MethylSPWNet and MethylCapsNet: Biologically Motivated Organization of DNAm Neural Network, Inspired by Capsule Networks

Joshua J. Levy*[1,2,3], Youdinghuan Chen[1,2], Nasim Azizgolshani[2], Curtis L. Petersen[2,4], Alexander J. Titus[5], Erika L. Moen[4,6], Louis J. Vaickus[3], Lucas A. Salas[2,7], Brock C. Christensen[2,7,8]

Author Affiliations:

1. Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
2. Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
3. Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center, Lebanon, NH 03756
4. The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, 03756
5. Department of Life Sciences, University of New Hampshire, Manchester, NH 03101
6. Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
7. Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
8. Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756

* To whom correspondence should be addressed. Contact: joshua.j.levy.gr@dartmouth.edu

Key Words: Deep Learning, DNA Methylation, Central Nervous System Tumors, Capsules, Prior Biological Knowledge, Neural Networks, WGCNA

Abstract

DNA methylation (DNAm) alterations have been heavily implicated in carcinogenesis and the pathophysiology of diseases through upstream regulation of gene expression. DNAm deep-learning approaches are able to capture features associated with aging, cell type, and disease progression, but lack incorporation of prior biological knowledge. Here, we present modular, user-friendly deep learning methodology and software, *MethylCapsNet* and *MethylSPWNet*, that group CpGs into biologically relevant capsules – such as gene promoter context, CpG island relationship, or user-defined groupings – and relate them to diagnostic and prognostic outcomes. We demonstrate these models' utility on 3,897 individuals in the classification of central nervous system (CNS) tumors. *MethylCapsNet* and *MethylSPWNet* provide an opportunity to increase DNAm deep learning analyses' interpretability by enabling a flexible organization of DNAm data into biologically relevant capsules.

Introduction

DNA methylation (DNAm) is a key epigenetic regulator of gene expression in health and disease states, processes of aging and cellular differentiation/stemness, and response to environmental exposures [1–3]. DNAm of cytosine in the context of cytosine-guanine dinucleotides (CpG) sites can be measured with standardized genome-scale oligonucleotide bead-arrays at hundreds of thousands of sites [4,5]. Though a CpG is either unmethylated or methylated, fluorescence signal intensities from array measures of bulk biospecimen DNA are used to derive a beta value measure that approximates the proportion of methylated DNA copies. Gene promoter CpG island methylation is associated with repression of transcription, whereas unmethylated CpG islands are permissive to gene transcription. Alterations to DNAm have a well-established role in

58 carcinogenesis and tumor progression, including inactivation of tumor suppressor genes, aberrant
59 oncogene expression, and loss of repression of repetitive element sequences that contribute to
60 genomic instability [6,7].
61
62 The World Health Organization Central Nervous System (CNS) tumor classification includes
63 over 38 tumor types defined by histopathological features [8]. Most of the 38 can be grouped
64 into the broader glioma, ependymoma, and embryonal tumor types. Within those three
65 categorizations, over 80 further delineations are specified by molecular subtyping. DNAm
66 alterations have been heavily implicated in the development and prognosis of CNS tumors. For
67 instance, epigenetic silencing of *MGMT* is associated with an improved response to
68 chemotherapy in glioblastoma patients through the deactivation of crucial DNA repair
69 mechanisms [9]. *IDH* mutations are associated with improved survival in glioma patients
70 through subsequent global hypermethylation of CpG island promoters, known as induction of the
71 CpG Island Methylator Phenotype (CIMP) [10–13]. Other examples include hypermethylation of
72 Wnt and Shh pathways in medulloblastoma patients [14]. The success of differential methylation
73 analyses in characterizing CNS tumors has recently led to the development of DNAm classifiers
74 of brain tumors as companion diagnostic tools to understand and correctly diagnose challenging
75 histologic cases and for the selection of targeted therapies [8].
76
77 While the development of this methylation-based brain-tumor machine learning classifier has
78 been heralded as an improvement, existing diagnostic framework clinically applicable classifiers
79 use only a small subset of measured CpGs (e.g., 10,000) [15]. Incorporating additional CpG
80 predictors may allow for the resolution of tumor classes otherwise not identified and help

understand relationships with outcomes [16]. This problem may be better approached using machine learning analyses by merit of their prohibitive dimensionality. Deep learning algorithms are a subclass of machine learning approaches that are based on the use of artificial neural networks (ANN) [17–19]. Multi-layer perceptrons (MLP) represent a subclass of neural networks that treat the input data as a one-dimensional vector and then pass the information from one set of nodes to a subsequent set of nodes through fully connected layers of weights/parameters. The information at the subsequent layer of nodes is transformed using non-linear transforms/activations/link functions. These types of analyses are common for deep learning analyses of DNAm data, where the input data is a list of beta values for each subject [20].

DNAm deep learning frameworks, e.g., MethylNet, can accurately characterize tissue, disease states and infer subject age and cell type proportions through unsupervised embedding, generation, classification, and regression tasks [20–24]. They also attempt to ascribe important methylated loci using model interpretability frameworks such as SHAP [25] or LIME [26]. While the inclusion of more CpGs presents an opportunity to expand the space of biologically testable hypotheses [20], statistical challenges (e.g., multicollinearity) with interpretations and generation of associations with pathways remain understudied [27].

Multicollinearity, the unusually high correlation between features, can be addressed with careful feature-selection or grouping [1,28]. Feature selection methods and statistical learning methods such as sparse Group LASSO and network regularization have identified important CpGs in highly complex data [29–33]. More recent work has called for a greater understanding of

104 DNAm-DNAm interactions' implications through the incorporation of Gaussian Graphical
105 Models, Canonical Correlation Analysis, and module discovery through weighted gene co-
106 methylation networks [34–50]. There is growing support for the use of novel deep learning
107 methods to aggregate, group, and select CpGs by their local context (e.g., genes) to connect and
108 interpret the data with clinical outcomes [51–53]. Incorporation of prior biological knowledge
109 improves the transparency and interpretability of the modeling approach and reduces noise while
110 increasing signal by meaningfully pruning redundant relationships between predictors [54].
111
112 Capsule networks have served as inspiration for methods that group CpGs to harness their
113 statistical interactions and relate predictors' groupings to clinical and biological outcomes [27].
114 Capsule networks explicitly model the relationships between constituent parts/groups of
115 predictors, or capsules, through parameterizing pose matrices (unitary transformations) and then
116 hierarchically associate each of these parts independently to higher-order targets of interest.
117 While capsule networks are primarily featured in the computer vision domain, evolving methods
118 within different biomedical specialties often utilize grouped organization of predictors in the
119 neural network design [55].
120
121 Here we provide a deep learning framework for methylation data that draws inspiration from
122 capsule networks. We investigated the organization of CpG features into DNAm capsules, which
123 represent local contexts that can be related to one another. *MethylSPWNet* and *MethylCapsNet*
124 organize sets of CpGs into a series of capsules defined by higher-order genomic contexts and
125 performs classification tasks (Figure 1A, B). To bring additional interpretability to existing deep
126 learning approaches while capturing hierarchical association networks, we propose and explore

MethylSPWNet (Sparse Pathway Network) and *MethylCapsNet* as deep learning analogs of these enrichment tests. We provide recommendations for developing these capsule and network-deriving models and provide open-source software for training these models. The *MethylCapsNet* framework proposes to expand the broad utility of these tools by allowing end-users to construct their unique capsules that represent an array of biologically plausible contexts that further explain their target of interest.

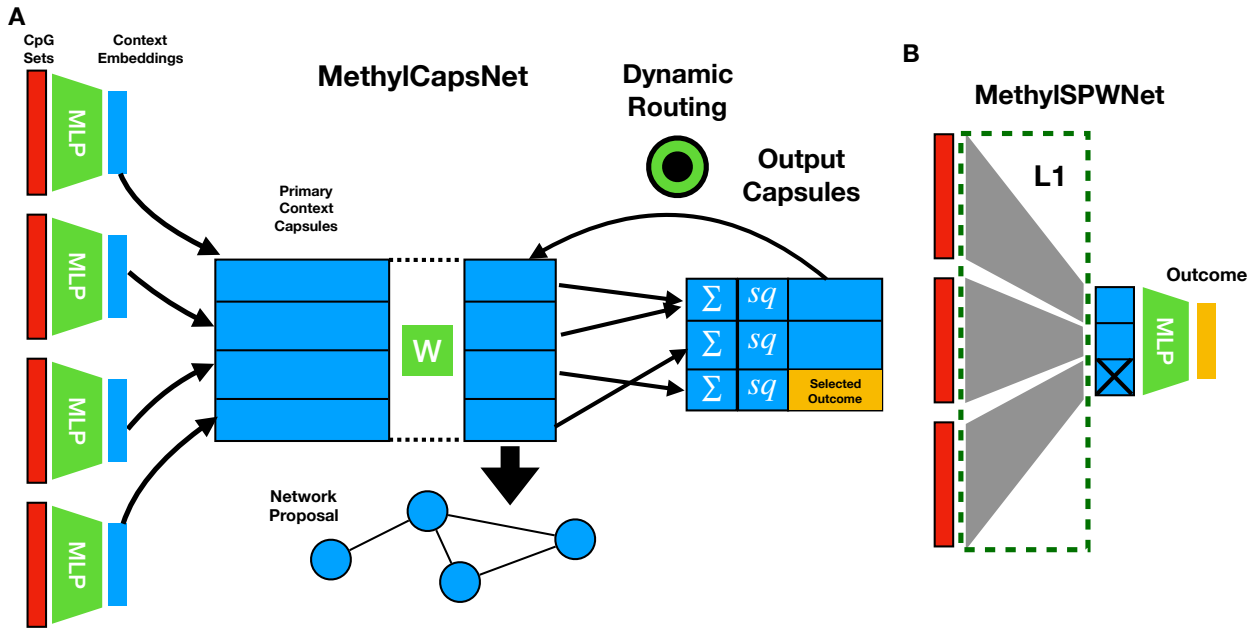


Figure 1: Description of Modeling Approaches: A) *MethylCapsNet* separate MLPs are utilized to form capsule level embeddings from beta values of CpG groupings, which are then associated with outcome targets of interest through dynamic routing of information; these embeddings may be studied to propose networks based on co-methylation on the individual level; B) *MethylSPWNet* aggregates the beta values of groups of CpGs through one locally connected layer; nodes of the resulting layer represent biologically meaningful units that are passed through an MLP for final prediction; group L1 penalties are applied to prune genes/capsules potentially unrelated to the outcome; red colors indicate beta values for CpGs, which serve as input to the algorithm

Results

145 To illustrate the potential utility of capsule-inspired neural network approaches, we revisited the
146 Capper et al. [8] dataset used to train a model that differentiates CNS tumors [56]. CNS tumor
147 histology is largely characterized by the presence or absence of morphologically distinct cells of
148 origin, including neuronal, astrocytic, microglial, oligodendrocytic, and Schwann cells. We
149 aimed to predict the 38 histological subtypes of CNS tumors (39 classes, including controls) as a
150 test case for the capsule-inspired neural network approaches. While distinct cell types may
151 characterize these histological subtypes, it was not our aim to classify these cell types through
152 this modeling approach, as methods for brain cell type estimation using DNAm data are still
153 under development. We compare the *MethylCapsNet* and *MethylSPWNet* frameworks for capsule
154 organization with the existing *MethylNet* framework (which does not account for Capsule
155 organized information [20]) and a Random Forest model fit on 10k important CpGs derived
156 using a previously established method (*Random Forest 10k*). Additionally, we provide a Random
157 Forest model on the capsule-organized information extracted from *MethylSPWNet* (*Random*
158 *Forest Capsules*). Additional details of modeling approaches, fitting procedures, and capsule
159 selection are in the Methods section.
160

161 **Capsule Generation for CNS Tumor Prediction**

162 Capsules may be supplied to the neural network approaches in the form of annotations and/or
163 gene sets from MSigDB and GSEA: 1) genes, 2) sites upstream/downstream of the gene, 3) the
164 following Illumina methylation array annotations – UCSC_RefGene_Name,
165 UCSC_RefGene_Accession, UCSC_RefGene_Group, UCSC_CpG_Islands_Name,
166 Relation_to_UCSC_CpG_Island, Phantom, DMR, Enhancer, HMM_Island,
167 Regulatory_Feature_Name, Regulatory_Feature_Group, and DHS – and 4) the following GSEA

gene sets: C5.BP, C6, C1, H, C3.MIR, C2.CGP, C4.CM, C5.CC, C3.TFT, C5.MF, C7, C2.CP, C4.CGN. Importantly, users can also input custom capsules into the pipeline through a dictionary that maps CpG to a context name of choice. Finally, capsule generation has been integrated with BedTools [57] (*genomic_binned* selection), which can break up the entire hg19 genome into overlapping windows of fixed width. CpGs in these windows will belong to these capsules. We utilized gene capsules for the primary classification study, though alternative methods for capsule formation are explored in section "Exploration of Alternative Capsule Formations and Cancer Subtypes."

Classification Study

Table 1: Classification Results for Random Forest Approaches, MethylNet, MethylSPWNet, and MethylCapsNet*

Approach	*Accuracy±SE*	*Recall±SE*	*Precision±SE*	*F1-Score±SE*
Random Forest 10k	0.96±0.007	0.98±0.012	0.93±0.013	0.95±0.012
Random Forest Capsules	0.94±0.0087	0.97±0.016	0.91±0.017	0.93±0.016
MethylNet	0.97±0.0061	0.99±0.0099	0.96±0.011	0.97±0.011
MethylSPWNet	0.98±0.0049	0.96±0.0088	0.96±0.0085	0.96±0.0089
MethylCapsNet	0.98±0.0044	0.98±0.012	0.98±0.01	0.98±0.011

*All scores are reported as macro-measures across all subtypes; 95% confidence intervals estimated using 1000-sample non-parametric bootstrap

We trained each of the modeling approaches to differentiate 38 histological subtypes of CNS tumors and compared their classification performance via a 1000-iteration non-parametric bootstrap of F1-scores over the test set, which balances sensitivity and specificity and reduces the bias in output. Our results indicate that *MethylNet*, *MethylSPWNet*, and *MethylCapsNet* can achieve very similar high performance on a common data set (Table 1). The neural network

187 approaches achieved marginally better performance than the Random Forest approaches. A
188 breakdown of classification scores for the capsule-inspired models has been included in the
189 supplementary material (Supplementary Table 1). Since all three neural network approaches
190 offer similar performance on classifying brain tumors, we next sought to uncover overlap or
191 complementary insights provided by each modeling approach based on their data organization.
192 The high predictive accuracy of both capsule approaches provided grounds for exploring the
193 factors related to its decision-making process for increased transparency and validation of our
194 approach.
195

196 **Clustering Gene-Level Brain Cancer Embeddings**

197 Until this point, our unit of analysis has been individual CpGs. Summarizing gene-level
198 methylation using median or mean methylation is generally not appropriate. The relationship
199 between methylation state and gene expression can vary depending on the genomic context (e.g.,
200 promoter and gene body). However, while training *MethylSPWNet* to predict tumor histological
201 subtype, the model learns to generate gene-level summaries of methylation by updating the
202 weight of each CpG when aggregating beta values of CpGs on the gene-level (See Methods
203 "Description of MethylSPWNet"). This gene-level aggregation can transform a design matrix of
204 samples by CpGs into samples by genes. Gene-level embeddings correlated with outcome (here
205 tumor histological subtype), can then be interrogated for their relationship with known pathways
206 and gene networks. To visualize gene-level embeddings, we generated cluster heatmaps, where
207 rows constitute observations and columns comprise genes, showing plots of the top 2000
208 variable genes from neural network gene-level embeddings (Supplementary Figures 1-2).
209

210 To assess the representation capacity of MethylNet, MethylCapsNet, and MethylSPWNet
211 embeddings, we clustered embeddings with histologic tumor subtype, cell of origin, and
212 histological subtype with the molecular subclass (Supplementary Table 2). Preliminary
213 clustering of the observations demonstrates, for instance, the inability to differentiate *IDH*
214 mutant subtypes of glioma when defined by median methylation versus the neural network
215 parameterization. There is observed concordance between hierarchical clustering in this
216 embedding space and the brain cancer subtypes. This concordance is defined by their molecular
217 subtypes, the original histological subtypes that the model was trained on, or higher-order cells
218 of origin (e.g., mesenchymal, ependymal, neuro-glial origin). *MethylSPWNet* had the highest
219 degree of concordance with histological and molecular subtypes within the gene-level
220 embedding space [58] (V-Measure 0.72±0.0059; Supplementary Table 2). The ability to
221 recapitulate relevant histological subtypes of CNS tumors through the embeddings alone is
222 further corroborated by embedding plots [58]. In these, MethylCapsNet appears to generate the
223 best separation and differentiation of subtypes (Silhouette Score: 0.52±0.0048), followed by
224 MethylNet (Silhouette Score: 0.25±0.01), and MethylSPWNet (Silhouette Score: 0.1±0.0087),
225 estimated using a 1000-sample non-parametric bootstrap. Since these three approaches were
226 trained to recognize histological subtypes, the signal of the cell of cancer origin and molecular
227 subtypes were less well captured.
228

229 **Gene-Level and Modularity Enrichment Analyses**

230 Next, we aimed to evaluate the utility of the group-regularized deep learning approach for
231 capsule-organized summaries of DNAm on the gene-level for pathways and gene network
232 analyses. We performed a preliminary analysis of pathway and module detection based on the

extraction of hypervariable genes across the neural network embeddings and Louvain clustering of networks of genes based on the pairwise correlation between the genes. Further description of the methods and results is provided in the supplementary material (Supplementary Table 3; Supplementary Figures 3-4).

A description and flow-chart showing an overview of methods for pathways and gene network analysis downstream from *MethylSPWNet* can be found in the Methods section "Description of Potential Downstream Analyses." We focused our presentation of results on three specific CNS tumor subtypes: glioblastoma (GBM), low-grade glioma (LGG), and medulloblastoma (MB). Gene-level embeddings (gene by sample) and pathways and gene network analysis results (derived from those embeddings) are shown in Figures 2, 3, and 4, respectively. Results on pathways and gene network analyses for these three CNS tumor subtypes (GBM, LGG, and MB) are provided in additional files 1-3. A description of the additional files may be found in the supplementary materials (section "Description of Additional Files").

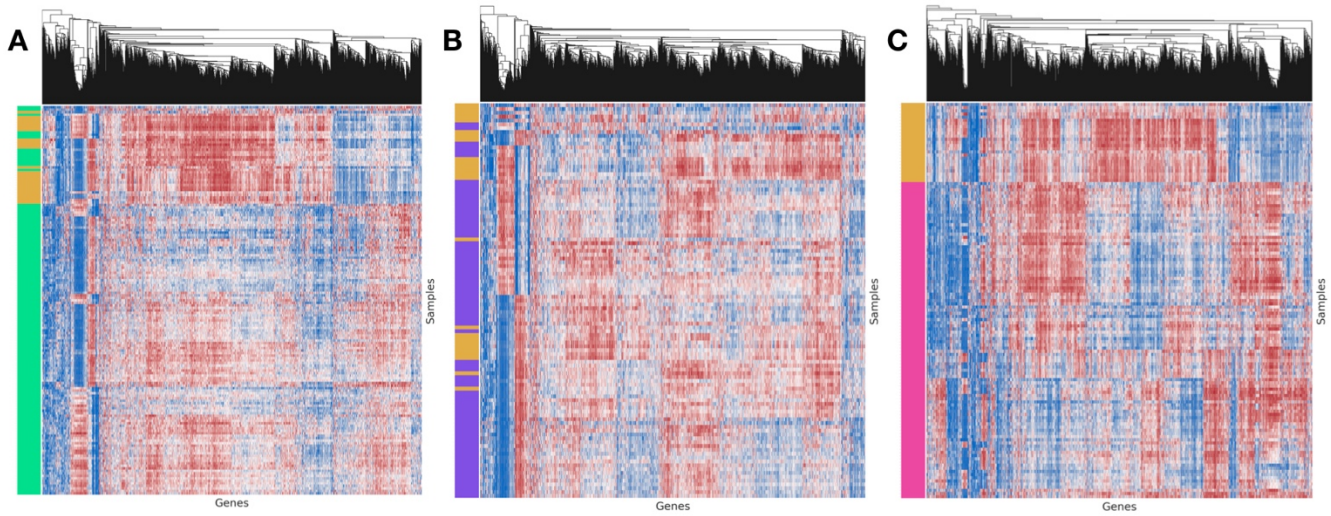


Figure 2: Clustermap of gene-level embeddings for 2000 most variable genes in: A) GBM; B) LGG; C) MB. The left color track indicates the presence of the subtype; yellow indicates the presence of

controls; green, purple, and pink indicates the presence of the GBM, LGG, and MB subtypes, respectively; columns have been standardized to highlight trends; in the heatmap, red indicates high *MethylSPWNet* embedding values, while blue indicates low embedding values

Subtype-Specific Pathways Discovery

Using the gene-level *MethylSPWNet* embedding values, we sought to calculate differentially embedded genes between disease and controls (empirical Bayes) and determine the associations of these genes with some of their correspondent pathways (see Methods "Description of Potential Downstream Analyses"). A visualization of pathway networks and output of gene-subtype associations and pathway enrichment statistics are provided in Figure 3. The empirical Bayes results of the differential embedding analyses for each subtype are provided as Additional Files.

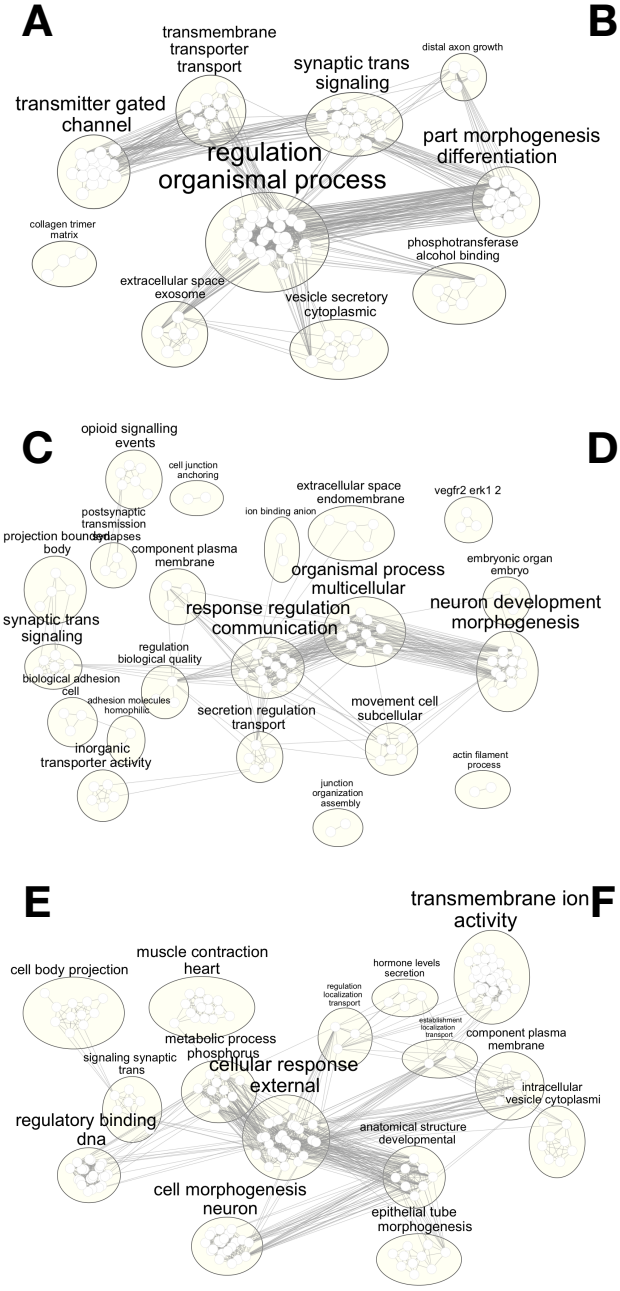
Our analysis of the selected three CNS tumor subtypes (GBM, LGG, and MB) found many top differential genes have been implicated for these subtypes in prior literature. For instance, *TACC* and *FGFBP2* (Figure 3B) are differentially embedded between GBM and controls and have been implicated with a tumorigenic gene fusion event (*TACC-FGFR3* [59–61]). *RASGRF2* (Figure 3B) has been linked to congenital GBM [62]. *LGI2* and *NPY5R* (Figure 3B) have both been related to changes in Sox2 expression [63], which promotes cellular plasticity in GBM. An interesting associated pathway for GBM was type 1 diabetes mellitus (Figure 3A-B), an autoimmune disease, of which its spurious association could be related to found associations with several immune evading markers. For LGG, *TLK1* (Figure 3D), a serine-threonine kinase associated with replication, focal adhesion, and cell cycle, has previously been implicated in gliomas [64]. Similarly, low *ELL2* expression (Figure 3D), regulated by microRNA (miRNA)-mediated gene silencing, was reported to be a marker for poorer survival in GBM patients [65].

276 *GNL1* (Figure 3D) was found in our analysis to be associated with LGG and has been identified
277 as being related to cell proliferation, given its role in the phosphorylation of Rb [66]. We also
278 found associations with the Opioid-signaling pathway and G alpha (i) signaling events [67], and
279 the tyrosine kinase receptor pathway VEGFR (Vascular Endothelial Growth Factor Receptor)
280 and downstream signaling pathway ERK (Figure 3C-D), largely involved with proliferation and
281 angiogenesis [68]. Regarding MB, as examples, we uncovered *NRBP2* (Nuclear Receptor
282 Binding Protein 2; Figure 3F), which had been shown to be downregulated in MB[69], and
283 *SOX14* (Figure 3F), part of the *SOX* family which largely determines cell fate and thus heavily
284 implicated across many CNS tumors[70]. Additionally, pathways such as Muscle Contraction
285 (Figure 3E-F) have been associated with specific molecular subtypes of MB [14,71].
286
287

288 **Grouped-Subtype Pathways Discovery**

289 Additionally, we investigated associations uncovered by grouping together a few select disease
290 subtypes. We would expect, at minimum, differences between these subtypes and healthy
291 controls to be related to pathways that are specialized to those larger histological groupings.
292 First, we compared melanoma-related CNS tumors (MELAN/MELCYT) to controls by
293 performing enrichment analyses of the top 40 differentially embedded genes, as defined by the
294 ranked p-values. As a few examples of potentially enriched gene sets across multiple databases
295 after Bonferroni adjustment, we found potential enrichment for MITF transcription factor targets
296 (*TRRUST*; p=0.06) and neural crest differentiation (*Wikipathways*; p=0.07), BMP signaling (*GO*
297 *Biological Processes*; p=0.03) and IL23-mediated signaling (*NCI-Cancer*; p=0.05). Of interest

from the ependymal tumors (EPN/SUBEPN) was that the top 40 genes had demonstrated overlap with genes related to the spinal cord (*Human Gene Atlas*; p=0.03).



B

Differential Embedding

gene	logFC	AveExpr	t	adj.P.Val	B
TACC3	-1.71	0.394	-23.0	1.30e-47	107.0
LGI2	-1.76	0.368	-20.6	3.32e-42	93.6
RASGRF2	-1.94	0.464	-20.0	5.56e-41	90.4
FGFBP2	-1.62	0.374	-19.8	1.44e-40	89.2
NPY5R	-1.81	0.548	-19.4	7.06e-40	87.3

Pathway Enrichment

Pathway	Description	FDR
REAC:R-HSA-112316	Neuronal System	0.00037
REAC:R-HSA-112315	Transmission across Chemical Synapses	0.00980
KEGG:05033	Nicotine addiction	0.00078
KEGG:04080	Neuroactive ligand-receptor interaction	0.00100
KEGG:04940	Type I diabetes mellitus	0.02200

D

Differential Embedding

gene	logFC	AveExpr	t	adj.P.Val	B
TLK1	1.180	-0.0191	12.5	2.63e-18	38.9
LIMS1	0.991	-0.3830	11.4	2.92e-16	33.5
ELL2	1.340	-0.9720	11.1	9.77e-16	31.9
TSR2	1.160	0.1090	10.9	2.03e-15	30.9
GNL1	1.220	-0.0613	10.7	5.22e-15	29.7

Pathway Enrichment

Pathway	Description	FDR
REAC:R-HSA-112314	Neurotransmitter receptors and postsynaptic signal transmission	3.5e-05
REAC:R-HSA-112316	Neuronal System	1.2e-04
REAC:R-HSA-112315	Transmission across Chemical Synapses	2.5e-03
REAC:R-HSA-111885	Opioid Signalling	1.2e-02
REAC:R-HSA-418594	G alpha (i) signalling events	2.0e-02

F

Differential Embedding

gene	logFC	AveExpr	t	adj.P.Val	B
VPS13B	1.99	-0.289	26.4	1.05e-48	109.0
NRBP2	2.55	-1.200	26.3	1.05e-48	109.0
PROP1	2.03	-0.585	25.8	4.32e-48	107.0
SOX14	2.64	-1.510	24.1	4.03e-45	100.0
TFCP2	2.21	-1.350	23.9	8.13e-45	99.1

Pathway Enrichment

Pathway	Description	FDR
REAC:R-HSA-112316	Neuronal System	9.4e-08
REAC:R-HSA-397014	Muscle contraction	4.4e-05
REAC:R-HSA-5576891	Cardiac conduction	2.4e-04
REAC:R-HSA-112315	Transmission across Chemical Synapses	1.3e-03
REAC:R-HSA-1296071	Potassium Channels	1.1e-02

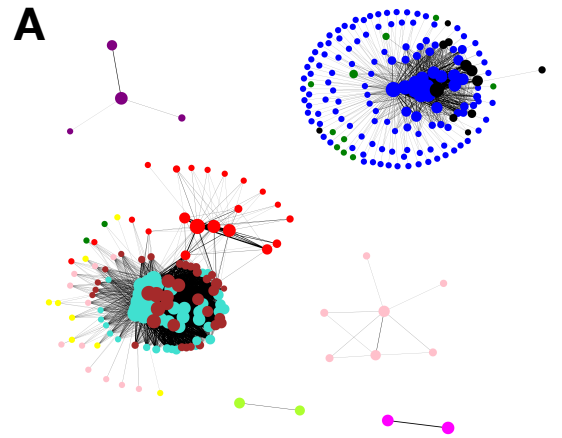
Figure 3: Example Output from Pathway Enrichment Analyses for: A-B) GBM-Control embeddings; C-D) LGG-Control embeddings; E-F) MB-Control embeddings; left-side plots: corresponds to summaries of GO/KEGG/Reactome enrichments using EnrichmentMap in Cytoscape, plotted via RCy3; each small node corresponds to a pathway; node and text size proportional to the number of overlapping genes after differential gene-level embedding analysis; edge between nodes corresponds to shared genes; large ellipses correspond to found clusters using Markov Chain clustering, three words extracted to annotate

307 cluster using a word cloud algorithm; right-side plots: example output of differential embedding (*limma*)
308 and pathway enrichment analyses (g:Profiler) on embeddings; top 5 genes and select pathways listed of
309 the many uncovered in each analysis.
310

311 Derivation of Weighted Gene Co-Embedding Networks

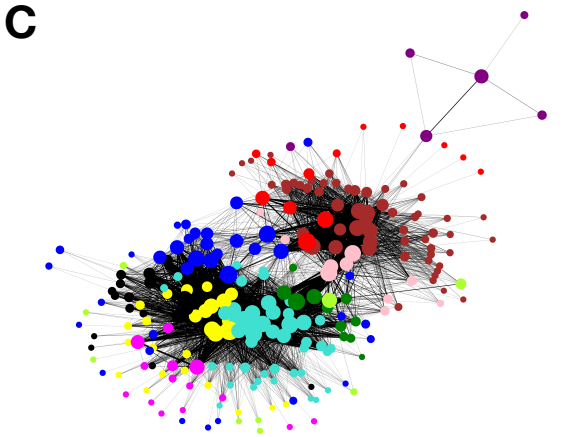
312 To investigate the gene-level embeddings for each of the 38 brain cancer subtypes (paired to
313 normal controls), we derived disease-specific modules of genes using the Weighted Gene
314 Correlation Network Analysis (WGCNA) R package (see Methods "Description of Potential
315 Downstream Analyses") [48].
316
317 We derived 606 modules of genes across the 38 subtypes (37 networks were derived, one
318 subtype was omitted for low sample count), 297 of which were significantly associated with
319 subtype (all P-values < 0.05). We have included as Additional Files the module membership of
320 each of the genes, module expression across the samples for the three example subtypes (GBM,
321 LGG, MB), hub genes for each module (genes located most centrally in each subnetwork), and
322 statistics that relate each module to the subtype. The connectivity of individual genes from the
323 generated WGCNA modules for GBM, LGG, and MB subtypes are shown in Figure 4. Tables of
324 top hub genes from selected modules strongly associated with each subtype are shown.
325
326 Some of the WGCNA modules' hub genes were found to be correspondent with prior knowledge
327 about their respective subtypes. In GBM, *RASGRF4* (blue module; Figure 4A-B) was previously
328 featured in Figure 3B. The silencing of *MGMT* (green module; Figure 4A-B) plays a significant
329 role in the progression of GBM through inactivation of its DNA repair mechanisms [9]. *MIR33B*
330 (brown module; Figure 4A-B) is also related to GBM progression by regulating cell
331 proliferation, invasion, migration, and MYC signaling [72,73]. Finally, the role of platelet factors

332 (*PF4*; blue module) and CpG island hypermethylation (homeobox gene *BARHL2*; turquoise
333 module) has previously been implicated with GBM [74,75]. Examples of hub genes in LGG
334 include *NRP1* (black module; Figure 4D) [76], *PTPRZ1* (pink module; Figure 4D) [77], and
335 *COL6A3* (green module; Figure 4D). *NRP1* has been shown to be related to poor prognosis in
336 gliomas and signals through microglia/macrophages. *PTPRZ1* has previously been related to
337 malignant growth in GBM. Finally, *COL6A3* is a member of genes serving to form the tumor
338 vasculature. Finally, of interest in MB were *SOX 14* and *SOX17* (green module; Figure 4F) [70],
339 *CD4 (*green-yellow module; Figure 4F*), SLIT3* and *GFPT2* (black module, Figure 4F) and
340 *SYNPO* (blue module, Figure 4F). *CD4* is a gene that codes for the membrane glycoprotein of the
341 CD4 T-cell, where its characterization could be corroborated by the immune infiltration patterns
342 of the stroma for MB. *SOX14* and *SOX17* are pertinent for cell fate lineage. *SLIT3* is a gene
343 characterized by axon guidance and consequently tumor growth, migration, and angiogenesis.
344 Interestingly, *GFPT2* (amino acid metabolism), implicated with higher expression and lower
345 GBM survival [78]). *SYNPO*, central to the black module of MB, were also central to the blue
346 module of GBM.



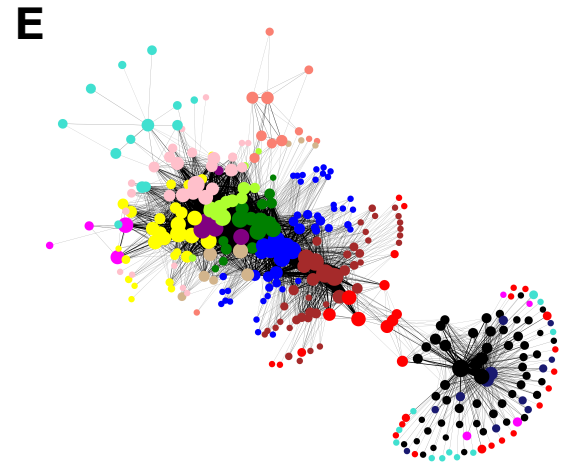
B GBM Significant Module Hub Genes

Description	blue	green	brown	turquoise
p-value	9.5e-37	6.5e-18	2.3e-11	3.5e-12
Correlation with GBM	0.81	0.62	-0.5	-0.52
Hub Gene 1	GFPT2	TNFRSF6B	MIR33B	LYRM4
Hub Gene 2	SYNPO	C3orf20	TSR3	PABPC3
Hub Gene 3	ENPEP	TSLP	LOC285796	LOC402778
Hub Gene 4	PF4	MYOM2	SERF2	BARHL2
Hub Gene 5	LOC285501	LOC220930	ADPRHL1	FAM71F1
Hub Gene 6	LRRC14B	PLA2G2D	ADCK2	DNAH1
Hub Gene 7	TRAM1L1	CDH23	NCS1	GALR2
Hub Gene 8	RASGRF2	MGMT	IL31	KLK6



D LGG Significant Module Hub Genes

Description	pink	green	blue	black
p-value	1.7e-09	1.4e-06	4.8e-08	9.7e-12
Correlation with LGG	0.55	0.45	-0.51	-0.61
Hub Gene 1	PTPRZ1	COL6A3	PABPC3	NRP1
Hub Gene 2	MLST8	FLJ36000	PC	TMPRSS4
Hub Gene 3	MAP10	ATP6V0A4	PALMD	IGDCC4
Hub Gene 4	FCGR2A	METTL9	CTBP1	WFIKKN2
Hub Gene 5	C17orf99	C10orf71	SLC25A18	COL17A1
Hub Gene 6	FOXI2	GABBR2	GFRA4	ADPRHL1
Hub Gene 7	DNAJB8	CTSF	MIR486-1	MELTF
Hub Gene 8	SULT1A2	ERBB4	ASXL2	MIR134



F MB Significant Module Hub Genes

Description	black	midnightblue	greenyellow	green
p-value	7.5e-47	1.9e-23	6.1e-29	2.9e-48
Correlation with MB	0.9	0.75	-0.8	-0.91
Hub Gene 1	GFPT2	GGACT	CD4	SOX17
Hub Gene 2	SLIT3	FOLR1	ST6GALNAC5	R3HDM2
Hub Gene 3	PDCD6	SCGB1D4	HSPB8	LHX1
Hub Gene 4	PCDHGA9	ADGRF2	RASSF9	AIFM2
Hub Gene 5	ENPEP	HEATR4	LIMS2	PPP4R3C
Hub Gene 6	ZAR1	COL7A1	MYPN	MIR196A2
Hub Gene 7	SYNPO	CPED1	LOC399815	SOX14
Hub Gene 8	C5orf52	CHD8	PANX2	ABCA2

Figure 4: Example Output from WGCNA Analyses for: A-B) GBM-Control embeddings; lack of connectivity between the blue/black modules and the other models suggests lack of functional relationship; C-D) LGG-Control embeddings; E-F) MB-Control embeddings; left-side plots: summarized gene-gene networks using the MAPPER algorithm; nodes indicate genes or dense grouping of genes; colored by discovered module via WGCNA; edges indicate co-methylation; right-side plots: example output from WGCNA analysis; the subset of gene-gene modules; p-value indicates whether significantly associated with tumor subtype; correlation value indicates directionality and strength of this relationship; select hub genes indicate genes with the highest centrality in each subnetwork/module.

Enrichment of Neural Network CpGs for Gene and Island Contexts

As Weighted Gene Correlation Network Analysis (WGCNA) identifies significant associations with known pathways and novel gene-gene co-methylation networks, we sought to investigate the CpG-specific parameters that corresponded to producing the embeddings to understand better why the neural network decided to upweight some CpGs, but not others. To elucidate the genomic contexts that *MethylSPWNet* found to be important, we explored the CpG island context and spatial relationship to the transcriptional start site (TSS) (Methods section "CpG Island/Gene Context Analysis"). CpG islands (CGI) are CpG dense regions. Approximately 60% of gene promoters contain CpG islands[79]. CpG shores immediately flank the CGIs by up to 2kb, shelves flank the shores by an additional 2kb, as regional CpG density decreases. Variables for the spatial relationship to the TSS include TSS1500 and TSS200, within 1500 bp and 200 bp of the TSS, respectively. Additional TSS variables are the 5'UTR immediately downstream of the TSS, the first exon, gene body, and 3'UTR.

Of note, we found that CpGs with positive weights (rank-ordered) were depleted for promoter island regions (defined as having TSS1500/TSS200 annotation and not open sea) (OR=0.69; p=0.04) as compared to sites not included in promoter-island regions (OR=1.45; p=0.04). However, when limiting the set of CpGs to only promoter regions (i.e., TSS1500/TSS200), we noted that positive weight CpGs were enriched for island context (i.e., not in an open sea region) (OR=1.21; p=0.03), while negative weight CpGs were depleted for the CGI context (OR=0.89; p=0.02). Furthermore, both the positive and negative weights of intragenic CpGs were depleted for association with the correspondent methylated promoters (as compared to unmethylated promoters) for their respective genes (positive weight OR=0.54, p<0.01; negative weight

380 OR=0.44, $p<0.01$). We have included tables for the relationship between CpG weight and
381 independently considered contexts in the Supplementary Materials (Supplementary Table 4;
382 Supplementary Figure 4).
383

384 **MethylCapsNet Module Enrichment**

385 From the embedding module discovery analysis and further contextualization of the neural
386 network CpG-weights, we observed that information encoded in *MethylSPWNet* corresponds to
387 key pathways associated with various CNS tumors and important genomic contexts.
388 *MethylCapsNet* offers the ability to infer more granular relationships between capsules on the
389 individual sample level when we can reduce the number of parameters specified. The primary
390 capability and emphasis of the capsule-inspired network approach is to compare capsules to each
391 other and directly relate them to particular outcomes of interest via the dynamic construction of a
392 bipartite network (gene-subtype relationships) as part of the training and prediction process. For
393 the *MethylCapsNet* analysis, we pre-selected a subset of genes previously shown to be implicated
394 in various types of brain cancer (see "Selection of Capsules for *MethylCapsNet* and
395 *MethylSPWNet*" in Methods). As such, we believe it would not be appropriate to test for
396 enrichment of these genes due to the pre-selection procedure that introduces a bias. Instead, we
397 derived modules of genes that the neural network deemed to have a coordinated DNAm response
398 in elucidating particular subtypes. Our modularity analysis projects the estimated bipartite graph
399 (gene-subtype) across samples into a univariate graph (gene-gene), then clusters the graph using
400 Louvain modularity to yield four modules of genes (green, red, blue, yellow) (Figure 5).
401

402 Here, we offered an example of the kinds of inferences that can be made from the resultant
403 unipartite network and subsequent clustering. For instance, the yellow module implicates
404 relationships between *WNT3A* and *EGFR*, heavily implicated in Igf and Wnt signaling. The red
405 module features the relationship between *FRZ* and *APC*, both of which are heavily involved in
406 WNT signaling (*APC* forms the complex to inhibit the accumulation of β-*catenin*, while WNT
407 binding to frizzled family receptors may degrade this inhibition and permit cell proliferation
408 [80]). Of the green module, *IDH3G* and *NPR3* (linked to energy metabolism, gene fusion, and
409 chromatin remodeling) were related to both *LRDD* (pro-apoptotic *MAPK* pathway) and *WIF1*
410 (both previously implicated WNT signaling suppressors) from the yellow module. *KIAA1549*,
411 related to astrocytomas and fused to *BRAF* for its progression to oncogenesis, was implicated
412 with *WNT1* in the blue module [81]. Insulin-like Growth Factor Binding Protein 2 (*IGFB2*;
413 glioma oncogene) of the red module appeared to be negatively correlated with many of the genes
414 across the yellow and green modules[82]. *TP53* (which lacked consequential methylation
415 patterns) and *MYC* [83], *MGMT,* and *TERT* share relationships with each other but not with the
416 other modules, perhaps highlighting how ubiquitous these somatic alterations are for
417 oncogenesis.

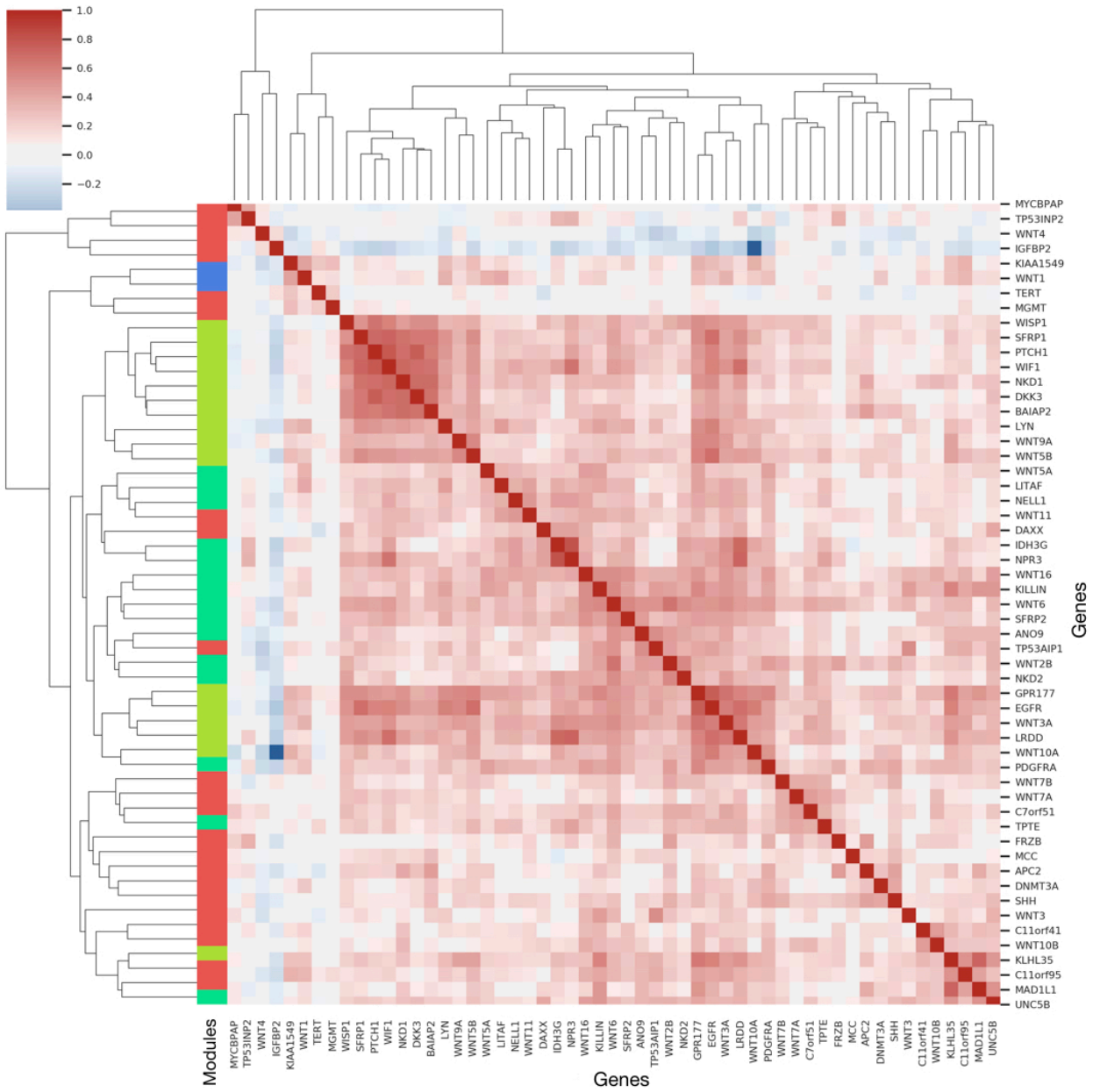


Figure 5: Clustermap of weighted bipartite network projection of select gene capsules' relationship to various CNS tumors using the *MethylCapsNet* modeling approach; rows and columns are genes, row colors denote modules (green, red, blue, yellow) of genes found using Louvain modularity; values in matrix denote Pearson's correlation between genes

Despite having highlighted potentially interesting relationships, we acknowledge that there will be a future opportunity to increase the search space of possible relationships between genes and their coordinated response for bringing about subtypes of brain cancer. In the Supplementary

427 Material, we provide the routing matrix that was used to align each gene to a correspondent CNS
428 subtype, which is the predicted gene-subtype bipartite network averaged across individuals
429 (Supplementary Figure 5). We have also provided a locally deployable web application that
430 allows the user to interrogate their uncovered capsules and form networks on the individual level
431 and aggregated across patient subgroups or disease subtypes using gene-capsule-specific
432 embeddings (Supplementary Figure 6). For instance, in our supplemental material, we
433 demonstrate how the *MethylCaps* web application can be used to derive individual networks for
434 LGG and GBM, identifying *ANO9* and *EGFR,* respectively, as implicated in these conditions
435 (Supplementary Figure 7-8). These genes and their associated pathways have been heavily
436 implicated in tumorigenesis and gliomas [84,85].
437

438 **Exploration of Alternative Capsule Formations and Cancer Subtypes**

439 In this section, we briefly present results from a few of many alternative means of forming
440 capsules. Particularly, we consider the following scenarios for CNS tumor classification: 1)
441 *MethylCapsNet* is fit when a) only half the genes are retained from the original list (selected
442 randomly) and b) none of the genes are retained and instead randomly sampled from all genes
443 and 2) *MethylCapsNet* is fit using binned genomics regions, utilizing CpGs encapsulated in 1Mb
444 bins. In addition to these analyses, we also explore an integrated breast cancer dataset utilized for
445 PAM50 molecular classification[20,86,87] using a few different capsule configurations: 1)
446 *MethylCapsNet* is fit using a curated list of genes, 2) *MethylCapsNet* is fit using binned genomics
447 regions, utilizing CpGs encapsulated in 700kb bins, and 3) *MethylSPWNet* is fit using capsules
448 organized by CpG island promoters, formed by intersecting CpG island with gene promoter

449 annotations from the Illumina 450k database. A complete set of results can be found in the
450 supplementary material (Supplementary Tables 5-6; Supplementary Figure 9).
451

452 **Discussion**

453 Recent reviews and initial explorations discussed the potential utility of capsule-inspired
454 networks to relate biologically-organized capsules to each other and known disease outcomes
455 [27,88–90]. In this work, we set out to perform a preliminary evaluation that shows the
456 feasibility and suitability of DNA methylation capsules for deep learning analyses as a means to
457 organize CpG information to higher-order contexts to improve prediction and transparency while
458 uncovering instances of coordinated gene-level methylation patterns. In our analyses, we
459 compared several state-of-the-art predictive modeling methods for DNA methylation
460 classification of brain tumors. We demonstrated that capsule-based deep learning approaches
461 could achieve performance on par with existing deep learning models and prove better than
462 existing traditional machine learning frameworks for analyzing DNA methylation data. Our work
463 demonstrated the potential for new insights compared to other existing methylation-based tumor
464 classification schemes currently used, which are often based on a small subset of CpGs, and lack
465 built-in interpretation of the loci selected[8,91]. We demonstrated the efficacy of increasing the
466 use of the available CpGs on the Illumina 450k Array, ultimately using 200,000 loci before sub-
467 setting by context.
468

469 DNA methylation capsules focused on the gene level can disentangle important CpGs that might
470 otherwise be down-weighted in a feature-by-feature deep learning unsupervised or supervised
471 learning approach. These CpGs demonstrated substantial overlap with genes known to be related

472 to tumorigenesis in the brain, such as *NOTCH1*, *PTEN*, and *GNAS*. This is consistent with
473 previous studies that demonstrated mutations common in brain tumors, such as *IDH1*, are
474 correlated with disruptions in methylation.[10,92–97]
475
476 The context-specific CpG weight enrichment analyses suggest that within promoter regions,
477 island context is important for differentiating different CNS tumor subtypes, but taken as a
478 whole, regions outside of CpG promoter islands are important for capturing this heterogeneity.
479 Furthermore, outside of the promoter context (supposedly regions that better capture tumor
480 heterogeneity), the ability of intragenic CpGs to distinguish tumor subtypes is still dependent on
481 the promoter methylation status of the respective gene. Clustering of CpGs with the highest
482 weights at CpG islands, shores, gene bodies, and transcription start sites will help us understand
483 where the most diagnostically relevant sites are in the genome, but demands additional
484 investigation.
485
486 We also presented a few examples of the potential downstream applications of capsule-based
487 approaches. In particular, our framework demonstrated the ability to relate derived gene-level
488 measures of *MethylSPWNet* to known disease pathways via differential methylation analysis of
489 the gene-level embeddings and gene-gene co-methylation networks via WGCNA. Additionally,
490 we provided a preliminary interpretation of bipartite (gene-subtype) and unipartite (gene-gene)
491 networks, which can be derived by *MethylCapsNet* web framework. Finally, we explored
492 alternative means from which to form capsules. We expected the curation of genes to lead to
493 more accurate models. Contrary to our initial hypothesis, for CNS classification, random
494 capsules' selection appeared to still produce a highly accurate model. These results suggest either

495 the potential to uncover novel associations between genes and subtype or that these genes may
496 be co-methylated with other genes that have well-established relationships. The binned
497 genomics, fit using *MethylCapsNet* for classification tasks in brain and breast, were similar to the
498 leading methods. The island promoter capsules slightly underperformed, suggesting that these
499 capsules' selection alone does not contain enough information to distinguish PAM50 molecular
500 subtypes.
501
502 There are a few limitations to this work, presenting room for future improvements in the
503 analytical method. First, the included CpG loci selection is biased on limited sites available from
504 the Illumina Array platform. At present, the utilization of an Illumina methylation array platform
505 is more tractable due to lower technology costs and expertise required [98] than Whole Genome
506 Bisulfite Sequencing (WGBS) that requires substantial sequencing depth on the order of 100x for
507 comparable precision [99].
508
509 Second, a reduced set of genes were fit using the capsule-inspired network. It remains
510 challenging to run MethylCapsNet at scale due to the heavy computational demand and the large
511 number of free parameters in its current formulation. Since *MethylCapsNet* can only analyze
512 approximately one thousand capsules at a time, the capsule selection step is critical to the
513 method's successful application. This parameter space should be reduced by finding some
514 marriage between the scalability enabled by *MethylSPWNet* and perhaps greater transparency
515 offered by *MethylCapsNet*. Presently, we advise end-users to utilize *MethylSPWNet* when the
516 number of contexts under evaluation is large (≥1000 capsules), or if the number of CpGs per
517 gene is small, and to utilize *MethylCapsNet* when the number of contexts under consideration is

518 smaller (<1000 capsules). Uncurated gene sets can be analyzed using *MethylSPWNet*, while
519 curated gene sets are best suited to *MethylCapsNet*, e.g., regions of the genome fragmented by
520 consistent windows or larger DNAm CpG modules that have been uncovered through methods
521 such as WGCNA. In addition, the adoption of capsule-inspired approaches that explicitly form
522 networks via their routing mechanisms presents a future area of research [89]. It is also assumed
523 that *MethylCapsNet* capsules that are more closely embedded are interacting, but it is not entirely
524 clear the nature of these interactions without incorporating gene expression data and methylated
525 Quantitative Trait Loci analyses.
526
527 While inspired by capsule networks, we also emphasize that these methods are not analogous to
528 capsule networks featured in computer vision tasks. While the new capsule-based approaches
529 were as accurate as fully connected approaches, this was done so under the constraint of sparse
530 connections, where such specification points to the validity of imposing these constraints. Given
531 the nature of the problem (classification amongst dozens of histological subtypes), intermediate
532 embeddings may reflect a more linearly separable subspace to subtypes of origin. Such a
533 subspace may require additional exploration/penalization to avoid potential biases pertaining to
534 the minimal redundant set of predictors to produce a subspace optimal for prediction. The
535 application of such methods does not preclude the potential for selection of genes due to
536 technical reasons such as noise, batch effects, and weight initializations, which are common to
537 many domains of application of neural networks. We attempted to account for such biases
538 through preprocessing methods on the data such as functional normalization and note that strict
539 interpretation of threshold cutoffs for methods devised for differential gene expression may not
540 be applicable. Thus, relaxation of the scope of features input into a pathway and other

541 enrichment analyses may potentially reduce bias so as long as limitations are appropriately
542 stated. Additionally, differences in tissue preparation (frozen, permanent) were not accounted
543 for, and however, given the high concordance between these preparation methods, we felt that
544 such adjustment was not necessary [100].
545
546 While DNAm deep learning methods with built-in interpretability do not yet exist, we hope these
547 methods, though constrained by potential limitations in design choice, may spur further research
548 into more interpretable capsule methods. Here, there is also an opportunity to further apply
549 concepts from topological data analysis (TDA), such as Mapper[101–105], to distill the key
550 functional relationships from high-dimensional, complex data.
551
552 Existing classification frameworks currently used in the clinical setting for aiding brain cancer
553 diagnosis only utilize a small subset of the total possible set of CpGs that can be measured.
554 Current modeling approaches can be difficult to trust or use to study new network biology until
555 they can consider a larger, more complete set of predictors. However, it is also important to note
556 that doing so would introduce additional noise into the modeling approach, but the incorporation
557 of prior biological knowledge can potentially help reduce noise while improving the detection of
558 biologically relevant signals. We note that underperformance could suggest selecting capsules
559 that may not be optimally aligned to the target task/dataset. By demonstrating that the
560 organization of CpGs into their respective genomic contexts, we present further opportunity to
561 reduce the feature space and disentangle correlation and collinearity between CpG sites to create
562 a new class of transparent, clinically tractable models. For instance, future classifiers should
563 include brain cell-type classification using DNAm data and incorporate it as covariates in the

564 prediction model, yet brain cell-type differentially methylated regions for deconvolution by
565 DNAm patterns is not well-established. The opportunity space of epigenetics research questions
566 is ample and poised to grow substantially as the field moves to expand reference-based
567 approaches to cell type deconvolution, include tandem assessments of other cytosine
568 modifications (hydroxymethylcytosine), and apply DNA methylation age clocks to questions of
569 biological aging. Despite having demonstrated the promising downstream analyses that users
570 may readily adopt through our framework, we acknowledge that there is ample opportunity to
571 develop related methods and their use cases further.
572

573 **Conclusions**

574 Here, we have demonstrated the feasibility and utility of DNAm based capsules for performing
575 disease classification and potentially determining dysregulated genes for these diseases. We
576 found that DNA methylation capsule methods can predict brain cancer subtypes with high
577 accuracy and present convenient means for organizing data over traditional techniques for
578 studying DNA methylation data. As such, we advocate for the organization of well-defined
579 DNAm capsules as a means to improve the accuracy, transparency, and broad applicability of
580 DNAm deep learning models. Future deep learning prognostic models that reimagine the
581 formation and incorporation of DNA methylation capsules, paired with cell type inference, gene
582 expression, and/or corroborating chromatin capture, may serve as grounds for the derivation of
583 unknown heterogeneity.
584

585 **Methods**

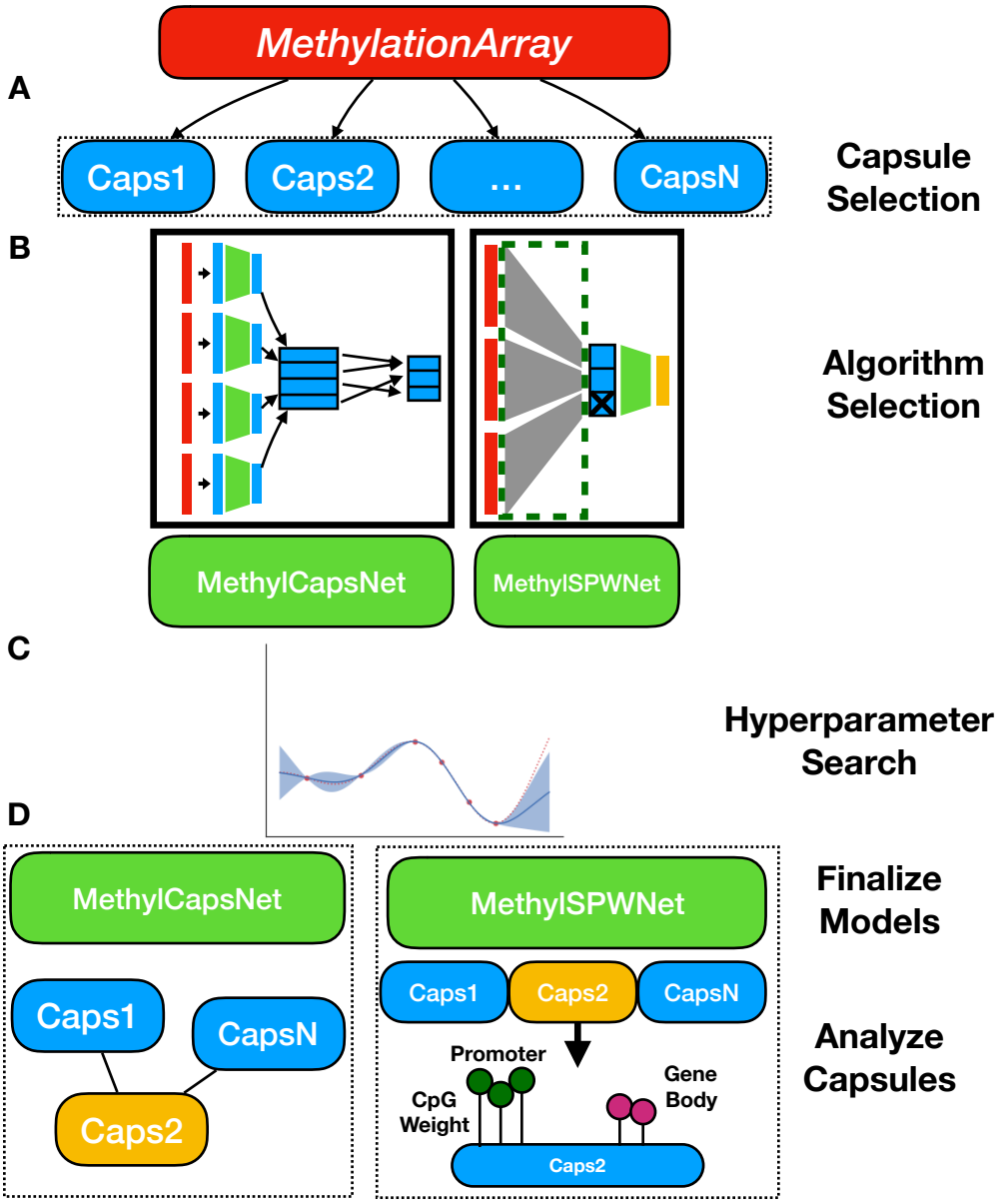
586 **Overview of Framework**

587 The *MethylCapsNet* methodology presents an extension of the *MethylNet* framework [20] and is
588 implemented as a command-line interface that allows the user to group CpGs into capsules and
589 then dynamically route the capsules to make a prediction and interpret the results. While this
590 approach draws inspiration from capsule networks featured in computer vision tasks,
591 *MethylCapsNet* is not explicitly a capsule network, as defined in previous works in this domain.
592
593 *MethylCapsNet* utilizes separate MLPs for every set of CpGs (one set per context) to derive
594 context-specific embeddings (separate context embeddings per each individual), and dynamic
595 routing processes force information from child capsules into disease/categorical outcomes
596 (Figure 1A). The information is hierarchical because each child capsule may only align with one
597 parent capsule. Once the capsule-inspired network is fit, graph structures that describe the
598 relationships between each individual's contexts can be derived by thresholding the correlation
599 between pairwise n-dimensional context embeddings. Highly conserved biological networks can
600 be derived by thresholding the number of individuals that share the same edge between the
601 contexts. The simplest genomic context considered are genes that the CpGs annotate to. Other
602 capsules can be defined by, for instance, genomic region or pathway/biological process
603 annotation. *MethylSPWNet* is a specialized neural network architecture that routes beta values
604 from the CpGs in each context into a single node representing the context (Figure 1B). Each
605 CpG is given a weight based on the importance of its contribution, both on the gene level and
606 towards the classification task as a whole. This information passes through additional neural
607 network layers that dynamically relate latent sets of predictors to outcomes of interest, whether
608 they be prognostic or diagnostic [53,106]. Much like Group LASSO approaches, group L1-

609 penalization can be utilized on the CpG weights routed to each gene to select relevant genes of

610 interest.

611



612

613 **Figure 6:** Description of Framework: a) User selects capsules to group sets of CpGs from their own
614 supplied list or a prespecified annotation set; b) User selects modeling approach, *MethylCapsNet*,
615 *MethylSPWNet*, *Group LASSO*; c) Hyperparameter search conducted to reveal ideal model specification;
616 d) Final models are fit, and capsules are interrogated for relationships with one another, relationships to
617 the outcome, and highly weighted CpGs; yellow indicates an important capsule; *MethylCapsNet* relates

618 the Capsule to each other per individual; *MethylSPWNet* locates important CpGs within a capsule and
619 contextualizes its location; input data into the algorithm is colored red
620

621 The software implementation (Figure 6) is comprised of modules pertaining to prediction and
622 interpretation tasks, which take into account the relationships and embeddings derived through
623 the training process.
624

625 **Data Preparation**

626 DNAm data from CNS tumors (n=3,897) were accessed from the GEO archive (GSE109381),
627 preprocessed using *PyMethylProcess* [21], and divided into 70%/10%/20% training, validation,
628 and testing sets (*MethylationArray* objects) via *PyMethylProcess*. The 200,000 most variable
629 CpG loci across the training samples were retained for analysis. Sets of CpGs were tracked to
630 genes, which were then selected to form capsules. The original set of 200,000 CpGs was used as
631 features for the *MethylNet* approach, the complete set of intersecting gene capsules with more
632 than five associated CpGs were used for the *MethylSPWNet* (n=10,341; 139,028 CpGs), and
633 Group LASSO approaches, and a reduced set of capsules (n=55), were utilized for
634 *MethylCapsNet* after manual curation and a hyperparameter search (see "Selection of Capsules
635 for *MethylCapsNet* and *MethylSPWNet*").
636

637 **Description of Potential Downstream Analyses**

638 After fitting a *MethylSPWNet* model (and *MethylCapsNet*), the user may further interrogate the
639 gene-level embeddings depending on the research question being addressed. The user may
640 explore how each gene relates to each outcome or how they relate to one another, the details of
641 which have been included in an informative flow diagram (Figure 7) (separate text describing

642 information that can be extracted after fitting *MethylCapsNet* can be found in the section
643 "Description of Capsule-Inspired Neural Network").
644
645 Differentially embedded genes (the extent to which gene-level embeddings vary between
646 subtypes versus normal) from gene-level values derived by the neural network embeddings in
647 each of GBM, LGG, and MB, were identified using the *limma* [107] package. This package
648 compares tumor to nontumor control tissue through least-squares regression and empirical Bayes
649 moderated F-tests, yielding FDR-adjusted p-values and log-odds ratios for the degree of
650 differential embedding. We profiled functionally enriched pathways using the g:Profiler package
651 [108] after selection of genes below an FDR-adjusted significance threshold and visualized the
652 results (relating pathways by the number of shared genes, clustering into higher-order pathways
653 via Markov clustering) using EnrichmentMap, as part of the Cytoscape network visualization
654 framework [109,110].
655
656 The pairwise correlation between *MethylSPWNet*-derived gene methylation was calculated using
657 Pearson's correlation coefficient. Weighted adjacency matrices were calculated from the
658 pairwise correlation matrices for each of the subtypes using the power adjacency function, which
659 takes the co-methylation to a power specified separately for each subtype. To further cluster the
660 genes, the weighted adjacency is transformed into a topological overlap matrix (*TOM*), defined
661 by the extent to which two genes share a third common gene. Finally, hierarchical clustering is
662 applied to derive the final modules of genes. Finally, to relate each module with the disease
663 subtype via the aforementioned least squares and empirical Bayes differential analyses methods,
664 we calculated eigengenes (1st principal component) for the genes in each module to further

665 reduce the design matrix (samples by modules) [107]. A large number of genes (on the order of
666 five thousand) for such a summary gene network plot may make plotting the individual genes
667 cumbersome and hard to understand, so we utilized Mapper[101,103,104], a tool from
668 Topological Data Analysis, to further summarize and portray the relationships between the genes
669 in the network summary plot.

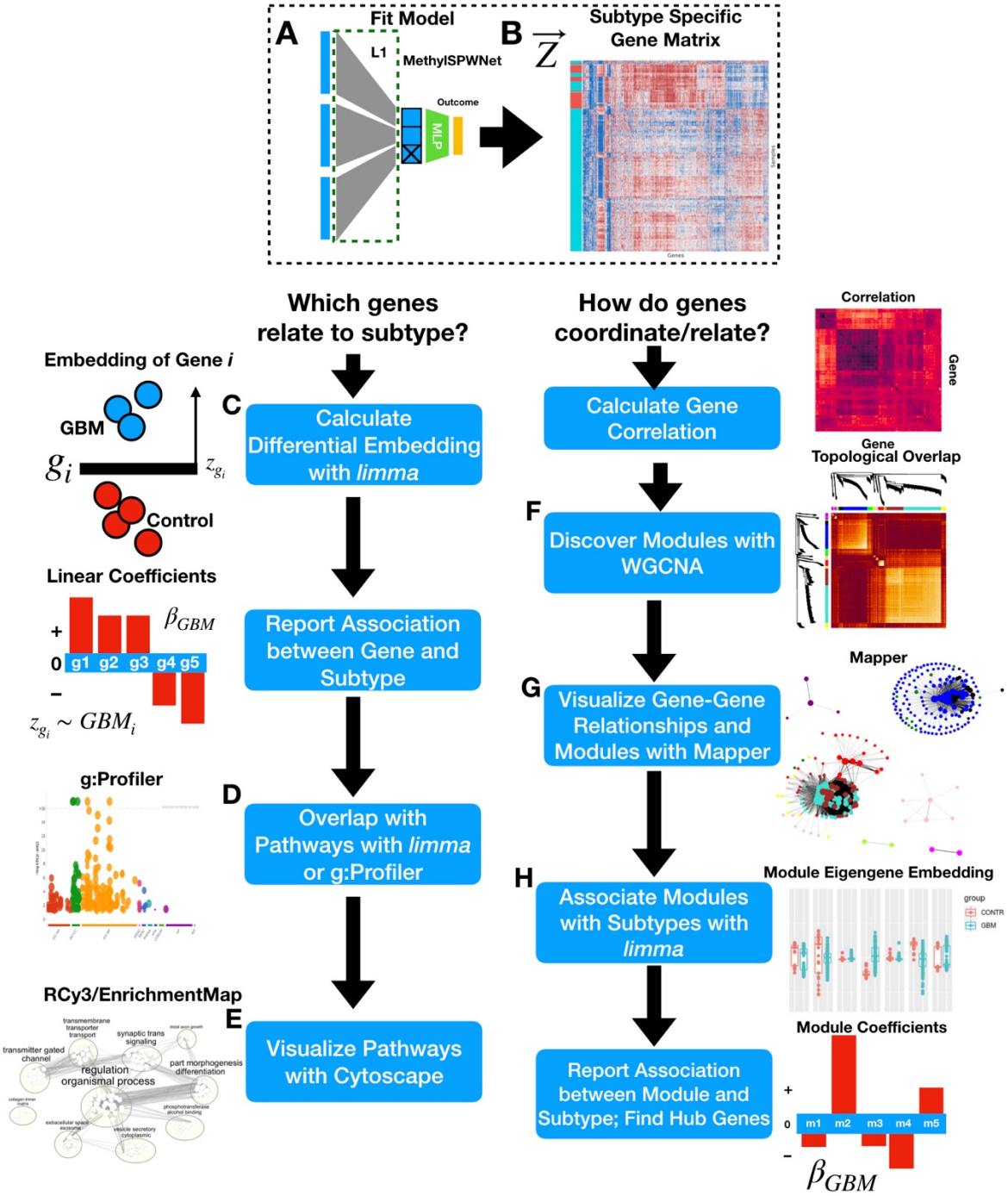


Figure 7: Flow Diagram for Possible Downstream Applications of *MethylSPWNet*: A) User fits *MethylSPWNet* model to predict brain cancer subtypes; B) gene embedding matrix (samples by genes; in this example GBM vs Control) is extracted from the model from the test samples; user decides whether they want to **(C-E)** associate gene-level embeddings with disease outcome, or **(F-H)** relate genes to one another to imply functional relationships; C) user opts for pathways analysis; *Empirical Bayes* method is used to identify genes with differential embeddings between GBM and controls using an empirical bayes moderated linear model; D) significant genes may be passed into g:Profiler or *limma*'s internal functions

camera, *goana*, and *kegga*, for pathways enrichment over GO, KEGG, Reactome, etc. databases; E) pathways are summarized and visualized using EnrichmentMap, accessed via RCy3; F) user opts for gene correlation analysis (WGCNA), first calculating correlation between genes, then calculating topological overlap over correlation transformed by power, hierarchical clustering is applied to deduce modules; G) gene-gene networks and module membership is visualized using Mapper; H) modules related to GBM are extracted using the *Empirical Bayes* method by comparing projections into first eigengene (samples by modules matrix) for each module for GBM vs controls

CpG Island/Gene Context Analysis

Using the MethylSPWNet, each CpG is assigned a weight that relates the CpG to its associated gene or genomic context. These CpG weights are learned by the neural network and can be used to rank genes based on their relative importance (rank assigned by maximum absolute CpG weight), an alternative measurement to the modularity analysis. Inspection of the weights of CpGs within each gene can provide insight into sites and contexts that are important for predicting brain cancer subtypes. Further, investigation of weights spatially across the genome may give rise to important patterns and motifs that could warrant future investigation. In the supplementary material, we first considered the contexts mentioned above independently and did not consider the joint impact of context (e.g., did not associate with island-promoter regions, which are generally considered to be regions more causally related to changes in their expression). We then considered sites that were associated with island promoter regions (including shore and shelf context; more causally associated with gene expression) and separately compared the overlap of the CpGs correspondent to top positive and negative weights to the CpGs that were unassociated with this context (open sea and not TSS200/1500). We separately considered CpGs within the promoter regions. Finally, we considered intragenic CpGs and whether or not their corresponding gene's promoter was methylated or unmethylated (as operationalized by calculating a beta-value methylation cutoff via local minima in the distribution of beta-values. The beta-value distribution, bimodally distributed, reflects the

705 distribution of proportion of methylated alleles across a bulk mixture of cells for individual CpG
706 sites. This distribution across CpG sites is typically estimated per individual(s). Beta values can
707 take on values between 0 and 1, but particularly concentrate closer to 0 or 1 to reflect that a site
708 is either "methylated" or "unmethylated". The intermediate proportions reflect scenarios from
709 which around half of the cells of the mixture are methylated at that site, which is uncommon, and
710 used to as the threshold to denote whether a site is methylated. Any CpG with a beta value above
711 the threshold was methylated. CpG methylation was averaged across each promoter and subject
712 to the threshold to determine methylation status). We calculated odds-ratios for
713 enrichment/depletion in these contexts using Fisher's exact tests. Without matching gene
714 expression information, we could not make any causal claims/inferences about how these
715 contexts modify gene expression to bring about these disease states.
716

717 **Description of Capsule-Inspired Neural Network**

718 The capsule-inspired network featured in this work operates by first finding representations of
719 the given CpG sets as denoted by the primary capsule formation. The features, CpGs, of the CpG
720 sets are fed into parallel implementations of a multi-layer perceptron, f_j, where the output
721 dimensions of each of the neural networks are the same. Thus, the dimensionality of the primary
722 capsules reflects the number of output neurons, a latent representation of each CpG set, times the
723 number of capsules, per individual. The mathematical formulation of this transformation is
724 presented below:

725 $$\overrightarrow{z_{gene_j}} = \mathrm{f_j}\left(\overrightarrow{x_j}\right)$$

726 For a single individual, the capsules, represented by row vectors, are stacked to form a capsule
727 matrix:

$$\overleftrightarrow{Z} = \begin{bmatrix} \overrightarrow{z_{gene1}} \\ \vdots \\ \overrightarrow{z_{genen}} \end{bmatrix}$$

An affine transformation transforms the primary capsules, a set of learnable parameters that seek to rotate, scale, shift the data, and encode information pertaining to the interactions between capsules.

$$\overleftrightarrow{Z_*} = \overleftrightarrow{W}\,\overleftrightarrow{Z}$$

Each primary child capsule's information is then dynamically routed to parent hidden or output capsules.

$$\overleftrightarrow{Y} = \sigma\left(\overleftrightarrow{C}\overleftrightarrow{Z_*}\right) = sq\left(\begin{bmatrix} \overrightarrow{y_{class\ 1}} \\ \vdots \\ \overrightarrow{y_{class\ m}} \end{bmatrix}\right)$$

Where:

$$\overrightarrow{y_{class\ j}} = sq\left(\sum_i C_{genei,classj}\,\overrightarrow{z_{gene\ i}}\right)$$

Dynamic routing aims to force the information encoded into each child to align with one parent capsule, thus utilized to calculate $\overleftrightarrow{C} = \{C_{ij}\}$, a bipartite network relating the child-parent-capsule. A vector of the same length represents each child as the output of the parent capsule. Analogous to the non-linear transformation of the sum of the information output from the previous layer of neurons for traditional neural networks, for each parent capsule, the child-capsule values are summed, and then a non-linear transform called a squash function, $sq(\vec{x}) = \frac{\|\vec{x}\|^2}{1+\|\vec{x}\|^2}\frac{\vec{x}}{\|\vec{x}\|}$, is applied to effectively zero-out, or *squash*, child capsules that do not agree with parent capsules.

Each child's contributions to a parent are weighted, but two constraints are imposed: First, the weights from each child to its parents must sum to 1. Second, a reward for the alignment of a child to exactly one parent is a dynamic routing by agreement mechanism. An iterative process updates the weights between the child and parent by adding their dot product. The update mechanism for calculating $\overleftrightarrow{C}$ is recapitulated below. After initializing $\overrightarrow{C_\iota} = softmax(\overrightarrow{\beta_1} = \vec{0})$ for $r \in \{1, 2, 3, \dots\}$ iterations:

$$\overrightarrow{C_\iota} = softmax(\overrightarrow{\beta_1})$$

$$\overrightarrow{Y_J} = sq\left(\sum_i C_{genei,classj}\, \overrightarrow{z_{gene\ \iota}}\right)$$

$$\beta_{ij} = \beta_{ij} + \overrightarrow{Y_{classJ}} \cdot \overrightarrow{z_{gene\ \iota}}$$

This formula is simplified from its original derivation and utilizes a few notational shortcuts.

Applying this operation for r iterations per batch per training epoch effectively prunes the other connections between the child and its parents as it converges on a single parent from which to send its information. Each output capsule per individual, $\overrightarrow{Y_J}$, is represented by a vector in some n-dimensional space. The output capsule with the highest L2-norm, $\|\overrightarrow{Y_J}\|_2$, is selected as the predicted class, and a margin loss is applied to penalize the model when it fails to either concretely have a very high ($m^+ = 0.9$) or very low probability ($m^- = 0.1$) of prediction.

$$L_{margin:\,classj} = \delta_{y_i,j}\, max\left(0, m^+ - \|\overrightarrow{Y_J}\|\right)^2 + \lambda\left(1 - \delta_{y_i,j}\right) max\left(0, \|\overrightarrow{Y_J}\| - m^-\right)^2$$

The Kronecker's delta $\delta_{y_i,j}$ is equal to one when the outcome for individual i is equal to the *j*-th class, thereby activating the left-hand margin loss that penalizes the model if the probability is below $m^+ = 0.9$. On the right-hand of the equation, $\delta_{y_i,j}$ is equal to zero when the outcome for

individual i is not equal to the *j*-th class, thereby activating the right-hand margin loss that penalizes the model if the probability is above $m^{-} = 0.1$.

The model is also penalized based on how much the original methylation array could be constructed from the true class's output capsule via a decoder neural network, $\hat{X} \sim p_{\phi}(\vec{X}|\vec{Y_j})$:

$$L_{reconstruct} \propto \left(\hat{X} - \vec{X}\right)^2$$

Of the most interest to a biologist may be the primary capsule embeddings per individual, $\overleftrightarrow{Z_*}$, which demonstrate interactions between these biological hypotheses and how the outcome of interest is separable within certain genomic context, and the weights between the primary and output capsules, $\overleftrightarrow{C}$, a bipartite graph demonstrates how these genomic regions are related hierarchically and have implications for parent processes. The coordinated response of capsules can also be derived through a bipartite projection of $\overleftrightarrow{C}$ into a unipartite network of capsules. Secondarily, of importance are the concatenation of the primary capsules, which demonstrate overall class separation, and the decoded output. Tweaking the embeddings or L2 norm of the output capsules and decoding can potentially effectively generate methylation data conditionally on outcomes of interest and interpolate between purified states, though this aspect was unexplored due to prohibitive dimensionality.

Description of MethylSPWNet

MethylSPWNet is the deep learning analog of a Group LASSO Regression model. The beta values for the CpGs for each gene, $\overrightarrow{x_j}$, is transformed into a single value, z_{genej}, through the multiplication of a set of gene-specific CpG weight matrices, $\overrightarrow{w_j}$. These weights are updated

throughout the training process to minimize the divergence between observed and expected outcomes. The magnitude of the weights dictates how much information from each CpG should be considered. The final gene-level summary value is given by:

$$z_{genej} = \sigma\left(\overrightarrow{w_j} \cdot \overrightarrow{x_j}\right)$$

The gene-level summary values are concatenated to form an array of gene-level summaries ($\vec{z} \in \mathbb{R}^n$):

$$\vec{z} = \left[z_{gene1}\ z_{gene2}\ z_{gene3} \ldots z_{genen}\right]$$

The final prediction for the network can be obtained using the following transformation via an MLP, f:

$$\hat{y} = f(\vec{z})$$

$$\hat{p} = softmax(\hat{y})$$

In the classification case, this predicted outcome is compared to the expected outcome via:

$$L_{CE} = \frac{-\sum_i \sum_c y_{i,c} log\left(\widehat{p_{i,c}}\right)}{N}$$

We applied Group L1 regularization to these weights to cause certain genes to drop out, returning genes important for the prediction of the cancer subtypes. The final LASSO penalty is given by:

$$L_{L1} = \sum_j \sqrt{d_j}\ \left\|\overrightarrow{w_{gene\ j}}\right\|_2$$

Where d_j is the number of CpGs assigned to that gene. An intermediate layer of the neural network, $\vec{z}$, stores gene-level summaries of DNAm information, and $\overrightarrow{w_{gene\ j}}$ contains the importance of each CpG for a particular gene.

Here, we contrast this summary measure, $z_{genej} = \sigma(\overrightarrow{w_j} \cdot \overrightarrow{x_j})$ to a more traditional summary measure, such as the median or mean methylation (mean displayed on the right): $\beta_{genej} = \frac{\sum_k \beta_{jk}}{d_j}$. Assuming the mean as our measurement, for simplification, it can be seen that each CpG is given equal weight $\frac{1}{d_j}$, while for *MethylSPWNet*, each CpG is given weight w_{jk}, which is learnable and reflective of the relative contribution of the given methylation beta value to the aggregate measure. A comparison between z_{genej} and β_{genej}can be found in the supplementary materials.

Hyperparameter Scans

MethylCapsNet includes the use of a hyperparameter optimization scheme, accessible through the *methylcaps-hypscan* module. Currently offered by the package is the availability to scan a number of hyperparameters, including the number of training epochs, length of the genomic region, the minimum number of CpGs to constitute a capsule, weighting schemes for reconstruction loss and survival loss, learning rate, in addition to other focused hyperparameters. Additionally, the search of *MethylNet* model architecture, randomized neural network topologies, was replaced by a framework that searches for the ideal number of neurons per neural network layer, conditional upon the choice of the number of layers. There are three search strategies for optimization, including randomized searches and Bayesian optimization techniques. This scheme differs from *MethylNet*, as both the neural network topology and set of hyperparameters can be optimized through the application of successive Gaussian processes to update some prior of losses over the set of hyperparameters. However, the results presented in this manuscript utilized the randomized search design. The jobs can be launched in parallel and scaled to meet the demands of a larger compute cluster.

833

834 ## Capsule Generation

835 Capsules specify the groupings of CpGs of the *MethylationArray* object. Capsule selection has
836 been incorporated into the *hyperparameter_scan* and the *methylcaps-model* subcommands.
837 Application Programming Interface (API) access to capsule selection and the building may be
838 accessed through the *build_capsules* script. As mentioned in the results section, prespecified
839 capsules include the following Illumina methylation array annotations – UCSC_RefGene_Name,
840 UCSC_RefGene_Accession, UCSC_RefGene_Group, UCSC_CpG_Islands_Name,
841 Relation_to_UCSC_CpG_Island, Phantom, DMR, Enhancer, HMM_Island,
842 Regulatory_Feature_Name, Regulatory_Feature_Group, and DHS. Additionally, the following
843 GSEA gene sets may be queried: C5.BP, C6, C1, H, C3.MIR, C2.CGP, C4.CM, C5.CC,
844 C3.TFT, C5.MF, C7, C2.CP, C4.CGN. Users can also specify their own capsules through the
845 presentation of a pickled dictionary containing a DataFrame that maps each CpG to a context
846 name of choice. Capsule generation may also be accomplished by breaking up the entire hg19
847 genome into overlapping windows of fixed width [57] (*genomic_binned* selection). We
848 recommend the utilization of the Circos tool [111] for visualization of derived capsule
849 relationships using the *genomic_binned* option.

850

851 ## Selection of Capsules for *MethylCapsNet* and *MethylSPWNet*

852 For the training of *MethylSPWNet*, we utilized all genes that overlapped with the 200,000 most
853 variable CpGs across the CNS tumors. For *MethylCapsNet*, we could not utilize the complete set
854 of genes due to the number of free parameters, a gene list of 650 genes was manually curated for
855 *MethylCapsNet* that included genes related to WNT, SHH, DKK1, beta-catenin, SFRP, NPR3,

amongst others. This list was reduced to 55 genes via recognition of genes by domain experts and thresholding of the minimum number of CpGs. As a further description, for both approaches, hyperparameter scans were utilized for pruning genes that did not contain a minimum number of CpGs (this threshold was varied via the hyperparameter scan), resulting in a lower number of genes than originally specified (n=10,341). In future iterations of the capsule-inspired network-based approach, gene selection constraints will be lifted via reduction of free parameters and the adoption of explicit network building approaches.

Co-Methylation Embedding Modules

MethylSPWNet derived gene-level methylation summaries/embeddings were correlated to each other and within their own set of top genes. Louvain modularity was performed on a k-nearest neighbors graph of *MethylSPWNet* gene-level embeddings to establish preliminary modules. For *MethylSPWNet,* the final gene co-methylation/embedding analysis was carried out on a subtype-specific basis, done so through the use of WGCNA.

For *MethylCapsNet*, capsule-level embeddings were averaged across all individuals to form overall embeddings. Though just as relevant, these approaches can be extended to capsule-level embeddings on the individual level or aggregated across meaningful subgroups. To derive final measures of coordinated response between capsules, we averaged the routing matrix coefficients across the individuals to form a weighted bipartite graph and calculated a bipartite projection of the graph to form a unipartite graph of capsules. We utilized the Louvain modularity algorithm to discover hubs in this network and performed enrichment analyses on the pathway level using enrichr [112] to describe these hubs.

879

880 **Random Forest Approaches in Comparison**

881 As a comparison to *MethylSPWNet* and *MethylCapsNet*, we adapted the Random Forest scheme
882 featured in a DNAm machine learning classification study. We selected 10k CpGs by first fitting
883 100 random forest models, each themselves fit on 10k randomly selected CpGs. Shapley
884 Additive Feature Explanations (SHAP) [113], was employed to determine the top CpGs from
885 each random forest run. The 10k CpGs with the highest average rank across the 100 random
886 forest models were selected for the final RF model. We note that we did not have access to the
887 original set of 10k CpGs featured in the previous classifier development study. The previous
888 study also utilized probability calibration methods to boost the model sensitivity and specificity,
889 which we avoided to ensure a proper comparison between methods.

890

891 **Analysis of CpG Weights derived from *MethylSPWNet***

892 *MethylSPWNet* derives CpG-specific weights, $\vec{w}$, that relates each CpG to its respective gene.
893 We rank-ordered, reverse rank-ordered, and absolute-value reverse rank-ordered these lists to
894 yield CpGs that were important to differentiate the tumor types. We subset the first 1000 CpGs,
895 marked which genes they corresponded to, and tested for enrichment using enrichr in our
896 preliminary weight analysis. Finally, weights were also rank-ordered and reverse-rank ordered to
897 yield the set of top negative and positive weights, respectively, the CpGs correspondent to the
898 top number of CpGs (selected to highlight tendencies of enrichment and depletion) were related
899 to the various island and gene context.

900

901 **Method to Cluster Gene-Level Brain Cancer Embeddings by Samples**

Recall that embeddings for individuals for *MethylSPWNet* were given in the following form (the design matrix is of dimensionality samples by genes):

$$\vec{z} = \left[z_{gene1}\ z_{gene2}\ z_{gene3} \dots z_{genen}\right]$$

For *MethylNet*, the embeddings are derived using the encoder:

$$\vec{z} = f(\vec{x})$$

Embeddings for individuals using the *MethylCapsNet* approach (of dimensionality samples by genes by latent dimensions) can be obtained by either averaging or concatenating (an aggregation, or *AGG* operator) the gene-level embeddings:

$$\overleftrightarrow{z} = AGG\left(\begin{bmatrix} \overrightarrow{z_{gene1}} \\ \vdots \\ \overrightarrow{z_{genen}} \end{bmatrix}\right)$$

Stacking these vectors for individuals would yield a design matrix that can be clustered using methods such as hierarchical clustering. We implemented hierarchical clustering using *scikit-learn* (>0.22) and found 14 clusters to compare against true labels of cell-of-origin, histological subtype, and histological and molecular subtype using the v-measure statistic and cluster separation using the Silhouette coefficient.

Web Application

We have developed a web application for the submission and investigation of *MethylCapsNet* outputs. The web application features three modules. The first is the network projection model, where capsules are related two to each other across subtypes, and network configurations can be changed by having some users tweak the relationships between the capsules and conservation. The second module displays routing information and the third module displays embedding information. Usage is detailed in the wiki.

Analysis Hardware and Software

The analyses run for this work were optimized utilizing K80 GPUs at the Dartmouth Research Computing Cluster. The algorithms were designed using Python 3.7, PyTorch version 1.1, and CUDA 9.0.

Dataset Preprocessing

We acquired data from GEO accession GSE109381 preprocessed data using *PyMethylProcess* and the subselected 200K of the most hypervariable CpGs (to focus on CpGs that may better differentiate CNS tumor subtypes) after functional normalization was applied to the data. SNPs and non-autosomal (sex chromosome) probes were omitted. Preprocessing steps have been detailed in using the pipeline of *PyMethylProcess* [21].

Abbreviations

ANN – Artificial Neural Network
CGI – CpG Island
CIMP – CpG Island Methylator Phenotype
CNS – Central Nervous System
CpG – Cytosine Guanine Dinucleotide
DNAm – DNA Methylation
GBM – Glioblastoma Multiforme
LGG – Low-Grade Glioma
MB – Medulloblastoma
MLP – Multi-Layer Perceptron
ReLU – Rectified Linear Unit
Tanh – Hyperbolic Tangent
TDA – Topological Data Analysis
TOM – Topological Overlap Matrix
WGCNA – Weighted Gene Correlation Network Analysis

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Competing Interests

Not applicable.

Code Availability

The software (*MethylCapsNet*, *MethylSPWNet*) is open source and can be found on GitHub at https://github.com/Christensen-Lab-Dartmouth/MethylCapsNet, on PyPI under the tag *methylcapsnet*, and on Docker at *joshualevy44/methylcapsnet*. While new features may be developed for the *MethylCapsNet* framework, community contributions are welcome in the form of GitHub pull requests and issues. A test pipeline is available in the software implementation, a Wiki, help documentation, and example R scripts for possible downstream analyses can be found in the GitHub repository.

Data Availability

Data used in this study was acquired from GEO accessions GSE109381, GSE84207, and GSE75067. Test data is available in our GitHub repository.

Correspondence Statement

Correspondence and request for materials should be addressed to Joshua J. Levy; email: joshua.j.levy.gr@dartmouth.edu .

Competing Interests

Not applicable.

Funding

This work was supported by NIH grants R01CA216265, R01DE022772, and P20GM104416 to BCC, Dartmouth College Neukom Institute for Computational Science CompX awards to BCC and LJV, and training fellowship support for AJT from T32LM012204. CLP and JJL are supported through the Burroughs Wellcome Fund Big Data in the Life Sciences at Dartmouth.

Authors' Contributions

The conception and design of the study were contributed by JJL and BCC. Implementation, programming, data acquisition, and analyses were by JJL. All authors contributed towards refining the analytic plan and direction. All authors contributed to the writing and editing of the manuscript. CLP and JJL tested the pipeline.

Acknowledgments

We would like to acknowledge Christian Haudenschild, Hildreth Robert Frost, and A. James O'Malley for their thoughtful discussions.

References

1. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. Genome Biol. 2019;20:249.

2. Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of cellular differentiation. Cell Cycle. 2010;9:3880–3.

3. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5:e1000602.

4. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3:771–84.

5. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8:389–99.

6. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nature Reviews Genetics. 2012;13:679–92.

7. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. The Lancet. 2018;392:777–86.

8. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555:469–74.

9. Hegi ME, Diserens A-C, Gorlia T, Hamou M-F, de Tribolet N, Weller M, et al. MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. New England Journal of Medicine. 2005;352:997–1003.

10. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483:479–83.

1024 11. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al.
1025 Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma.
1026 Cancer Cell. 2010;17:510–22.

1027 12. Christensen BC, Smith AA, Zheng S, Koestler DC, Houseman EA, Marsit CJ, et al. DNA
1028 Methylation, Isocitrate Dehydrogenase Mutation, and Survival in Glioma. J Natl Cancer Inst.
1029 2011;103:143–53.

1030 13. J. Dabrowski M, Wojtas B. Global DNA Methylation Patterns in Human Gliomas and Their
1031 Interplay with Other Epigenetic Modifications. Int J Mol Sci [Internet]. 2019 [cited 2020 Jul
1032 21];20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6678179/

1033 14. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral
1034 Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2017;31:737-754.e6.

1035 15. Maros ME, Capper D, Jones DTW, Hovestadt V, von Deimling A, Pfister SM, et al. Machine
1036 learning workflows to estimate class probabilities for precision cancer diagnostics on DNA
1037 methylation microarray data. Nature Protocols. 2020;15:479–512.

1038 16. Rauschert S, Raubenheimer K, Melton PE, Huang RC. Machine learning and clinical
1039 epigenetics: a review of challenges for diagnosis and classification. Clinical Epigenetics.
1040 2020;12:51.

1041 17. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

1042 18. Hinton GE. Connectionist learning procedures. Artificial Intelligence. 1989;40:185–234.

1043 19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional
1044 Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in
1045 Neural Information Processing Systems 25 [Internet]. Curran Associates, Inc.; 2012 [cited 2019
1046 Jun 9]. p. 1097–105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-
1047 with-deep-convolutional-neural-networks.pdf

1048 20. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. MethylNet: an automated
1049 and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics.
1050 2020;21:108.

1051 21. Levy JJ, Titus AJ, Salas LA, Christensen BC. PyMethylProcess - convenient high-throughput
1052 preprocessing workflow for DNA methylation data. Bioinformatics. 2019;

1053 22. Titus AJ, Wilkins OM, Bobak CA, Christensen BC. Unsupervised deep learning with
1054 variational autoencoders applied to breast tumor genome-wide DNA methylation data with
1055 biologic feature extraction. bioRxiv. 2018;433763.

1056 23. Titus AJ, Bobak CA, Christensen BC. A New Dimension of Breast Cancer Epigenetics -
1057 Applications of Variational Autoencoders with DNA Methylation. 2018 [cited 2019 Jun 15]. p.

1058 140–5. Available from:
1059 https://www.scitepress.org/PublicationsDetail.aspx?ID=kxmWGvGaZh0=&t=1

1060 24. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA
1061 methylation states using deep learning. Genome Biology. 2017;18:67.

1062 25. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I,
1063 Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural
1064 Information Processing Systems 30 [Internet]. Curran Associates, Inc.; 2017 [cited 2019 Jun 9].
1065 p. 4765–74. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-
1066 interpreting-model-predictions.pdf

1067 26. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of
1068 Any Classifier. arXiv:160204938 [cs, stat] [Internet]. 2016 [cited 2020 Feb 9]; Available from:
1069 http://arxiv.org/abs/1602.04938

1070 27. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine
1071 Learning for Biological Networks. Cell. 2018;173:1581–92.

1072 28. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol.
1073 2013;14:R115.

1074 29. Handl L, Jalali A, Scherer M, Eggeling R, Pfeifer N. Weighted elastic net for unsupervised
1075 domain adaptation with application to age prediction from DNA methylation data.
1076 Bioinformatics. 2019;35:i154–63.

1077 30. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data
1078 with case-control studies. Bioinformatics. 2012;28:1368–75.

1079 31. Zhou W, Lo S-H. Analysis of genotype by methylation interactions through sparsity-inducing
1080 regularized regression. BMC Proceedings. 2018;12:40.

1081 32. Choi J, Kim K, Sun H. New variable selection strategy for analysis of high-dimensional DNA
1082 methylation data. J Bioinform Comput Biol. 2018;16:1850010.

1083 33. Dong NT, Khosla M. Revisiting Feature Selection with Data Complexity. bioRxiv.
1084 2019;754630.

1085 34. Sun L, Namboodiri S, Chen E, Sun S. Preliminary Analysis of Within-Sample Co-methylation
1086 Patterns in Normal and Cancerous Breast Samples. Cancer Inform.
1087 2019;18:1176935119880516.

1088 35. Rickabaugh TM, Baxter RM, Sehl M, Sinsheimer JS, Hultin PM, Hultin LE, et al. Acceleration
1089 of age-associated methylation patterns in HIV-1-infected adults. PLoS ONE. 2015;10:e0119201.

1090 36. Zhang J, Huang K. Pan-cancer analysis of frequent DNA co-methylation patterns reveals
1091 consistent epigenetic landscape changes in multiple cancers. BMC Genomics. 2017;18:1045.

1092 37. Gomez L, Odom GJ, Young JI, Martin ER, Liu L, Chen X, et al. coMethDMR: accurate
1093 identification of co-methylated and differentially methylated regions in epigenome-wide
1094 association studies with continuous phenotypes. Nucleic Acids Res. 2019;47:e98–e98.

1095 38. Lien TG, Borgan Ø, Reppe S, Gautvik K, Glad IK. Integrated analysis of DNA-methylation and
1096 gene expression using high-dimensional penalized regression: a cohort study on bone mineral
1097 density in postmenopausal women. BMC Medical Genomics. 2018;11:24.

1098 39. Ng B, Jafarzadeh S, Cole D, Goldenberg A, Mostafavi S. DNA Methylation Network
1099 Estimation with Sparse Latent Gaussian Graphical Model. 2018.

1100 40. Davies M, Volta M, Pidsley R, Lunnon K, Dixit A, Lovestone S, et al. Functional annotation of
1101 the human brain methylome identifies tissue-specific epigenetic variation across brain and
1102 blood. Genome biology. 2012;13:R43.

1103 41. Cui Z-J, Zhou X-H, Zhang H-Y. DNA Methylation Module Network-Based Prognosis and
1104 Molecular Typing of Cancer. Genes (Basel) [Internet]. 2019 [cited 2020 Feb 9];10. Available
1105 from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6722866/

1106 42. Mallona I, Aussó S, Díez-Villanueva A, Moreno V, Peinado MA. Modular dynamics of DNA
1107 co-methylation networks exposes the functional organization of colon cancer cells' genome.
1108 bioRxiv. 2018;428730.

1109 43. Tremblay BL, Guénard F, Lamarche B, Pérusse L, Vohl M-C. Network Analysis of the Potential
1110 Role of DNA Methylation in the Relationship between Plasma Carotenoids and Lipid Profile.
1111 Nutrients. 2019;11:1265.

1112 44. Mallik S, Bandyopadhyay S. WeCoMXP: Weighted Connectivity Measure Integrating Co-
1113 Methylation, Co-Expression and Protein-Protein Interactions for Gene-Module Detection.
1114 IEEE/ACM Trans Comput Biol Bioinform. 2018;

1115 45. Wang F, Xu H, Zhao H, Gelernter J, Zhang H. DNA co-methylation modules in postmortem
1116 prefrontal cortex tissues of European Australians with alcohol use disorders. Scientific Reports.
1117 2016;6:1–11.

1118 46. Bartlett TE, Olhede SC, Zaikin A. A DNA Methylation Network Interaction Measure, and
1119 Detection of Network Oncomarkers. PLoS One [Internet]. 2014 [cited 2020 Feb 10];9. Available
1120 from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3882261/

1121 47. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K, et al. Aging effects on DNA
1122 methylation modules in human brain and blood tissue. Genome Biol. 2012;13:R97.

1123 48. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis.
1124 BMC Bioinformatics. 2008;9:559.

1125 49. Akulenko R, Helms V. DNA co-methylation analysis suggests novel functional associations
1126 between gene pairs in breast cancer samples. Hum Mol Genet. 2013;22:3016–22.

1127 50. Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, et al. Nucleotide
1128 distance influences co-methylation between nearby CpG sites. Genomics. 2020;112:144–50.

1129 51. Hao J, Kim Y, Kim T-K, Kang M. PASNet: pathway-associated sparse deep neural network for
1130 prognosis prediction from high-throughput data. BMC Bioinformatics. 2018;19:510.

1131 52. Hao J, Masum M, Oh JH, Kang M. Gene- and Pathway-Based Deep Neural Network for Multi-
1132 omics Data Integration to Predict Cancer Survival Outcomes. In: Cai Z, Skums P, Li M, editors.
1133 Bioinformatics Research and Applications. Cham: Springer International Publishing; 2019. p.
1134 113–24.

1135 53. Borisov V, Haug J, Kasneci G. CancelOut: A Layer for Feature Selection in Deep Neural
1136 Networks. In: Tetko IV, Kůrková V, Karpov P, Theis F, editors. Artificial Neural Networks and
1137 Machine Learning – ICANN 2019: Deep Learning. Cham: Springer International Publishing; 2019.
1138 p. 72–83.

1139 54. Crawford J, Greene CS. Incorporating biological structure into machine learning models in
1140 biomedicine. Current Opinion in Biotechnology. 2020;63:126–34.

1141 55. Xie G, Dong C, Kong Y, Zhong JF, Li M, Wang K. Group Lasso Regularized Deep Learning for
1142 Cancer Prognosis from Multi-Omics and Clinical Features. Genes (Basel) [Internet]. 2019 [cited
1143 2020 Mar 28];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6471789/

1144 56. Barthel FP, Johnson KC, Wesseling P, Verhaak RGW. Evolving Insights into the Molecular
1145 Neuropathology of Diffuse Gliomas in Adults. Neurol Clin. 2018;36:421–37.

1146 57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
1147 Bioinformatics. 2010;26:841–2.

1148 58. Artemenkov A, Panov M. NCVis: Noise Contrastive Approach for Scalable Visualization. 2020
1149 [cited 2020 May 6]; Available from: https://arxiv.org/abs/2001.11411v1

1150 59. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11:
1151 protein–protein association networks with increased coverage, supporting functional discovery
1152 in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13.

1153 60. Babic I, Mischel PS. Multiple functions of a glioblastoma fusion oncogene. J Clin Invest.
1154 2013;123:548–51.

1155 61. Parker BC, Annala MJ, Cogdell DE, Granberg KJ, Sun Y, Ji P, et al. The tumorigenic FGFR3-
1156 TACC3 gene fusion escapes miR-99a regulation in glioblastoma. J Clin Invest. 2013;123:855–65.

1157 62. Macy ME, Birks DK, Barton VN, Chan MH, Donson AM, Kleinschmidt-DeMasters BK, et al.
1158 Clinical and molecular characteristics of congenital glioblastoma. Neuro Oncol. 2012;14:931–41.

1159 63. Berezovsky AD, Poisson LM, Cherba D, Webb CP, Transou AD, Lemke NW, et al. Sox2
1160 Promotes Malignancy in Glioblastoma by Regulating Plasticity and Astrocytic Differentiation.
1161 Neoplasia. 2014;16:193-206.e25.

1162 64. Ibrahim K, Abdul Murad NA, Harun R, Jamal R. Knockdown of Tousled-like kinase 1 inhibits
1163 survival of glioblastoma multiforme cells. International Journal of Molecular Medicine.
1164 2020;46:685–99.

1165 65. Huang Q, Zhang X-W, Ma Y-S, Lu G-X, Xie R-T, Yang H-Q, et al. Up-regulated microRNA-299
1166 corrected with poor prognosis of glioblastoma multiforme patients by targeting ELL2. Jpn J Clin
1167 Oncol. 2017;47:590–6.

1168 66. Krishnan R, Boddapati N, Mahalingam S. Interplay between human nucleolar GNL1 and
1169 RPS20 is critical to modulate cell proliferation. Scientific Reports. 2018;8:11421.

1170 67. Friesen C, Hormann I, Roscher M, Fichtner I, Alt A, Hilger R, et al. Opioid receptor activation
1171 triggering downregulation of cAMP improves effectiveness of anti-cancer drugs in treatment of
1172 glioblastoma. Cell Cycle. 2014;13:1560–70.

1173 68. Pearson JRD, Regad T. Targeting cellular pathways in glioblastoma multiforme. Signal
1174 Transduct Target Ther. 2017;2:17040.

1175 69. Xiong A, Roy A, Spyrou A, Weishaupt H, Marinescu VD, Olofsson T, et al. Nuclear Receptor
1176 Binding Protein 2 Is Downregulated in Medulloblastoma, and Reduces Tumor Cell Survival upon
1177 Overexpression. Cancers (Basel) [Internet]. 2020 [cited 2020 Aug 9];12. Available from:
1178 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7352854/

1179 70. de la Rocha AMA, Sampron N, Alonso MM, Matheu A. Role of SOX family of transcription
1180 factors in central nervous system tumors. Am J Cancer Res. 2014;4:312–24.

1181 71. Rivero-Hinojosa S, Lau LS, Stampar M, Staal J, Zhang H, Gordish-Dressman H, et al.
1182 Proteomic analysis of Medulloblastoma reveals functional biology with translational potential.
1183 acta neuropathol commun. 2018;6:48.

1184 72. Qi Y, Gao Y. Clinical significance of miR-33b in glioma and its regulatory role in tumor cell
1185 proliferation, invasion and migration. Biomarkers in Medicine. 2020;14:539–48.

1186 73. Wang X, Huang Z, Wu Q, Prager BC, Mack SC, Yang K, et al. MYC-regulated Mevalonate
1187 Metabolism Maintains Brain Tumor Initiating Cells. Cancer Res. 2017;77:4947–60.

1188 74. Marx S, Xiao Y, Baschin M, Splittstöhser M, Altmann R, Moritz E, et al. The Role of Platelets
1189 in Cancer Pathophysiology: Focus on Malignant Glioma. Cancers (Basel) [Internet]. 2019 [cited
1190 2020 Aug 9];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6521321/

1191 75. Wu X, Rauch TA, Zhong X, Bennett WP, Latif F, Krex D, et al. CpG island hypermethylation in
1192 human astrocytomas. Cancer Res. 2010;70:2718–27.

1193 76. Caponegro MD, Moffitt RA, Tsirka SE. Expression of neuropilin-1 is linked to glioma
1194 associated microglia and macrophages and correlates with unfavorable prognosis in high grade
1195 gliomas. Oncotarget. 2018;9:35655–65.

1196 77. Xia Z, Ouyang D, Li Q, Li M, Zou Q, Li L, et al. The Expression, Functions, Interactions and
1197 Prognostic Values of PTPRZ1: A Review and Bioinformatic Analysis. J Cancer. 2019;10:1663–74.

1198 78. Panosyan EH, Lin HJ, Koster J, Lasky JL. In search of druggable targets for GBM amino acid
1199 metabolism. BMC Cancer [Internet]. 2017 [cited 2020 Aug 9];17. Available from:
1200 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331648/

1201 79. Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat
1202 Med. 2011;17:330–9.

1203 80. Chenn A. Wnt/β-catenin signaling in cerebral cortical development. Organogenesis.
1204 2008;4:76–80.

1205 81. Testa U, Castelli G, Pelosi E. Genetic Abnormalities, Clonal Evolution, and Cancer Stem Cells
1206 of Brain Tumors. Med Sci (Basel) [Internet]. 2018 [cited 2020 Jul 1];6. Available from:
1207 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6313628/

1208 82. Li T, Forbes ME, Fuller GN, Li J, Yang X, Zhang W. IGFBP2: integrative hub of developmental
1209 and oncogenic signaling network. Oncogene. 2020;39:2243–57.

1210 83. Ishak G, Leal MF, dos Santos NPC, Demachki S, Nunes CAM, do Nascimento Borges B, et al.
1211 Deregulation of MYC and TP53 through genetic and epigenetic alterations in gallbladder
1212 carcinomas. Clin Exp Med. 2015;15:421–6.

1213 84. Jun I, Park HS, Piao H, Han JW, An MJ, Yun BG, et al. ANO9/TMEM16J promotes
1214 tumourigenesis via EGFR and is a novel therapeutic target for pancreatic cancer. Br J Cancer.
1215 2017;117:1798–809.

1216 85. Hatanpaa KJ, Burma S, Zhao D, Habib AA. Epidermal Growth Factor Receptor in Glioma:
1217 Signal Transduction, Neuropathology, Imaging, and Radioresistance. Neoplasia. 2010;12:675–
1218 84.

1219 86. Fleischer T, Tekpli X, Mathelier A, Wang S, Nebdal D, Dhakal HP, et al. DNA methylation at
1220 enhancers identifies distinct breast cancer lineages. Nature Communications. 2017;8:1379.

1221 87. Holm K, Staaf J, Lauss M, Aine M, Lindgren D, Bendahl P-O, et al. An integrated genomics
1222 analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to
1223 chromatin states in normal mammary cells. Breast Cancer Res. 2016;18:27.

1224 88. Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules. arXiv:171009829 [cs]
1225 [Internet]. 2017 [cited 2020 Jan 18]; Available from: http://arxiv.org/abs/1710.09829

1226 89. Venkatraman S, S B, Sarma R. Building Deep, Equivariant Capsule Networks. 2019.

1227 90. Wang L, Miao X, Zhang J, Cai J. MultiCapsNet: a interpretable deep learning classifier
1228 integrate data from multiple sources. bioRxiv. 2019;570507.

1229 91. Danielsson A, Nemes S, Tisell M, Lannering B, Nordborg C, Sabel M, et al. MethPed: a DNA
1230 methylation classifier tool for the identification of pediatric brain tumor subtypes. Clin
1231 Epigenetics. 2015;7:62.

1232 92. Hovestadt V, Jones DTW, Picelli S, Wang W, Kool M, Northcott PA, et al. Decoding the
1233 regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature.
1234 2014;510:537–41.

1235 93. Baeza N, Weller M, Yonekawa Y, Kleihues P, Ohgaki H. PTEN methylation and expression in
1236 glioblastomas. Acta Neuropathol. 2003;106:479–85.

1237 94. Capaccione KM, Pine SR. The Notch signaling pathway as a mediator of tumor survival.
1238 Carcinogenesis. 2013;34:1420–30.

1239 95. Fan X, Matsui W, Khaki L, Stearns D, Chun J, Li Y-M, et al. Notch pathway inhibition depletes
1240 stem-like cells and blocks engraftment in embryonal brain tumors. Cancer Res. 2006;66:7445–
1241 52.

1242 96. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, et al. PTEN, a putative protein tyrosine
1243 phosphatase gene mutated in human brain, breast, and prostate cancer. Science.
1244 1997;275:1943–7.

1245 97. He X, Zhang L, Chen Y, Remke M, Shih D, Lu F, et al. The G protein α subunit Gαs is a tumor
1246 suppressor in Sonic hedgehog-driven medulloblastoma. Nat Med. 2014;20:1035–42.

1247 98. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical
1248 evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA
1249 methylation profiling. Genome Biology. 2016;17:208.

1250 99. Zhou L, Ng HK, Drautz-Moses DI, Schuster SC, Beck S, Kim C, et al. Systematic evaluation of
1251 library preparation methods and sequencing platforms for high-throughput whole genome
1252 bisulfite sequencing. Scientific Reports. 2019;9:10383.

1253 100. Moran S, Vizoso M, Martinez-Cardús A, Gomez A, Matías-Guiu X, Chiavenna SM, et al.
1254 Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the
1255 Infinium HumanMethylation450 Microarray. Epigenetics. 2014;9:829–33.

1256 101. Bodnar C, Cangea C, Liò P. Deep Graph Mapper: Seeing Graphs through the Neural Lens.
1257 arXiv:200203864 [cs, stat] [Internet]. 2020 [cited 2020 Feb 20]; Available from:
1258 http://arxiv.org/abs/2002.03864

1259 102. Veen HJ van, Saul N, Eargle D, Mangham SW. Kepler Mapper: A flexible Python
1260 implementation of the Mapper algorithm. Journal of Open Source Software. 2019;4:1315.

1261 103. Wang T, Johnson T, Jie Z, Huang K. Topological Methods for Visualization and Analysis of
1262 High Dimensional Single-Cell RNA Sequencing Data. Pacific Symposium on Biocomputing.
1263 2019;24:350–61.

1264 104. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, et al. Single-cell
1265 topological RNA-Seq analysis reveals insights into cellular differentiation and development. Nat
1266 Biotechnol. 2017;35:551–60.

1267 105. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al.
1268 Extracting insights from the shape of complex data using topology. Scientific Reports.
1269 2013;3:1236.

1270 106. Singh D, Yamada M. FsNet: Feature Selection Network on High-dimensional Biological
1271 Data. arXiv:200108322 [cs, stat] [Internet]. 2020 [cited 2020 Feb 22]; Available from:
1272 http://arxiv.org/abs/2001.08322

1273 107. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential
1274 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.
1275 2015;43:e47–e47.

1276 108. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server
1277 for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids
1278 Res. 2019;47:W191–8.

1279 109. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software
1280 Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res.
1281 2003;13:2498–504.

1282 110. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. RCy3: Network biology using Cytoscape
1283 from within R. F1000Res [Internet]. 2019 [cited 2020 Jul 2];8. Available from:
1284 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6880260/

1285 111. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An
1286 information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

1287 112. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and
1288 collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.

1289 113. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local
1290 explanations to global understanding with explainable AI for trees. Nature Machine
1291 Intelligence. 2020;2:56–67.

1292