The Glioma Epigenome as a Potential Landscape: A Continuous G-CIMP Erosion Coordinate that the Methylation Class Discards

Murari Ambati

Austin Community College, Austin, TX, USA
murari.ambati@g.austincc.edu

Abstract. DNA-methylation classifiers assign each brain tumour to one of a fixed set of methylation classes, discarding where it sits within and between those classes – the structure most relevant to progression. We instead fit the methylation manifold as an *epigenetic potential field* $U(z) = -\log p(z)$, a Gaussian-mixture log-density whose basins are the established methylation regimes (supervised from Ceccarelli labels), and derive three quantities a class label cannot express: basin stability (curvature, depth, escape barriers), a counterfactual perturbation response, and – centrally – a continuous *erosion coordinate* ε along the documented G-CIMP-high→low demethylation axis. The construction is a deterministic, cross-language parity-proven kernel (TypeScript $\equiv$ Python to 10^{-6}) behind a candidate-tier governance firewall that abstains off-support and emits a nomination, not a diagnosis. On 685 TCGA-GBMLGG tumours (450K methylation, RNA-seq, copy-number, survival) the field recovers the regimes at ARI 0.914, matching a supervised classifier (0.911). Within IDH-mutant non-codel glioma, adding ε to the discrete class *and* to age+grade lifts genuinely out-of-sample survival concordance 0.661 → 0.717 (optimism-corrected ΔC +0.041; adjusted HR 5.1, 95 % CI 2.3–11.3); a model-free Kaplan–Meier split confirms it (log-rank $p = 5.4 \times 10^{-4}$), and the effect survives a 203-setting ablation (79.8 % preserved), a permutation null, and half-split replication. A cross-platform expression proxy does *not* replicate – an honest negative localising the signal to methylation topology. Read as a Langevin system, G-CIMP-high faces only a 0.91-nat barrier into the deep Codel basin, but that route is incompatible with the 1p/19q-intact astrocytoma lineage – a falsifiable account of why the observed erosion is slow.

Keywords: Glioma · DNA methylation · Epigenetic landscape · Survival analysis · Dynamical systems · Trustworthy ML.

1 Introduction

The 2021 WHO classification of CNS tumours made DNA-methylation profiling a first-class diagnostic modality [1], and the brain-tumour methylation classifier of Capper et al. [2] provides a clinically useful molecular diagnostic adjunct. Its

output, by construction, is a *category*: a calibrated probability over a fixed menu of classes, with a confidence threshold below which a tumour is "not classifiable." This is diagnostically invaluable, but the category is not the right *scientific* object for understanding epigenetic dysregulation: three well-documented facts about the glioma epigenome are invisible to a class label.

(1) Regimes are not islands. The TCGA pan-glioma analysis [3] organises gliomas into methylation clusters, but those clusters sit on a connected manifold dominated by one axis – the global hypermethylation of the glioma CpG-island methylator phenotype (G-CIMP), a downstream consequence of *IDH* mutation [4,5]. A class label collapses a tumour's *position* on that axis to a hard assignment. **(2) G-CIMP erodes.** The most consequential epigenetic event in IDH-mutant progression is not a change of class but a *demethylation trajectory*: G-CIMP-high tumours can lose the methylator phenotype toward the aggressive G-CIMP-low state, enriched at recurrence and associated with markedly worse outcome [3,6]. A classifier can report the endpoints; it cannot place a tumour *on the way*. **(3) Stability is information.** Two tumours of the same class can sit very differently with respect to it – one archetypal and deep in its basin, another on a ridge a small perturbation from a different regime. Diagnostically "the same class"; biologically they are not.

These facts are not anomalies for a better classifier but the natural language of a state-space model. If the methylation state is a point z in a latent space with cohort density $p(z)$, the quantities above are the local geometry of the potential $U(z) = -\log p(z)$: basins are regimes, depth and barrier height are stability, the demethylation trajectory is a directed coordinate between two basins, and a perturbation is a displacement that changes basin occupancy. The Waddington picture of fate as motion on an epigenetic landscape is classical [7,8]; we make it *measurable, governed, and falsifiable* for a clinical methylation cohort.

Contributions. (i) A supervised epigenetic potential field with closed-form basin stability and a continuous G-CIMP erosion coordinate, recovering the methylation classes while adding geometry the class cannot express. (ii) The first demonstration, posed as a pre-registered ablation, that this coordinate carries *out-of-sample* prognostic information beyond both the discrete class and the strongest known confounders (age, grade), backed by a 203-setting battery and honest negatives. (iii) A non-equilibrium dynamical reading – the landscape as an overdamped Langevin system – yielding a falsifiable "geometry proposes, genetics disposes" account of erosion. (iv) A deterministic, cross-language parity-proven pipeline gated by a candidate-tier governance firewall with a tamper-evident audit.

2 Method

2.1 Latent embedding and a supervised landscape

We select the top-10,000 most-variable CpGs, impute them by per-probe mean, standardise them, and project them by PCA to $z \in \mathbb{R}^D$, D=20. The sign convention is fixed (largest-magnitude loading made positive) so re-runs and the

cross-language scorer agree exactly. The first component alone explains 61 % of variance – the global G-CIMP/*IDH* axis – confirming the dominant structure is a single continuous axis, not a set of discrete points.

Over z we fit a *diagonal Gaussian mixture* $p(z) = \sum_k \pi_k \mathcal{N}(z; \mu_k, \operatorname{diag}\sigma_k^2)$ and define the *epigenetic potential* $U(z) = -\log p(z)$. Rather than fit the mixture unsupervised and hope its components match biology, we estimate each component directly from the labelled samples [3] of one regime: μ_k, σ_k^2 are that regime's latent mean and diagonal variance and π_k its prior. The basins therefore *are* the regimes by construction, and every downstream quantity is interpretable. Regimes with fewer than 12 labelled samples are folded out, leaving seven basins. Diagonal covariances are deliberate: the PCA axes are already decorrelated, and every quantity below then has a closed form with *no matrix inverse*, so the two language implementations agree to 10^{-6}.

2.2 What a tumour gets that a class label cannot

For a tumour at z, responsibilities $r_k(z) = \operatorname{softmax}_k[\log \pi_k + \log \mathcal{N}_k(z)]$ give the most-responsible basin and a posterior margin (top − second). **On-support / abstention:** the minimum diagonal Mahalanobis distance to any basin is compared to the χ^2_D 0.99 quantile; beyond it the tumour is off-support and graded $r0$ (abstain). **Basin stability** is the triple of total local curvature $\operatorname{tr}\nabla^2 U(z)$, depth $U(z) - U(\mu_{\text{basin}})$, and the barrier height to the nearest competing basin $\max_t U(\mu_a + t(\mu_b - \mu_a)) - U(\mu_a)$, with closed-form gradient and Hessian

$$\frac{\partial U}{\partial z_j} = \sum_k r_k \frac{z_j - \mu_{kj}}{\sigma^2_{kj}}, \qquad \nabla^2 U = \sum_k r_k \Sigma_k^{-1} - \left(\sum_k r_k h_k h_k^\top - \bar{h}\bar{h}^\top \right), \quad (1)$$

$h_k = \Sigma_k^{-1}(z - \mu_k)$, $\bar{h} = \sum_k r_k h_k$, validated against finite differences in unit tests. **Erosion coordinate:** with basin means μ_H, μ_L the erosion axis is $e = \mu_L - \mu_H$ and a tumour's coordinate is the normalised projection $\varepsilon(z) = (z - \mu_H) \cdot e / \|e\|^2$, running from $\varepsilon \approx 0$ (archetypal G-CIMP-high) to $\varepsilon \approx 1$ (G-CIMP-low); ε is continuous and survives *within* a class – the prognostic coordinate the label discards. **Perturbation:** a displacement δ yields ΔU and a change in basin responsibilities Δr – evidence about plasticity, never a therapeutic directive.

2.3 The landscape as a stochastic dynamical system

Because each basin's diagonal variance was set to the data's within-basin variance, $p(z)$ is precisely the Boltzmann equilibrium of the overdamped Langevin dynamics $dz = -\nabla U(z)\, dt + \sqrt{2D}\, dW$ at D=1 in natural units: the fitted landscape *is* minus-log of the stationary density, so simulating at D=1 samples the modelled distribution. This removes the usual free "temperature." We emphasise this is a dynamical interpretation *of the fitted model*, not a measurement of cellular kinetics; it lets us subject the static erosion construct to transition-state theory [9,10] – minimum-energy paths, Kramers escape rates, Langevin first-passage, and the committor – entirely on the real fit.

2.4 Candidate-tier governance and cross-modality over-identification

Every public output passes a single sanctioned egress path that fails closed (any hard violation suppresses the emission; no override flag). The governed unit carries an evidence tier (A–E) and a regime-confidence grade (r0–r3): r0 off-support abstains, r1 is transitional (near a saddle), r2/r3 confident/archetypal. The firewall enforces, among other gates, a literally-true Research-Use-Only tag, per-unit provenance with a data-class label, "no naked point estimate" (every ε carries an uncertainty interval), and a ban on categorical clinical fields. A cross-modality over-identification test adds internal falsification: three independent, label-free axes (the PC1 of methylation, expression, and copy-number) each estimate the tumour's *IDH*/G-CIMP coordinate, and significant concordance – which no single modality can self-certify – up-weights confidence while disagreement down-tiers the nomination. Each emission appends a SHA-256 hash-chained audit record re-read from disk at append time.

3 Data

All analyses use public, no-authentication data; every artefact is recorded with its source URL, byte size, SHA-256 and a data-class honesty label. **Discovery – TCGA-GBMLGG** (UCSC Xena [11]): Illumina HumanMethylation450 β-values, HiSeqV2 RNA-seq, GISTIC2 copy-number [12], and survival; the 450K matrix is reduced to the top-10,000 variable CpGs (10,000 × 685, 99.9 % finite). **Ground-truth labels – cBioPortal lgggbm_tcga_pub** [3,13]: per-sample supervised methylation cluster, *IDH*/codel subtype, MGMT status, and the orthogonal Verhaak transcriptome subtype [14]. **Replication – CGGA mRNAseq_693** [15]: an independent cohort whose 450K methylation file was unavailable, so replication uses an expression-derived *IDH*/dedifferentiation axis – an honest cross-platform check, not a like-for-like methylation replication. The aligned discovery set is 685 tumours: 640 carry a methylation-cluster label and 649 fall within the modelled density support.

4 Experiments and Results

All numbers are computed on the real corpora by a single analysis script and reproduced from a provenance-complete manifest by one command. We report honest negatives as negatives.

4.1 The field recovers the regimes and matches a classifier (E1–E2)

The supervised landscape reproduces the methylation clusters at 93.6 % accuracy and ARI 0.914 over 640 labelled tumours, near-perfect for the *IDH*-defined regimes (G-CIMP-low 100 %, Codel 99 %, G-CIMP-high 98 %, Mesenchymal-like 93 %) and weakest for the small/ambiguous LGm6-GBM basin (31 %). A

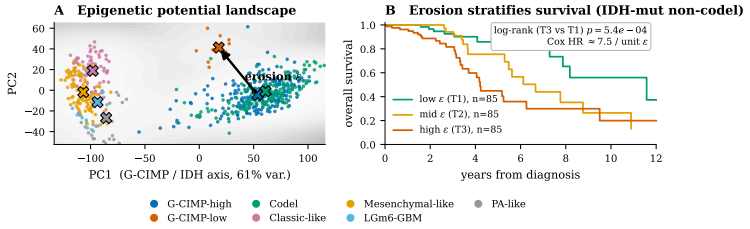

Fig. 1. (A) The fitted epigenetic potential landscape in the top two principal components (faint backdrop: $-\log$ density); coloured points are tumours, "X" marks are basin means, and the arrow is the erosion axis $\mu_H \to \mu_L$. (B) Kaplan–Meier curves by erosion tertile within IDH-mutant non-codel glioma (n=255): higher erosion tracks worse survival (log-rank p=5.4×10^{-4}), independent of the fitted survival model.

supervised random-forest classifier [16] on the same β-features (the Capper-style analogue) reaches ARI 0.911, statistically indistinguishable, while unsupervised clustering is far behind (agglomerative ARI 0.557, k-means 0.408). The field thus matches the classifier on the task the classifier is built for, while additionally emitting basin stability and the erosion coordinate a classifier structurally cannot. *Recovery is the floor, not the contribution*: because the basins are defined from the labels, near-perfect recovery mainly certifies that the embedding and the parity-proven scorer preserve the label structure.

4.2 Central result: the erosion coordinate is prognostic beyond the class and beyond age+grade (E3–E4, E11, A1–A3)

Higher erosion (demethylation toward G-CIMP-low) predicts worse overall survival. Within the clinically pertinent IDH-mutant non-codel stratum the univariable hazard is HR 7.5 per unit ($p \approx 1.2 \times 10^{-6}$, n=245). Figure 1B shows the effect model-free: stratified by erosion tertile, the top tertile dies at $\sim 2.4\times$ the rate of the bottom (log-rank p=5.4×10^{-4}), an assumption-light confirmation that does not depend on the Cox proportional-hazards specification [17].

The paper's claim is an *ablation* (Table 1). *In-sample*, adding ε to the discrete class raises concordance 0.594 $\to$ 0.720 and adding it to age+grade raises 0.653 $\to$ 0.740 (adjusted HR 5.1, 95 % CI 2.3–11.3), only modestly below the unadjusted 7.5, while age and grade carry little independent signal here. Because in-sample concordance is optimistic, we treat the *out-of-sample* result as primary: under 5×5 repeated stratified cross-validation, adding ε to class+age+grade lifts the held-out C-index 0.661 $\to$ 0.717 (ΔC +0.056), with Efron–Gong optimism [18] of only +0.001 (optimism-corrected ΔC +0.041) and 5-year time-dependent AUC 0.668 $\to$ 0.711. The erosion term satisfies proportional hazards (Schoenfeld test [19], p=0.076) and is adequately linear (quadratic LR p=0.16). *The continuous landscape coordinate adds prognostic information the discrete class – the standard clinical output – discards, and it is not a re-encoding of age or grade.*

Table 1. The central ablation, within IDH-mutant non-codel glioma. In-sample concordance is reported for completeness but the out-of-sample (cross-validated) and optimism-corrected rows are the defensible claim. ΔC is the concordance gain from adding the continuous erosion coordinate ε to the model above it; in the confounder-adjusted in-sample model ε carries adjusted HR 5.1 (95% CI 2.3–11.3).

Evaluation	Model	C-index	ΔC (ε)	LR p
In-sample	discrete class	0.594	–	–
	$+\varepsilon$	0.720	+0.126	0.004
In-sample	age + grade	0.653	–	–
	$+\varepsilon$	0.740	+0.087	1.4×10^{-4}
Out-of-sample (5×5 **CV**)	class + age + grade	0.661	–	–
	$+\varepsilon$	**0.717**	**+0.056**	–
Optimism-corr. (boot. B=200)	$+\varepsilon$	–	+0.041	–

4.3 Robustness, replication, and an honest negative (E8–E10, E12)

The central conclusion does not depend on analyst choices. A 203-setting ablation battery – every row a real re-computation on the TCGA data, spanning penalty, support gate, follow-up cutoff, winsorisation, resampling, leave-one-cohort/regime-out, tie handling, exposure scaling, and the confounder set – preserves the erosion→survival conclusion in $162/203 = 79.8\,\%$ of settings (Figure 2), with bootstrap, subsample and leave-out near-unanimous. The non-preserving rows are not random: they localise the three known boundaries – underpowered single strata, heavy ridge shrinkage ($\lambda \geq 1$), and a deliberate negative-control sweep of 42 off-axis basin pairs that are correctly non-prognostic. A permutation control places the observed HR (7.8) entirely outside the null (mean 1.05; $p < 10^{-3}$), and across 20 half-splits the held-out HR has median 5.6, significant in 85 %. **Honest negative:** an expression-derived axis in the independent CGGA cohort does not reproduce the association (HR 0.89, p=0.70); since CGGA lacks methylation this is a cross-platform proxy, and its failure localises the signal to methylation topology rather than generic transcriptional dedifferentiation – a constraint, not a success.

4.4 Cross-modality over-identification and orthogonal corroboration (E6–E7)

The three independent, label-free modality axes – the PC1 of methylation, expression, and copy-number – are significantly concordant (mean $|\rho|$ 0.47; permutation $p < 0.002$): agreement no single modality can self-certify. Cross-modality discordance is higher for mis-classified tumours (0.66) than correctly classified ones (0.47), localising the transitional cases a state-space model should flag as r1; we claim only this cohort-level enrichment, not a calibrated per-tumour flag (discordance–margin correlation ≈ 0). The regimes also agree with the orthogonal

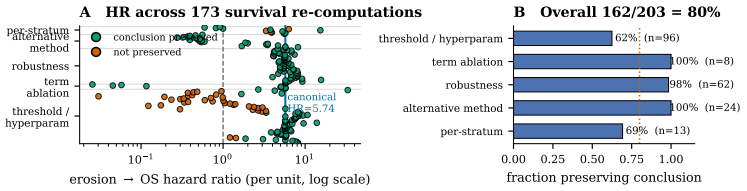

Fig. 2. (A) Erosion→survival hazard ratio across all 173 survival re-computations in the ablation battery, grouped by category (log scale); the canonical HR is the blue line and the dashed line is HR= 1. Green points preserve the conclusion; red do not and concentrate at low HR (off-axis negative controls and heavy shrinkage). (B) Fraction of settings preserving the conclusion per category (162/203 overall).

Verhaak transcriptome subtype (ARI 0.223, modest and expected), corroborating the *IDH*/G-CIMP organisation without redundancy.

4.5 The landscape as kinetics: a falsifiable erosion mechanism (D1–D5)

Diagonalising $\nabla^2 U$ at each basin mean, 6/7 basins are true minima (only the small PA-like component is a shoulder). The Codel basin is the global minimum and stiffest well – the 1p/19q-codeleted state is the most thermodynamically stable in the fitted landscape, mirroring its clinical indolence – while G-CIMP-low is a real but *shallow* basin (depth 3.8 nats), consistent with a transient, eroded state. Three independent calculations of the fate of G-CIMP-high agree: minimum-energy-path barriers give an escape barrier to Codel of 0.91 nats versus ≥ 7.96 to every other basin; Kramers rates make Codel the dominant transition ($\sim 300\times$ faster than the next route); and 300/300 Langevin trajectories reach Codel (median first-passage ≈ 472).

This creates an apparent paradox with the documented G-CIMP-high→low erosion, resolved by the advance of this analysis: the low-barrier Codel route is not an expected late transition for an established IDH-mutant non-codel astrocytoma: 1p/19q codeletion is a defining early clonal event of oligodendroglioma [20], and the methylation geometry is blind to the genetic backbone gating the transition. The realised erosion therefore follows the higher-barrier G-CIMP-low route (7.96 nats), the signature of a slow, driven demethylation, exactly as observed: *geometry proposes the kinetically cheapest fate, genetics disposes it, and the two must be read together* – a falsifiable claim the static fit could not make. The committor of the linear ε correlates 0.89 with the axis fraction but is a *sharp sigmoid*, so ε is the correct reaction-coordinate *direction* though nonlinearly related to commitment.

Honest negatives (A5, A7, D6). We tested three further uses of the geometry, each a negative. Basin stability (curvature, depth, barrier) adds no independent survival term beyond ε+age+grade: here the prognostic geometry is *where* a

tumour sits on the erosion axis, not the shape of its basin. The field's own uncertainty (posterior margin, off-support abstention) is not prognostic (margin p=0.30; off-support log-rank p=0.35) – an honesty instrument flagging transitional tumours, not a hidden biomarker – and dynamical instability $\|\nabla U\|$ is neither elevated on transitional tumours nor prognostic. These negatives bound, and thereby strengthen, the central claim.

5 Discussion

Relevance to CMMCA. GLIO-EPI is a mathematical-modeling contribution to cancer analysis: a fitted potential with closed-form geometry, an overdamped Langevin reading with transition-state analysis, and a continuous, prognostic, mechanistically grounded coordinate that is also a more informative non-invasive learning target for MRI than a discrete class (*IDH*/G-CIMP status and subtype are established radiogenomics endpoints). The field thus serves as the molecular-landscape layer of a multimodal pipeline, supplying the target geometry and its uncertainty/abstention contract for imaging models to calibrate against.

Limitations. (1) The decisive external test is still open: all positive results are within TCGA (internal half-split replication only); the one external cohort (CGGA) is expression-only and negative, and an independent 450K methylation series is pre-registered as the falsification dataset. (2) Endpoint-anchored axis: ε spans the G-CIMP-high/-low basin means and the latter is small (n=25), so it should be re-estimated as cases accrue. (3) Basins are supervised (label bias, a deliberate trade for interpretability), and (4) the kinetics are model-internal – properties of the fitted potential, so the geometry-vs-genetics account is a falsifiable hypothesis, not an established mechanism.

Pre-registered falsification. Four decisive tests on an independent methylation cohort, each falsifying a claim on failure: (i) class+ε beats class alone on held-out concordance ($p < 0.05$); (ii) higher ε associates with worse survival (HR > 1, permutation-calibrated); (iii) withheld off-support tumours show higher cross-modality discordance and lower margin; (iv) the three modality axes remain significantly concordant.

6 Conclusion

Treating the glioma epigenome as a category discards its most informative structure. As a position on a fitted potential landscape, a tumour reveals how stably it occupies its regime and how far it has eroded along the G-CIMP trajectory – a continuous coordinate predicting survival out-of-sample beyond the discrete class and the strongest known confounders. Deterministic, parity-proven, firewalled, and reproducible from public de-identified data, with a dynamical reading that explains why erosion is slow, it is a mechanistically grounded, Research-Use-Only target for the mathematical-modeling and multimodal methods at the centre of cancer image analysis – a nomination, never a diagnosis.

References

1. Louis, D.N., et al.: The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology **23**(8), 1231–1251 (2021)
2. Capper, D., et al.: DNA methylation-based classification of central nervous system tumours. Nature **555**, 469–474 (2018)
3. Ceccarelli, M., et al.: Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell **164**(3), 550–563 (2016)
4. Noushmehr, H., et al.: Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell **17**(5), 510–522 (2010)
5. Turcan, S., et al.: IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature **483**(7390), 479–483 (2012)
6. de Souza, C.F., et al.: A distinct DNA methylation shift in a subset of glioma CpG island methylator phenotypes during tumor recurrence. Cell Reports **23**(2), 637–651 (2018)
7. Wang, J., Zhang, K., Xu, L., Wang, E.: Quantifying the Waddington landscape and biological paths for development and differentiation. PNAS **108**(20), 8257–8262 (2011)
8. Moris, N., Pina, C., Arias, A.M.: Transition states and cell fate decisions in epigenetic landscapes. Nature Reviews Genetics **17**(11), 693–703 (2016)
9. Hänggi, P., Talkner, P., Borkovec, M.: Reaction-rate theory: fifty years after Kramers. Reviews of Modern Physics **62**(2), 251–341 (1990)
10. E, W., Vanden-Eijnden, E.: Transition-path theory and path-finding algorithms for the study of rare events. Annual Review of Physical Chemistry **61**, 391–420 (2010)
11. Goldman, M.J., et al.: Visualizing and interpreting cancer genomics data via the Xena platform. Nature Biotechnology **38**, 675–678 (2020)
12. Mermel, C.H., et al.: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology **12**, R41 (2011)
13. Cerami, E., et al.: The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery **2**(5), 401–404 (2012)
14. Verhaak, R.G.W., et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell **17**(1), 98–110 (2010)
15. Zhao, Z., et al.: Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. Genomics, Proteomics & Bioinformatics **19**(1), 1–12 (2021)
16. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
17. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
18. Efron, B., Gong, G.: A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician **37**(1), 36–48 (1983)
19. Schoenfeld, D.: Partial residuals for the proportional hazards regression model. Biometrika **69**(1), 239–241 (1982)
20. Zhang, J., et al.: Distinct aneuploid evolution of astrocytoma and glioblastoma during recurrence. npj Precision Oncology **7**, 97 (2023)

Disclosure of Interests. The author has no competing interests to declare.