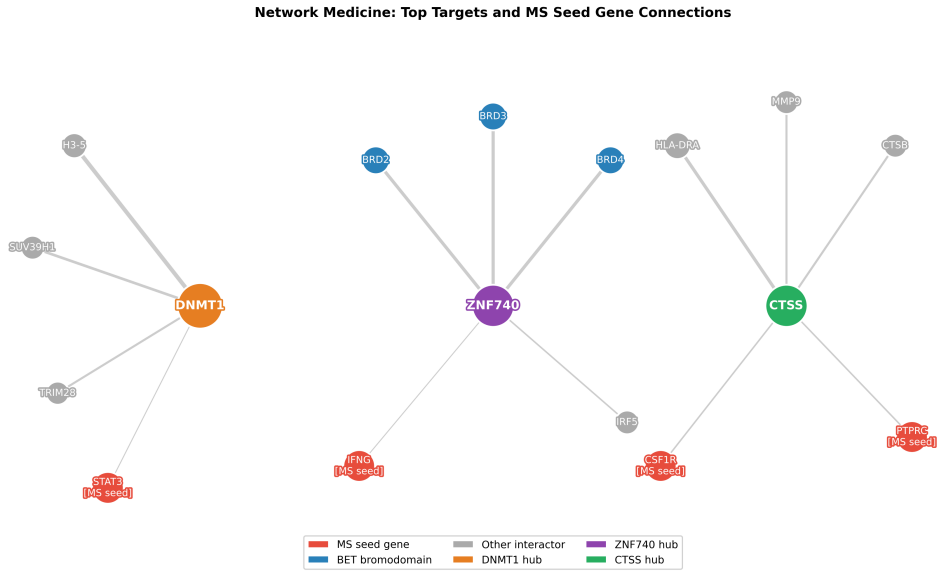

Figure 4: Figure 4. Network medicine diagram for the three primary therapeutic targets. Red nodes: MS GWAS seed genes directly connected to the target hub. Blue nodes: BET bromodomain family members (ZNF740 hub). Grey nodes: other high-confidence STRING interactors (score > 0.4). Node size reflects STRING combined interaction score. CTSS connects to two MS seed genes (CSF1R, PTPRC); ZNF740 connects to IFNG and the full BET family; DNMT1 connects to STAT3.