



NET4BRAIN / SCIENTIFIC TRAINING

Final activity report

Almagro, Spain | 15–18 September 2026

Palacio de los Condes de Valdeparaíso, University of Castilla-La Mancha

Scientific coordination: Prof. Gabriel F. Calvo, MOLAB / UCLM

15

trainees

12

trainers and
contributors

10

one-hour
lectures

5

supervised
mini-projects

5

final group
presentations

Executive summary

The **Net4Brain Training School on Mathematical Oncology** brought together 15 trainees and 12 trainers and academic contributors in Almagro from 15 to 18 September 2026. All trainees were Net4Brain members, with ten countries represented. Organised at the University of Castilla-La Mancha and sponsored by Net4Brain COST Action CA22103 and Reichman University, the School combined specialist lectures with sustained work on five brain-tumour research problems.

Ten one-hour lectures occupied the mornings from Tuesday to Thursday. The programme began with brain tumour biology, clinical classification, imaging and treatment, then introduced mechanistic mathematical models and individual-based simulations. Sessions on mechanism-aware deep learning, multimodal artificial intelligence and virtual clinical trials showed how different computational approaches can address biomedical questions. Thursday's lectures extended the discussion to fusion genes, chimeric proteins and liquid-biopsy biomarkers. The sequence kept the biological problem in view as the methods became more specialised.

The mini-projects gave the School its practical focus. Five teams of three trainees each worked on a single problem throughout the three afternoons, supported by their designated supervisors and an additional project hour on Thursday. Their subjects ranged from glioblastoma spatial transcriptomics and early tumour evolution to radiotherapy resistance, FGFR3–TACC3 sequence and structure, and brain-metastasis growth measured by MRI. Core tasks made each problem manageable within the time available; more advanced questions left scope for further investigation. On Friday morning, all five teams presented their main findings to the other trainees and trainers.

The feedback was particularly positive about the quality of supervision. Trainees valued clear explanations, the opportunity to ask questions and the care taken to help them understand the concepts behind the work. Several comments described a new, more interdisciplinary perspective on the problems. The main suggestion for improvement was to allow more time for an ambitious programme of lectures and practical analysis.

As a follow-up, Prof. Annalisa Occhipinti has proposed inviting the five groups to an online Net4Brain early-career researcher meeting in October. This would give them an opportunity to develop their analyses and share their progress with the wider community. The School thus concluded with five group presentations and a concrete prospect of continued collaboration, supported by a set of project materials that can be used beyond the event.



1 Participation and organisation

1.1 A shared setting for different research backgrounds

For three and a half days, the Palacio de los Condes de Valdeparaíso in Almagro was the setting for an intensive programme on the mathematical and computational study of brain tumours. The School brought clinical and biological questions into the same discussion as equations, simulations and machine-learning analyses. Its purpose was to help trainees move between those perspectives: to understand the problem before choosing a method, and to return to the biology when interpreting a result.

The cohort comprised 15 trainees, all members of Net4Brain, with ten countries represented. Small teams made it possible to give each participant a substantial role in the practical work. The final team allocations are shown below.

Table 1: The 15 trainees and their mini-project teams.

No.	Trainee	Country	Mini-project
1	Aleyna Baltacioglu	Turkey	1
2	David Ben Dayan	Israel	5
3	Olawumi Giwa	Israel	5
4	Ewelina Jałonica	Poland	3
5	Panagiotis Karanikolaos	Greece	1
6	Michelle Klein	The Netherlands	3
7	Carles López-Mateu	Spain	2
8	Bárbara Marques	Portugal	5
9	Avital Morgenstern	Israel	4
10	Thu Nguyen	Belgium	2
11	Iryna Oliynyk	Ukraine	4
12	Patrycja Rosa	Poland	3
13	Ceyda Sönmez	Turkey	1
14	Mengyuan Wang	UK	4
15	Julia Zakaszewska	Poland	2

The teams remained with their projects throughout the School. This continuity mattered: a question encountered while examining the data on Tuesday could be revisited during Wednesday's analysis and explained in Friday's presentation. The format allowed trainees to combine their experience while receiving advice tailored to the particular problem they were addressing.

1.2 The training and supervision team

Twelve trainers and academic contributors supported the programme through lectures, mini-project supervision and the evaluation of the final presentations. The team brought together Hospital Universitario de Salamanca, the University of Castilla-La Mancha, Teesside University and Reichman University. Seven lecturers delivered the ten morning sessions; the broader group of mentors ensured that the practical work received sustained attention.

Table 2: Trainers, academic contributors and their responsibilities.

Name	Affiliation	Contribution
Luis A. Pérez-Romasanta	Radiotherapy Oncology Service, Hospital Universitario de Salamanca, Spain	Lectures 1 and 2.
Gabriel F. Calvo	Mathematical Oncology Laboratory (MOLAB), UCLM, Spain	Scientific coordination; Lecture 3; Mini-Project 5.
Juan Jiménez-Sánchez	MOLAB, UCLM, Spain	Lecture 4; Mini-Project 2.
Claudio Angione	School of Computing, Engineering and Digital Technologies, Teesside University, UK	Lecture 5; Mini-Project 1.
Annalisa Occhipinti	School of Computing, Engineering and Digital Technologies, Teesside University, UK	Lecture 6; Mini-Project 1.
Miguel Perales-Patón	MOLAB, UCLM, Spain	Lecture 7; Mini-Project 2.
Milana Frenkel-Morgenstern	Scojen Institute of Synthetic Biology, Dina Recanati School of Medicine, Reichman University, Israel	Lectures 8–10; Mini-Project 4.
Suraj Verma	Teesside University, UK	Mini-Project 1.
Kaveh Shahhosseini	Teesside University, UK	Mini-Project 1.
José Antonio Romero Rosales	MOLAB, UCLM, Spain	Mini-Project 3.
David G. Aragonés	MOLAB, UCLM, Spain	Mini-Project 3.
Víctor M. Pérez-García	MOLAB, UCLM, Spain	Evaluation of the Friday mini-project presentations and defences.

The supervision arrangements reflected the needs of each project. Claudio Angione and Annalisa Occhipinti worked with Suraj Verma and Kaveh Shahhosseini on the spatial-transcriptomics project. Juan Jiménez-Sánchez and Miguel Perales-Patón supervised the individual-based modelling project, while José Antonio Romero Rosales and David G. Aragonés guided the work on phenotypic resistance. Milana Frenkel-Morgenstern and Gabriel F. Calvo supervised the fusion-gene and brain-metastasis projects, respectively.

Several supervisors also taught the morning lectures. Trainees could therefore return to an idea introduced in a lecture with a specific question arising from their own analysis, rather than treating the theoretical and practical parts as separate activities.

2 The programme in Almagro

The School opened on Tuesday, 15 September, at 09:30. Its daily rhythm combined morning lectures with extended afternoon project sessions. There were three lectures on Tuesday, four on Wednesday and three on Thursday. This arrangement established the clinical and mathematical foundations early, while leaving time for the groups to develop their work before Friday's presentations.

Table 3: The ten morning lectures. All times are local (CEST).

Slot	Date / time	Lecture title	Lecturer
L1	Tue 15 September 10:00–11:00	Introduction to brain tumour biology I	Luis A. Pérez-Romasanta
L2	Tue 15 September 11:30–12:30	Introduction to brain tumour biology II	Luis A. Pérez-Romasanta
L3	Tue 15 September 12:30–13:30	Introduction to mechanistic mathematical models in neuro-oncology	Gabriel F. Calvo
L4	Wed 16 September 09:00–10:00	Agent-based and hybrid mathematical models for brain tumours	Juan Jiménez-Sánchez
L5	Wed 16 September 10:00–11:00	Mechanism-aware deep learning for biomedicine	Claudio Angione
L6	Wed 16 September 11:30–12:30	Integrating multimodal AI and biology	Annalisa Occhipinti
L7	Wed 16 September 12:30–13:30	From mathematical models to virtual clinical trials in neuro-oncology	Miguel Perales-Patón
L8	Thu 17 September 09:00–10:00	Fusion genes and chimeric proteins in brain cancers	Milana Frenkel-Morgenstern
L9	Thu 17 September 10:00–11:00	Liquid biopsy for diagnostic in brain cancers	Milana Frenkel-Morgenstern
L10	Thu 17 September 11:30–12:30	Biomarkers in liquid biopsy	Milana Frenkel-Morgenstern

The afternoon mini-project sessions ran from 15:30 to 19:00 on Tuesday, Wednesday and Thursday. An additional hour on Thursday, from 12:30 to 13:30, brought the scheduled project time to 11.5 hours before Friday's final preparation. Coffee breaks, lunch and dinner were also included in the daily programme.

Friday began with final preparation from 09:00 to 10:00. Presentations followed from 10:00 to 11:00 and resumed after the coffee break, with the closing session ending at 13:00. A farewell lunch from 13:00 to 14:00 concluded the School.

Training School timetable

15–18 September 2026 | Almagro, Spain | All times are local (CEST)



Hour/Day	Tuesday 15	Wednesday 16	Thursday 17	Friday 18
9.00-10:00h	Opening (From 9:30-10:00h)	Lecture 4: Juan Jiménez Sánchez	Lecture 8: Milana Frenkel-Morgenstern	Final preparation for Mini-Projects
10.00-11:00h	Lecture 1: Luis Pérez Romasanta	Lecture 5: Claudio Angione	Lecture 9: Milana Frenkel-Morgenstern	Presentation of Mini-Projects
11.00-11:30h	Coffee Break	Coffee Break	Coffee Break	Coffee Break
11.30-12:30h	Lecture 2: Luis Pérez Romasanta	Lecture 6: Annalisa Occhipinti	Lecture 10: Milana Frenkel-Morgenstern	Presentation of Mini-Projects and Closing (From 11:30-13:00h)
12.30-13:30h	Lecture 3: Gabriel F. Calvo	Lecture 7: Miguel Perales	Mini-Projects	Farewell Lunch (From 13:00-14:00h)
13.30-15:30h	Lunch Break	Lunch Break	Lunch Break	
15.30-19:00h	Mini-Projects	Mini-Projects	Mini-Projects	
19.00-20:30h	Free time	Free time	Free time	
20:30-22:00h	Dinner	Dinner	Dinner	



3 The lecture programme

3.1 Understanding the tumour and what imaging reveals

Lecture 1 | Luis A. Pérez-Romasanta | Tuesday, 15 September

Tuesday's teaching began with the clinical vocabulary needed for the rest of the School. Luis A. Pérez-Romasanta introduced brain tumour classification and radiology, showing how histology, molecular information and imaging contribute to an integrated diagnosis. Tumour grade, IDH status and 1p/19q co-deletion were discussed alongside the distinction between adult-type diffuse gliomas and central nervous system metastases. These were different biological and clinical entities, not interchangeable examples for a mathematical exercise.

Imaging made that distinction tangible. The lecture compared appearances of lower-grade glioma, glioblastoma and brain metastases, and covered MRI, CT, diffusion and perfusion imaging, magnetic resonance spectroscopy and PET. The examples raised a question that would recur in later sessions: what does a measured tumour volume actually represent?

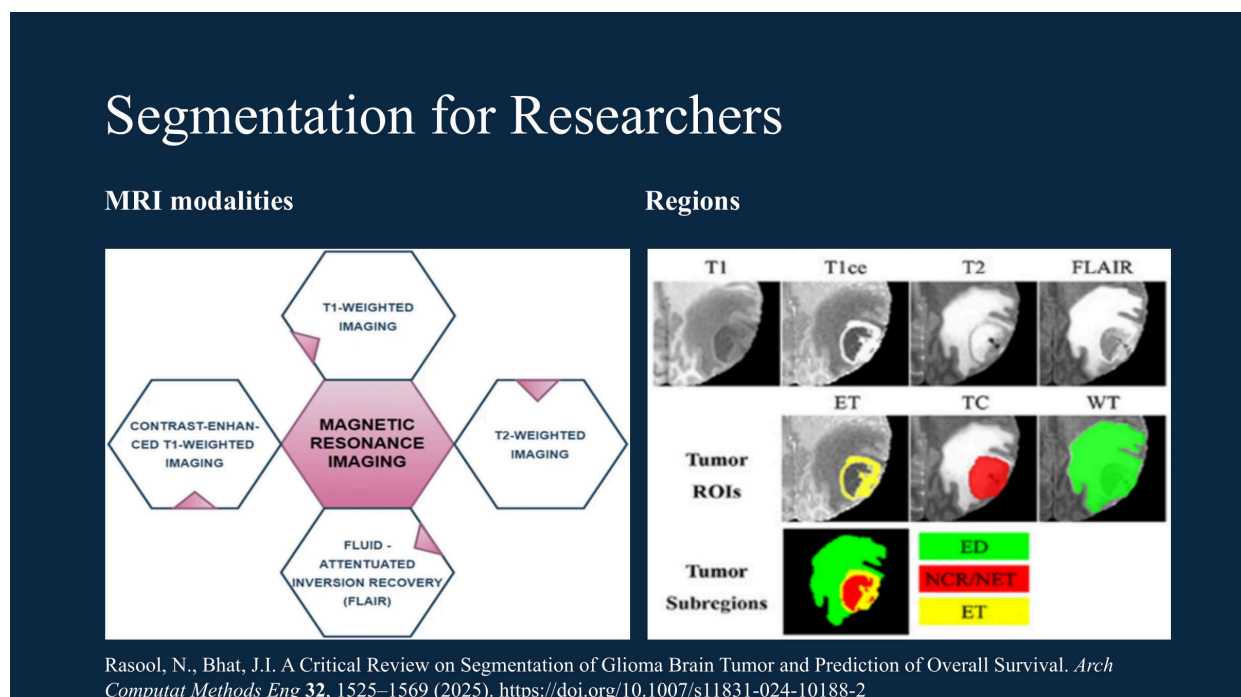


Figure 1: MRI sequences reveal different aspects of a lesion. The distinction between enhancing tumour, tumour core and whole tumour determines which image-derived quantity is being analysed or modelled.

A segmented region depends on the acquisition and on the definition used to delineate it. Enhancement, a tumour core and a whole-tumour region cannot simply be substituted for one another, nor do they directly count a single cell population. The distinction between segmentation for radiotherapy planning and segmentation for research further showed how the purpose of an analysis affects its measurements.

For trainees coming from mathematics or computing, this was an important starting point. Before fitting a model to an imaging variable, they needed to understand what was observed and how that observation related to the clinical question. The same issue later reappeared in the projects on tumour core and periphery and on serial MRI measurements of brain metastases.

3.2 Biology, microenvironment and treatment

Lecture 2 | Luis A. Pérez-Romasanta | Tuesday, 15 September

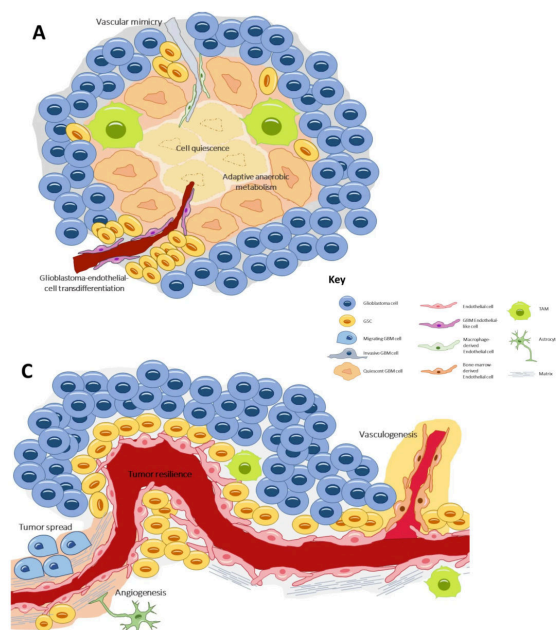
The second lecture examined the processes underlying the clinical and imaging appearances. Invasion through brain tissue, migration along white-matter structures and blood vessels, the neurovascular unit, hypoxia and vascular organisation all illustrated how strongly tumour behaviour depends on its surroundings. Necrotic and perivascular niches offered particularly clear examples of different pressures acting within the same lesion.

The molecular discussion included IDH-associated metabolic changes and the roles of markers such as ATRX, TP53, CDKN2A/B, EGFR, TERT and MGMT. Material on brain metastases brought in interactions with astrocytes, immune cells and neurons, as well as vascular interactions and metabolic adaptation. This breadth explained why the biological assumptions appropriate for a primary glioma cannot be transferred uncritically to a secondary brain lesion.

Local agresiveness and tumor microenvironments

- Agresiveness = Proliferation + Migration
- Driver: **Hypoxia**
- Result of:
 - Rapid proliferation and expansion of the tumor mass
 - The vascular system, although strongly stimulated to grow*, is unable to reach
- Microenvironments
 - the necrotic niche (A) – **Metabolic pressure**
 - the perivascular niche (C) – **Migration and plasticity pressure**
 - gradient-like configuration

*Neo-angiogenesis



Biomedicines 2022, 10(6), 1285; <https://doi.org/10.3390/biomedicines10061285>

Figure 2: Necrotic and perivascular niches expose tumour cells to different metabolic, migratory and plasticity-related pressures. These local conditions help explain why a spatially heterogeneous tumour may not be adequately described by one average growth rate.

Treatment was considered in that biological setting. Surgery, radiotherapy, temozolomide and other therapeutic approaches were linked to targeting, fractionation, tissue response and resistance. The discussion gave substance to processes that later appeared as mathematical terms: therapy acts on cells and tissues with different properties, and the resulting change in an image need not coincide with the moment of cell damage.

Together, the two opening lectures established a common point of reference for the School. Subsequent models of invasion, resistance and treatment response could be understood as attempts to isolate and investigate particular aspects of this more complex disease, while the clinical distinctions remained in view.



3.3 Mechanistic models: choosing what to represent

Lecture 3 | Gabriel F. Calvo | Tuesday, 15 September

Gabriel F. Calvo concluded Tuesday's morning programme by introducing the construction and use of mechanistic mathematical models in neuro-oncology. The starting point was a biological question and an explicit set of assumptions. From there, model formulation, calibration and validation formed an iterative process: a discrepancy between prediction and observation could prompt a change in the assumptions, the measurements or the model itself.

Simple growth laws—exponential, logistic, Gompertz and power-law—illustrated how different hypotheses lead to different descriptions of tumour volume. Scaling examples related macroscopic growth to heterogeneity and evolution. Moving from a total volume to a spatially distributed population then introduced the proliferation–invasion equation,

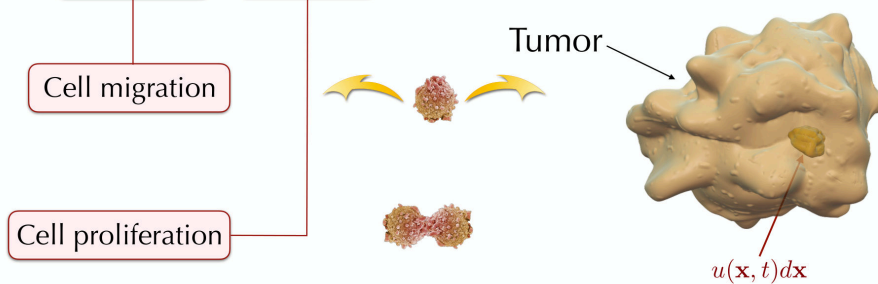
$$\frac{\partial u}{\partial t} = \nabla \cdot (D(\mathbf{x}) \nabla u) + \rho \left(1 - \frac{u}{K}\right) u,$$

where u denotes cell density, D the migration-related coefficient, ρ the proliferation rate and K a limiting density.

Proliferation-Invasion Mathematical Model

🔗 PDE capturing both tumor cell migration and growth:

$$\frac{\partial u}{\partial t} = \nabla \cdot (D(\mathbf{x}) \nabla u) + \rho \left(1 - \frac{u}{K}\right) u \quad \text{where } u = u(\mathbf{x}, t) \text{ is the cell density}$$



🔗 The solution of the **Proliferation-Invasion (PI) PDE** provides the **spatiotemporal evolution** of the tumor.

Figure 3: In the proliferation–invasion model, migration redistributes cells in space and the growth term changes their local density. The equation links a biological description of cell behaviour to the evolution of a tumour's spatial extent.

The separation of those terms made it possible to discuss invasion fronts and the relation between local cell dynamics and observable tumour extension. Treatment entered through additional model terms or discrete events, including radiation survival described by the linear–quadratic expression. Examples of treatment spacing, phenotypic variation and imaging biomarkers showed why a model must be chosen for a particular question: growth, spread, selection, response or prognosis.

This session provided the mathematical starting point for the afternoon work. Trainees were encouraged to explain the meaning of their variables and parameters before calculating with them, and to judge a model by the question it could answer rather than by its complexity.



3.4 Individual-based models: from cell rules to tumour behaviour

Lecture 4 | Juan Jiménez-Sánchez | Wednesday, 16 September

Wednesday began by changing the scale of description. Juan Jiménez-Sánchez introduced individual-based models, in which cells or other discrete entities follow local rules. Examples from the Ising model and cellular automata established how interactions between simple components can generate collective behaviour. Lattices, neighbourhoods, states and transition rules were defined before the examples turned to tumour populations.

Stochastic implementation was a substantial part of the lecture. Division, death, migration and quiescence could be represented as alternative events selected through Monte Carlo sampling. Showing how their probabilities divide the sampling interval made the connection between a biological event and a line of simulation code explicit.

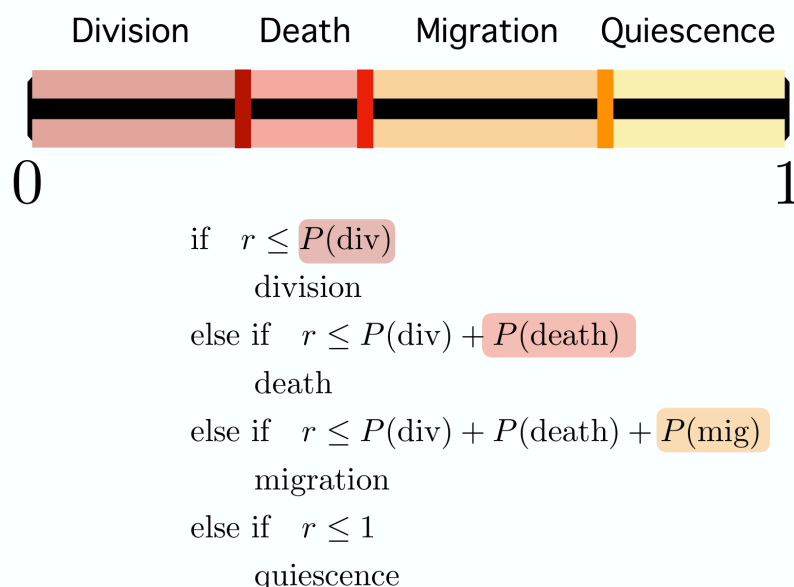


Figure 4: A Monte Carlo rule assigns division, death, migration or quiescence according to a random draw. The probabilities and the order in which updates are applied are part of the model and must be specified alongside the biological assumptions.

The discussion went beyond the event probabilities. Synchronous and asynchronous updating, daughter-cell placement, spatial representation and initial conditions can alter a simulation's behaviour. Approximate Bayesian computation illustrated one route to parameterisation, while mesoscopic approaches addressed the challenge of retaining relevant cell-level processes at a manageable computational cost.

These choices were especially relevant to the group studying early tumour evolution. A simulated increase in proliferation could only be interpreted once the rules governing inheritance, mutation and competition had been understood. The lecture therefore encouraged trainees to inspect how their results arose, rather than regard a plausible-looking spatial pattern as sufficient evidence. It also showed how individual-based and continuum descriptions could complement one another when different questions required different levels of resolution.

3.5 Mechanism-aware deep learning

Lecture 5 | Claudio Angione | Wednesday, 16 September

Claudio Angione examined how biological knowledge can enter a machine-learning analysis. Metabolic networks provided a central example: known biochemical relationships and model-derived quantities can shape the representation used by a learning system. The lecture presented mechanistic modelling and machine learning as approaches that can be combined in several distinct ways, each with implications for prediction and interpretation.

Some combinations use machine learning to analyse simulation outputs, approximate an expensive mechanistic model or learn from features constructed using that model. Others place biological information directly within a learning architecture or its optimisation. Distinguishing these arrangements was useful because an interpretation added after fitting is not the same as biological structure built into the analysis from the outset.

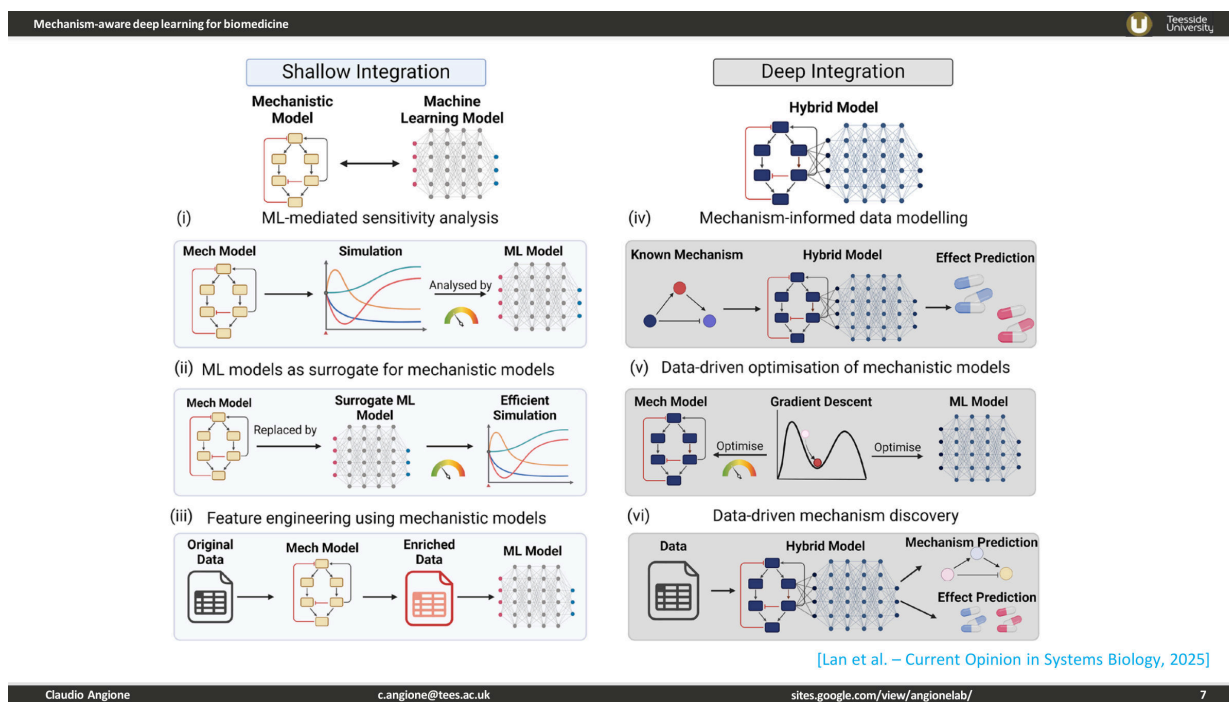


Figure 5: Different ways to combine mechanistic models and machine learning, from analysing simulations and constructing surrogate models to incorporating biological structure within a learning system. The choice determines how biological knowledge influences the prediction.

The examples included multimodal imaging and transcriptomic integration, learning when some modalities are missing, and spatially resolved metabolic modelling with hypergraph representations. They extended beyond brain cancer, allowing trainees to consider which methodological ideas could transfer to their own problems and which biological assumptions would need to change.

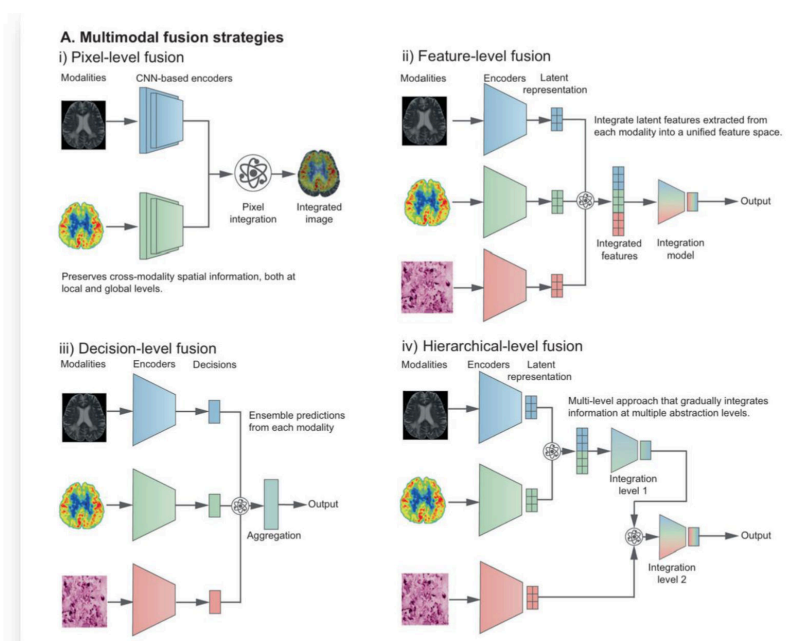
For the glioblastoma spatial-data project, the connection was immediate. Measured expression and transcriptome-inferred metabolic variables offered related but different descriptions of the same tissue. The practical question was whether the metabolic representation added useful information and whether the resulting predictions could be interpreted biologically. The lecture placed that exercise within a wider discussion of how mechanistic knowledge can guide learning without removing the need for careful testing.

3.6 Multimodal AI: making the analytical choices explicit

Lecture 6 | Annalisa Occhipinti | Wednesday, 16 September

Annalisa Occhipinti approached multimodal biomedical analysis through an accessible introduction to machine learning. A cell-classification exercise, short quizzes and worked examples clarified supervised and unsupervised learning, classification and regression, and the separation of training and test data. Support vector machines, decision trees, neural networks, nearest-neighbour methods and logistic regression illustrated the range of available approaches.

That introduction was particularly useful in a School whose participants did not all share the same computational background. It established the meaning of inputs, labels, features and testing before considering more complex combinations of imaging and omics. The lecture then distinguished pixel-level, feature-level, decision-level and hierarchical fusion, showing that multimodal analysis involves a choice about where and how information is combined.



Doan, L.M.T., Shahhosseini, K., Verma, S. et al. Bridging modalities with AI: a review of AI advances in multimodal biomedical imaging. *Commun Eng* 5, 30 (2026). <https://doi.org/10.1038/s44172-026-00602-x>

Figure 6: Pixel-, feature-, decision- and hierarchical-level fusion offer different ways to combine biomedical information. Each requires a clear definition of the inputs and of the stage at which one modality informs another.

Applications to classification, risk prediction, tissue analysis and explainability illustrated the consequences of these choices. Bulk, single-cell and spatial examples also showed how a finding could be examined at different biological scales. The value of an additional modality depended on what it contributed to the question, rather than on the number of data types assembled.

Metabolic modelling and fluxomic representations connected the lecture with the paired gene-expression and inferred-metabolism data used in Mini-Project 1. Together with Claudio Angione's preceding session, this provided the group with both a broad methodological perspective and concrete decisions to make in its own analysis: what to learn from, what to withhold for testing and how to interpret a predictive association.

3.7 From fitted models to virtual clinical trials

Lecture 7 | Miguel Perales-Patón | Wednesday, 16 September

Miguel Perales-Patón brought Wednesday's lectures back to a clinical question: how can a mathematical model be used to compare treatment strategies? The session followed the steps from experimental observations and biological knowledge to a calibrated model, validation, virtual patients and simulated treatment comparisons. A model had to be suitable for its intended use; representing every known mechanism was neither necessary nor a substitute for testing the chosen description.

A worked example followed sensitive, quiescent, transient-persister, persister, resistant and damaged tumour states. Chemotherapy and radiation acted through defined transitions and treatment events. Preclinical observations, including bioluminescence and survival curves, showed how these internal model states could be related to measured endpoints during calibration and validation.

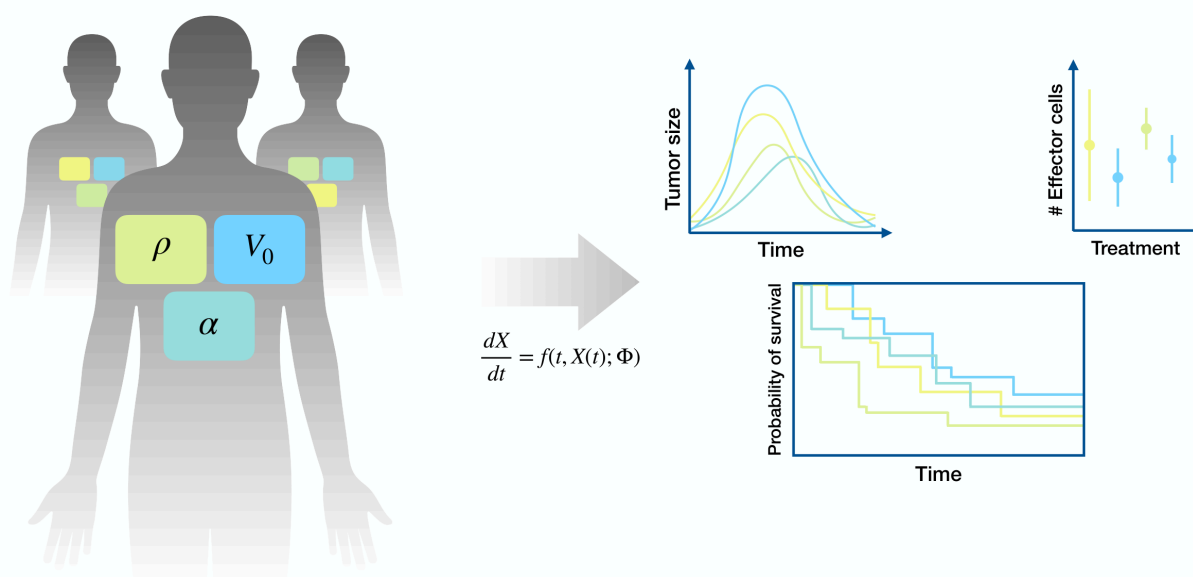


Figure 7: Virtual patients differ in parameters such as growth rate, initial volume and radiation response. Simulating these differences produces tumour trajectories and model-derived outcomes for comparing interventions across a virtual cohort.

Parameter sampling then extended the analysis from one fitted trajectory to a virtual cohort. Alternative treatment schedules could be compared within the same modelling framework. This made clear why fitting existing observations and predicting a response to a new intervention are different tasks, even when they use the same equations.

The lecture provided a translational connection between several parts of the School. The state-transition model related to the resistance mechanisms explored in Mini-Project 3, while the distinction between fitting and forecasting was central to the final-MRI prediction exercise in Mini-Project 5. Virtual trials were presented as a way to investigate treatment questions through calibrated models, with experimental evidence remaining essential to their interpretation.

3.8 Fusion genes and chimeric proteins

Lectures 8–10 | Milana Frenkel-Morgenstern | Thursday, 17 September

Thursday's three lectures, delivered by Milana Frenkel-Morgenstern, followed a connected set of molecular questions: how fusion genes arise, how their products can be identified and what they may offer as biomarkers. This shifted the focus from populations and tissue measurements to the relationship between genomic organisation, RNA sequences and protein architecture.

Chimeric RNAs and their possible origins introduced the subject. BCR–ABL1 provided a familiar general example before the discussion turned to FGFR3–TACC3 in glioblastoma. Attention to parental domains, fusion junctions and retained kinase and coiled-coil regions showed why identifying a fusion by name is only the beginning of a molecular analysis.

Variability of the fusion junction site of FGFR3-TACC3 gene

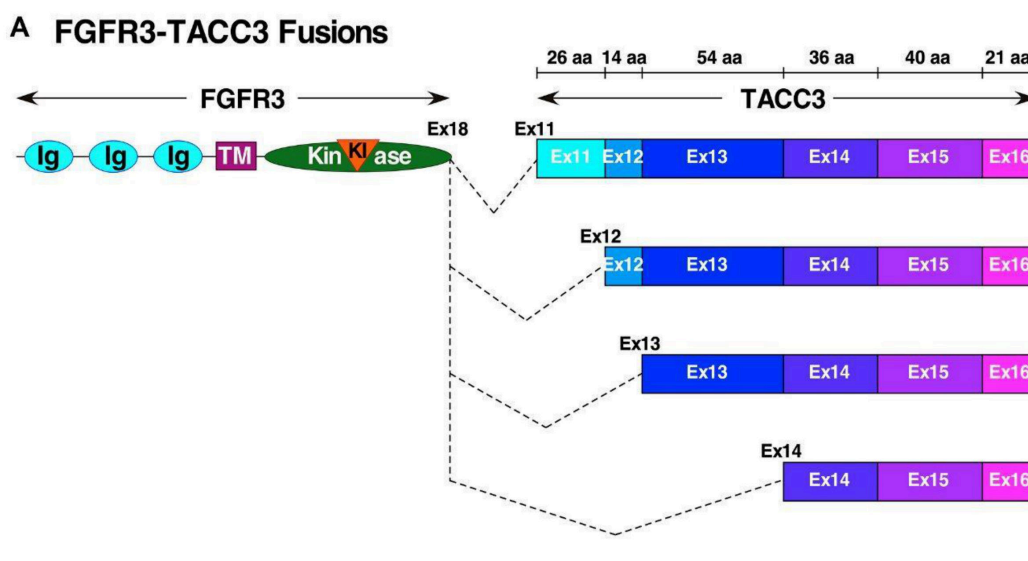


Figure 8: FGFR3–TACC3 fusions can differ in their junctions and retained sequence. A structural interpretation therefore begins with a specific transcript and coding sequence, rather than with the fusion name alone.

The variation between junctions was particularly relevant to Mini-Project 4. Its trainees were asked to determine which sequence they were analysing, whether it supported the proposed reading frame and which parental domains were retained before interpreting a predicted protein structure. These checks tied computational analysis to an identifiable molecular object.

The lecture thus extended the School's treatment of heterogeneity to another scale. Differences between tumours can involve not only cell composition, growth or treatment response, but also the precise architecture of an altered gene product. Understanding those differences requires attention to the sequence and its biological consequences, alongside the tools used to retrieve, align, translate and model it.

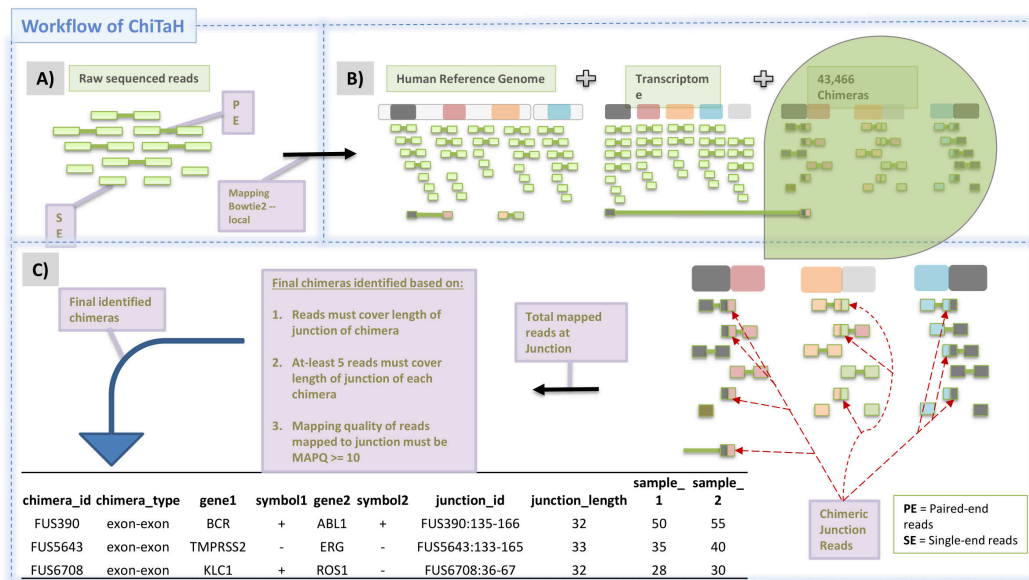


3.9 Fusion detection and liquid-biopsy biomarkers

Lectures 8–10, continued | Milana Frenkel-Morgenstern | Thursday, 17 September

The Thursday sessions also examined how a candidate chimeric sequence is recognised in sequencing data. The ChiTaH workflow followed the path from raw reads and reference sequences to candidate junctions, supporting reads and a structured output. Examples of PCR/Sanger validation and detection at single-cell level connected a computational candidate to further molecular evidence.

Our Method for Identification of Chimeric RNAs- ChiTaH



Detroja R., et al, Frenkel Morgenstern M., NAR Genomics and Bioinformatics 2022

Figure 9: The ChiTaH workflow links read mapping and reference chimeras to junction-level support. Organising that evidence is an essential step between detecting a candidate sequence and assigning it biological significance.

This analytical perspective complemented the earlier discussion of protein architecture. An attractive biological explanation does not establish that a fusion has been detected reliably; conversely, detecting a sequence does not by itself establish all of its functional consequences. The session showed how different kinds of evidence contribute to these separate questions.

Liquid-biopsy applications placed the molecular work in a wider diagnostic setting alongside tissue examination, molecular testing and imaging. Peripheral blood, cell-free DNA, mutations and fusions were considered in relation to early diagnosis, monitoring and personalised treatment. These were presented as research directions and potential applications, with the distinction between a promising molecular signal and a clinically useful biomarker remaining important.

The three lectures rounded out the School's scientific range. From an MRI-defined lesion to a phenotypic distribution or a fusion junction, each measurement supported particular conclusions and left other questions open. The afternoon projects gave trainees a setting in which to work through those distinctions for themselves, using data and models at several different biological scales.

4 Five projects, three afternoons

4.1 From a research question to a group presentation

The mini-projects were at the centre of the School's teaching format. Each team stayed with one problem from Tuesday to Thursday, with close guidance from the supervisors responsible for that project. This gave trainees time to move beyond a first attempt: to inspect the data, make choices, encounter difficulties and return to their assumptions before presenting the work.

The five subjects differed considerably. They involved spatial gene expression, individual cells, phenotypic distributions, fusion sequences and MRI-derived lesion volumes. Each nevertheless required a group to connect a biological question with a tractable analysis and explain what its results could mean. The modelling cycle introduced on Tuesday offered a useful way to organise that process.

Mathematical Modelling in Biomedicine

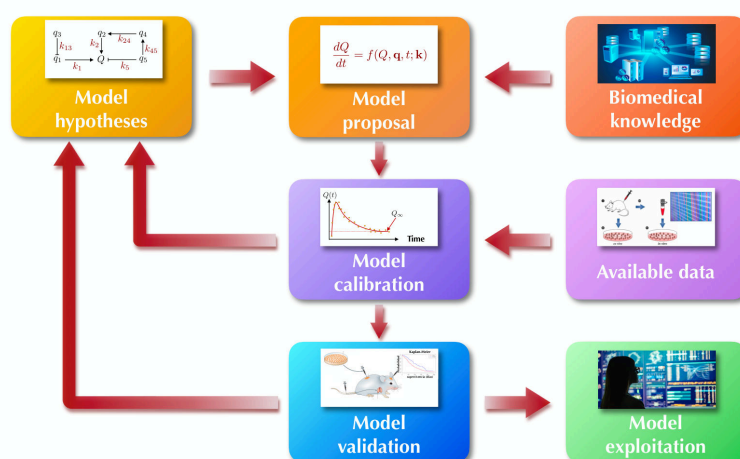


Figure 10: Modelling as an iterative activity: biological knowledge and hypotheses inform a model, observations constrain it, and validation can lead back to revised assumptions. The practical sessions gave trainees time to revisit these steps within one problem.

Supervision combined an introduction to the problem with continuing discussion and targeted help. Core tasks gave each group a feasible starting point; optional extensions allowed further exploration without making completion of every task a condition for a worthwhile presentation. The aim was an analysis that the trainees could understand and defend, with its assumptions and limitations made explicit.

The additional project hour on Thursday supported the transition from analysis to presentation. On Friday, all five groups shared their main findings with the whole School. The following sections describe the questions and methods around which that work was organised. The five complete project briefs are included as appendices for continued study.

4.2 Mini-Project 1: glioblastoma spatial transcriptomics

Title: Developing a Machine Learning Pipeline to Characterise Tumour Core and Periphery on Glioblastoma Spatial Data.

Supervisors: Claudio Angione, Annalisa Occhipinti, Suraj Verma and Kaveh Shahhosseini.

Team: Aleyna Baltacioglu, Panagiotis Karanikolaos and Ceyda Sönmez.

Research question. Which transcriptomic and transcriptome-inferred metabolic programmes distinguish tumour core from periphery in an unseen patient, and how robust and biologically interpretable is that distinction?

This project made the spatial heterogeneity introduced in the opening lectures the subject of a machine-learning problem. The teaching dataset covered six patients and 15 tissue sections, with approximately 30,000 spatial spots, 27,573 gene features and 168 inferred metabolic-module scores. Coordinates, patient and section identifiers, and regional labels linked each measurement to its tissue setting. Core and periphery were operational labels within a richer biological architecture, rather than two homogeneous populations.

Aleyna, Panagiotis and Ceyda were asked to compare expression-based and metabolism-based representations, using a majority-class baseline and regularised logistic regression as accessible starting points. A supporting Python notebook reduced the initial programming burden, leaving time to consider the scientific choices. The metabolic variables were inferred from the same expression profiles, not independently measured metabolites; understanding that relationship was essential to interpreting any benefit from combining them.

The central test was whether the analysis could generalise to an unseen patient. Preprocessing and model selection therefore belonged within patient-grouped training folds, with patients withheld from fitting used for evaluation. Comparing this design with a random split of spots offered a way to expose an important source of overoptimistic performance: nearby or same-patient observations can be highly correlated even when they appear as different rows in a dataset.

Interpretation extended beyond a classification score. Spatial maps of errors, calibration, per-patient performance and feature stability could reveal where and why a model struggled. Excluding genes involved in the original annotation rules was another useful check: recovering the rules behind a label is not equivalent to finding an independent biological signature. The project brought Claudio and Annalisa's lectures into a concrete glioblastoma analysis while asking the team to judge a prediction by its biological meaning and its performance on genuinely withheld data.

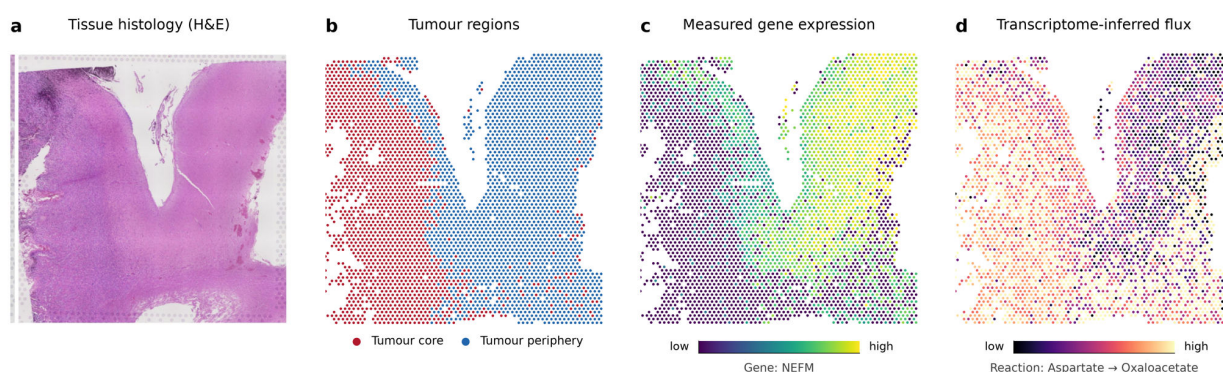


Figure 11: Histology, regional labels, gene expression and inferred metabolic activity provide related views of the same tissue. The analysis asks which signals distinguish core from periphery across patients.

4.3 Mini-Project 2: early tumour evolution

Title: Early stages of brain tumour development through the lens of individual-based models.

Supervisors: Juan Jiménez-Sánchez and Miguel Perales-Patón.

Team: Carles López-Mateu, Thu Nguyen and Julia Zakaszewska.

Research question. How do heritable phenotypic changes and local competition for space shape tumour growth, diversity and the spatial distribution of proliferative traits?

Carles, Thu and Julia approached tumour development from the scale of individual cells. Their starting point was a small population on a two-dimensional lattice, with each cell carrying a phenotypic trait that affected its division rate. Division, inheritance, phenotypic change and placement of daughter cells were enough to pose a substantial evolutionary problem. Migration, death and more elaborate spatial descriptions could be added later, once the basic simulator was understood.

Building that simulator required decisions with biological consequences. How should neighbours be defined? What happens at a boundary or when surrounding sites are occupied? Are cells updated together or one at a time? Which changes are inherited? These questions linked the practical work directly to Juan Jiménez-Sánchez's lecture: the computational rules were part of the explanation, not details that could be hidden behind a simulation.

The comparative experiments centred on two contrasts. One concerned frequent small changes versus rare large changes in the proliferative trait. The other concerned contact inhibition versus replacement of neighbouring cells through competition. Cell number, mean phenotype, diversity indices and the location of the fittest cells relative to the tumour centre provided complementary ways to examine the resulting populations. Repeated simulations and changes in parameters were needed to distinguish persistent patterns from stochastic variation.

The planned progression moved from an initial implementation to comparisons between scenarios and then to interpretation. A working model did not have to reconstruct the history of a particular patient's tumour to be informative. Its value lay in showing how a proposed cellular rule could affect growth and heterogeneity, and in making that dependence open to testing. For the final presentation, the challenge was to explain the relationship between the rules chosen, the numerical observations and the evolutionary questions that motivated them.

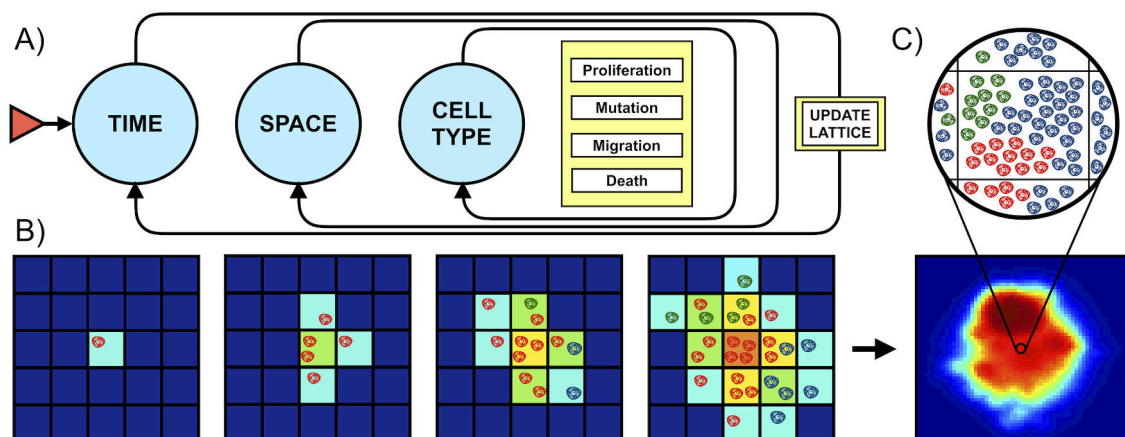


Figure 12: Cell-level rules determine how a simulated population expands and diversifies. Comparing rules and repeated runs helps connect spatial patterns to the processes that generate them.

4.4 Mini-Project 3: phenotypic dynamics and resistance

Title: Designing a Model to Distinguish Mechanisms of Therapy Resistance Acquisition through Phenotypic Measurements Across Time and Radiation Doses.

Supervisors: José Antonio Romero Rosales and David G. Aragonés.

Team: Ewelina Jałonica, Michelle Klein and Patrycja Rosa.

Research question. Can changing phenotypic distributions help distinguish differential proliferation, differential death and reprogramming as explanations for enrichment of a resistant glioblastoma state?

The third project examined a difficulty at the heart of resistance modelling: a population can change because some cells survive or proliferate more successfully, because surviving cells alter their state, or through a combination of both. Ewelina, Michelle and Patrycja considered this question through the vessel-co-opting resistant, or VC-Resist, state and Nestin-associated measurements in MGG4 glioblastoma cells.

Their teaching data consisted of 42 smoothed probability profiles, covering six radiation-dose conditions from 0 to 12 Gy over seven days. These profiles described the distribution of a phenotypic measurement rather than a complete record of individual cells. The modelling task was to formulate a phenotype-structured partial differential equation using the base-10 logarithm of the Nestin-associated variable as its coordinate.

Proliferation, death and phenotypic convection or reprogramming provided candidate terms in the equation. The team had to decide how those processes should depend on phenotype, time and dose, and which quantities could reasonably be fitted. An analytical or numerical solution simple enough for repeated comparison with the data was an important practical requirement. Positivity, smoothness, normalisability and plausible dose dependence were part of judging the formulation.

The distinction between a normalised distribution and a population size was especially important. Matching the shape of a profile does not automatically identify the separate contributions of growth, death and state change. The project therefore asked for a biological interpretation of the fitted description, sensitivity to its assumptions and consideration of further measurements that could distinguish competing mechanisms. It connected the continuous models introduced on Tuesday with the treatment-state transitions discussed on Wednesday, while leaving the trainees responsible for choosing and justifying their own mathematical formulation.

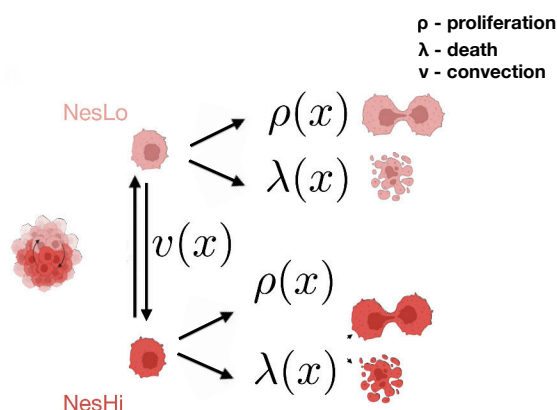


Figure 13: Proliferation, death and phenotypic change contribute differently to the evolution of Nestin-associated cell distributions.

4.5 Mini-Project 4: the FGFR3–TACC3 fusion

Title: Exploration of the FGFR3-TACC3 fusion gene in glioblastoma.

Supervisor: Milana Frenkel-Morgenstern.

Team: Avital Morgenstern, Mengyuan Wang and Iryna Oliynyk.

Research question. Which sequence and structural features of FGFR3–TACC3 support its role as an oncogenic fusion in glioblastoma, and which conclusions require further experimental evidence?

Avital, Mengyuan and Iryna followed a molecular change from its sequence to its possible structural consequences. FGFR3–TACC3 combines regions from two parental proteins, making the retained receptor domains, the kinase region and the TACC3-derived coiled-coil region central to its interpretation. The project connected directly with Milana Frenkel-Morgenstern's lectures, but began with an essential practical step: establishing exactly which fusion sequence was being analysed.

The initial task was to assemble a panel of at least six suitable records, subject to availability and checks on their origin and annotation. A complete transcript, a coding sequence and a short fragment spanning a junction do not support the same analysis. Engineered constructs also needed to be distinguished from other records. Alignments with the parental references, junction checks and translation into a justified open reading frame provided the basis for deciding what protein sequence each record actually supported. Multiple database entries were not to be counted automatically as different biological variants.

The translation stage also required a choice that can easily be overlooked when using an automated tool. Six-frame translation could reveal candidate coding regions, but orientation, parental coding phase, stop codons and junction coverage had to justify the chosen open reading frame. A partial record could support a local junction without supporting a complete protein model. This kept the structural interpretation tied to the sequence actually available.

Detailed structural work focused on one well-supported complete fusion sequence. Domain annotation and predicted structure were to be considered together with local confidence, relative domain placement and comparison with experimentally determined parental or homologous domains. Secondary-structure assignment and domain-level classification offered further ways to interpret the model. A sequence-matched backup structure allowed the analysis to proceed if a live prediction run proved impractical within the School.

The assignment required care in moving from an image of a protein to a claim about its function. A plausible local fold does not by itself demonstrate kinase activity, an oligomeric assembly or drug sensitivity. Sequence tables, alignments, translated sequences, annotated structural views and confidence information were therefore intended to support a reasoned molecular account. In a programme largely concerned with cells and tissues, this project showed how the same attention to assumptions and evidence applies at the scale of a gene product.

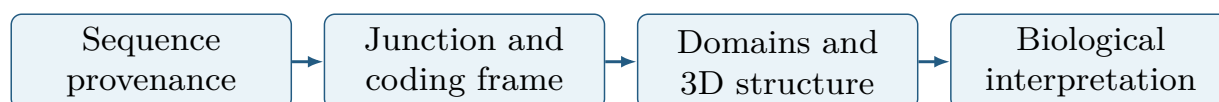


Figure 14: Sequence identity and annotation support each subsequent step, from checking a coding frame to interpreting a predicted structure.

4.6 Mini-Project 5: brain-metastasis growth and radiotherapy

Title: Modelling Brain-Metastasis Growth and RT Response.

Supervisor: Gabriel F. Calvo.

Team: David Ben Dayan, Olawumi Giwa and Bárbara Marques.

Research question. Can a simple ODE model explain enhancing and central non-enhancing lesion volumes over time and predict a later MRI measurement after radiotherapy?

The fifth project returned to the clinical questions in the opening lectures. David, Olawumi and Bárbara investigated lesion-level growth and radiotherapy response with prepared MRI-derived volume and treatment-history extracts from OpenBTAI, a resource containing 75 patients and 260 brain metastases.

Before an equation could be fitted, patient, lesion and visit identifiers had to be matched, units checked and treatment dates placed on a consistent timeline. The unit of analysis was a single lesion followed through time. A later scan could provide an initial condition, but could not make earlier treatment exposure disappear. Patient-wide therapies also had to be distinguished from treatment of an individual metastasis.

The modelling problem separated total, contrast-enhancing and central non-enhancing volumes. A key biological observation motivated that separation: radiation may remove a cell's capacity to reproduce before damaged tissue disappears from an image. Candidate compartments could therefore distinguish proliferative and lethally damaged populations, with an explicit account of how they contributed to the observed volumes. Equally, an image region labelled “necrosis” was not, on that basis alone, a histological diagnosis of radiation necrosis.

The task moved from checking trajectories and estimating a simple growth baseline to formulating balances, fitting a small number of parameters and testing a prediction. The final MRI was to be withheld from fitting and compared with the forecast. Persistence—carrying the preceding measurement forward—provided a deliberately simple benchmark. Total volume was not an independent third measurement when it was already the sum of the enhancing and non-enhancing components.

A robustness check and biological interpretation completed the proposed analysis. A model that failed to improve on persistence could still be informative if the group could explain why: reproducing earlier observations and making a useful prediction are different tests.

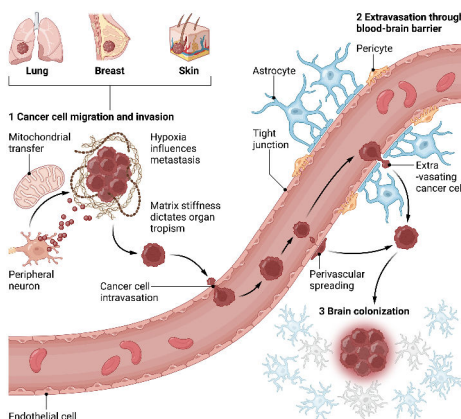


Figure 15: The metastatic cascade establishes the biological context. This modelling exercise began with lesions already visible on MRI and followed their subsequent growth and treatment response.

5 Friday: presenting the work and reflecting on the experience

Friday morning brought the five teams together to present their main findings. After the final preparation period, each group addressed the full audience of trainees and trainers. The work developed in the smaller supervised sessions was now shared across the School, from spatial classification and tumour evolution to phenotypic resistance, fusion structure and lesion-level response. Víctor M. Pérez-García contributed to the evaluation of the presentations and defences.

The presentations gave each team responsibility for explaining its problem and analysis to colleagues who had worked on something different. They also brought the practical work back into the broader scientific discussion established by the lectures. All five groups presented before the closing session and farewell lunch.

5.1 What trainees valued

A brief survey after the presentations invited trainees to reflect on their experience. The comments repeatedly returned to the mini-projects and the quality of supervision. One response described an interesting problem and supervisors who encouraged the group to explore it more deeply. Another explained why the mentoring had mattered:

“The mentors were incredibly kind, supportive, and clear in their explanations. They put a great deal of effort into helping us truly understand the concepts rather than simply completing the tasks. Working on this project also gave us a completely different perspective and helped us approach the problem in a new and more interdisciplinary way.”

A further comment emphasised that the helpful and supportive teaching made it easier to engage with unfamiliar topics and ask for clarification. These observations capture an important part of the School's approach: time spent explaining a method was as valuable as time spent implementing it.

5.2 The value of more time

Positive comments about the organisation and the combination of theory and practice were accompanied by a constructive suggestion:

“The Training School was very well organized, and I really enjoyed both the theoretical and practical sessions. I think the course could be even more beneficial if it were held over a longer period, as there were many interesting and valuable topics to cover in a relatively short time.”

The request for more time is particularly relevant to future editions. The School covered a broad scientific range, and the project format invited questions that could not all be resolved in three afternoons. These qualitative responses support retaining the close supervision and balance of lectures and practical work, while considering a longer period for project development.

6 Continuation, lessons and acknowledgements

6.1 Continuing the work beyond Almagro

The trainers' messages after the School echoed the positive trainee feedback. Claudio Angione reported that students had been extremely happy and genuinely thankful for the opportunity to attend. Annalisa Occhipinti also reported many favourable comments and praised the work on the mini-projects.

Annalisa proposed inviting the groups to present at an online Net4Brain early-career researcher meeting in October. The proposal would give the five teams several further weeks to develop their analyses, revisit questions raised by the Friday presentations and share an update with the wider community. It offers a practical continuation of the School without requiring every open question to have been settled before leaving Almagro.

6.2 What to retain and what to develop

The small-group format and sustained supervision were central to the experience described by trainees. Keeping a team with one focused question made it possible to follow the consequences of its decisions, rather than move rapidly between unrelated exercises. Accessible starting tasks and optional extensions helped accommodate different levels of experience while preserving room for independent work.

The clearest opportunity for improvement is time. A longer School, or a planned period of follow-up after it, would allow more room for refinement, interpretation and discussion. Collecting the groups' presentations and available analyses would support that continuation and provide useful material for later training activities.

The main achievement of the Almagro School was to bring a broad set of approaches into a shared working setting. Clinical and biological lectures gave meaning to the quantities being analysed; mathematical models made assumptions about growth and treatment explicit; machine learning offered ways to investigate complex molecular and spatial data; and the fusion-gene sessions added a detailed molecular perspective. The five projects turned those themes into problems that trainees could work on together and present to their peers. The positive feedback was a recognition of both the scientific programme and the care taken in its delivery.

6.3 Acknowledgements

The School was a collective effort. Sincere thanks are due to the 15 trainees for their engagement and collaborative work, and to the 12 trainers and academic contributors for their lectures, supervision and participation in the final presentations. The contributions of Hospital Universitario de Salamanca, Teesside University, Reichman University and the University of Castilla-La Mancha brought complementary expertise to the programme.

Particular thanks go to the local organising and administrative team at UCLM and the Mathematical Oncology Laboratory, including Staff Manager María Jesús Pérez Moraga and the Fundación General de la Universidad de Castilla-La Mancha. Their work supported the practical arrangements throughout the School. The financial support of **Net4Brain COST Action CA22103 and Reichman University** is gratefully acknowledged.



Net4Brain
Let's beat brain cancer together
Training School on Mathematical Oncology

Net4Brain Training School on Mathematical Oncology

15-18 September 2026, Palacio de los Condes de Valdeparaíso, University of Castilla-La Mancha, Almagro (Spain)



Figure 16: After four days of modelling brain tumours, one final fitting problem: getting everyone into the same frame! The **Net4Brain Training School** participants in the garden courtyard of the Palacio de los Condes de Valdeparaíso, Almagro.



Mini-Project: *Developing a Machine Learning Pipeline to Characterise Tumour Core and Periphery on Glioblastoma Spatial Data*

Supervisors: *Claudio Angione, Annalisa Occhipinti, Suraj Verma & Kaveh Shahhosseini*

Motivation

Glioblastoma (GBM) is the most aggressive diffuse primary brain tumour in adults. Its poor outcome is linked not only to rapid growth but also to extensive infiltration: malignant cells disperse beyond the tissue that can be identified and removed surgically, remain within the surrounding brain and contribute to near-inevitable recurrence. A specimen described as “tumour core” therefore coexists with a spatially extended periphery containing variable mixtures of infiltrating malignant cells, neurons, oligodendrocytes, astrocytes, microglia, vascular cells and extracellular tissue components [1,2].

This spatial heterogeneity matters biologically. Local nutrient and oxygen availability, vascular remodelling, immune-cell composition and interactions with normal neural lineages can favour different malignant states and treatment responses. Bulk sequencing averages these signals over an entire specimen. Single-cell RNA sequencing resolves cell populations but normally loses their original tissue coordinates. Spatial transcriptomics addresses this gap by measuring gene expression at thousands of positions while preserving where each measurement was made.

The human GBM atlas developed by Sonpatki and colleagues combines spatial gene-expression profiles, high-resolution histology, matched single-cell data and subcellular spatial measurements to describe anatomically organised tumour niches. The atlas reveals a structured transition from leading-edge and infiltrating regions towards cellular tumour, vascular, immune-glial and hypoxic niches, rather than a homogeneous mass [2,3]. This creates a natural quantitative problem: can a machine-learning (ML) model recognise an operational distinction between core and periphery in tissue from a patient that was not used to train the model?

This Mini-Project uses two related feature representations at each tissue spot: measured gene expression and metabolic-module activity computationally inferred from that expression. The aim is not merely to obtain a high classification score. It is to determine whether the result generalises across patients, which biological signals drive the prediction, whether the metabolic representation adds useful information and where the classifier fails in tissue space. No prior coding experience is assumed: the project is supported by an easy-to-follow Python notebook in which the principal technical steps are demonstrated and then reproduced or adapted by each group.

Central question: Which transcriptomic and transcriptome-inferred metabolic programmes distinguish tumour core from tumour periphery in an unseen patient, and how robust and biologically interpretable is that distinction?

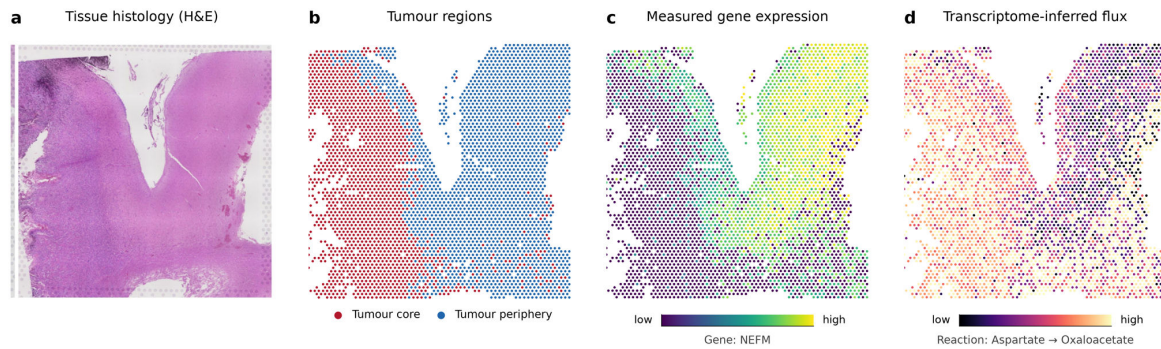


Figure 1: Project data modalities: (a) H&E tissue image; (b) binary regional labels supplied for the project; (c) measured spatial expression of the example gene *NEFM*; and (d) an example transcriptome-inferred metabolic flux score. The metabolic layer is derived computationally from the expression profile of the same spot.

Specific biological context of the problem

A binary task extracted from a richer spatial architecture. The complete atlas distinguishes cellular tumour (CT), infiltrating tumour (IT), leading edge in grey or white matter, immune-glial niches, vascular regions and hypoxic or perinecrotic compartments. The underlying regional framework is based on the Ivy Glioblastoma Atlas Project (Ivy GAP), which defines histologically recognisable regions of GBM tissue, and was refined by Sonpatki and colleagues for the tissue sections in their atlas. The present teaching dataset reduces part of this architecture to the operational classes tumour core and tumour periphery. The exact mapping from the original annotations to these two labels must be recorded in the accompanying data dictionary. “Periphery” must not be interpreted as purely healthy brain, nor “core” as a single homogeneous cell population.

Core and peripheral tissue differ in composition. In matched specimens, core samples contain a larger malignant-cell fraction. Proliferating and mesenchymal-like tumour states, blood-derived macrophages, proliferating microglia and regulatory T cells are more prominent in core tissue, whereas resting microglial populations are more prevalent in peripheral tissue. The atlas also identifies an oligodendrocyte state, *Oligo_2_3_2*, enriched in core and perivascular regions and associated with adverse clinical features [2]. Along the spatial transition from white-matter leading edge to infiltrating tumour and cellular tumour, normal neural-lineage signals progressively coexist with reactive astrocytes, microglia and emerging malignant-cell programmes.

A spot is a tissue mixture, not an individual cell. A Visium spot has a diameter of approximately $55\mu\text{m}$ and can contain transcripts from several cells. The classifier therefore predicts a tissue-region label from a mixed molecular profile; it does not identify the phenotype of one cell. Important predictors may reflect malignant-cell programmes, local non-malignant composition or both.

The labels have provenance. In the source atlas, anatomical-feature assignment combines histology, marker-gene scores, inferred cell states and hierarchical rules. Some predictor genes may consequently be related to the rules used to define the labels. A robustness analysis that removes annotation-marker genes is therefore required before interpreting a gene as an independently discovered regional signature.

The metabolic variables are inferred, not directly measured. The 168 metabolic-module values are estimated with scFEA, a graph neural network that infers reaction-level metabolic flux from transcriptomes while constraining its predictions through the stoichiometry of a metabolic module network [4]. They may compress correlated expression into a biologically meaningful representation, but they are not independent metabolomic measurements and should be described as transcriptome-inferred activity or flux scores.

Source article, anatomical atlas and scFEA resources

The links for direct access:

- Sonpatki *et al.*, *Nature Communications* (2026): <https://www.nature.com/articles/s41467-026-69716-2>
- Ivy Glioblastoma Atlas Project: <https://glioblastoma.alleninstitute.org/>
- scFEA method paper: <https://genome.cshlp.org/content/31/10/1867>
- scFEA software repository: <https://github.com/changwn/scFEA>

Dataset and machine-learning formulation

The project subset contains IDH-wild-type GBM tissue from six patients and 15 sections (2–5 sections per patient), profiled with the 10x Genomics Visium platform, with approximately 30,000 spatial spots. It is derived from the spatial atlas of Sonpatki *et al.* [2]. Every spot has a patient identifier, section identifier, spatial coordinates, a binary regional label, a normalised gene-expression profile and 168 transcriptome-inferred metabolic-module scores.

Data component	Scale	Dimension	Role in the project
Metadata	Spot	6 patients; 15 sections	Patient, 2–5 sections per patient, spatial coordinates, preservation or batch variables when available, and core/periphery label
Gene expression	Spot	27,573 genes	Normalised transcript counts measured with 10x Genomics Visium, together with the spot locations; a filtered subset will normally be used for the core analysis
Metabolic representation	Spot	168 modules	Activity of each metabolic module, estimated computationally from the expression profile of the same spot using scFEA

The files that the group will use will be pre-aligned by spot identifier and provided in ready-to-load tabular and sparse-matrix form, so that no afternoon is lost to raw-data wrangling. A data dictionary will document the anatomical-label mapping, coordinates, feature names and available technical covariates.

For spot i , let $y_i = 1$ denote core and $y_i = 0$ denote periphery. Let $\mathbf{x}_i$ contain the selected gene features, metabolic features or their combination. The core classifier is a regularised logistic regression,

$$p_i = \Pr(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp[-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})]}.$$

An elastic-net fit estimates the coefficients by minimising the training loss plus a mixture of L^1 and L^2 penalties,

$$\mathcal{J}(\beta_0, \boldsymbol{\beta}) = -\frac{1}{n_{\text{tr}}} \sum_{i \in \mathcal{I}_{\text{tr}}} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1 - \alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right].$$

Here λ controls the overall regularisation and α balances sparsity against coefficient shrinkage [5]. The principal comparison uses three feature configurations: genes only, metabolic modules only, and genes plus modules.

Non-negotiable validation rule

All spots and all sections from the same patient must remain in the same fold. Scaling, missing-value handling, gene filtering, feature selection and hyperparameter tuning must be learned from training patients only. A random split of neighbouring spots would leak patient- and section-specific information and would not test generalisation to a new patient.

Recommended baseline hierarchy

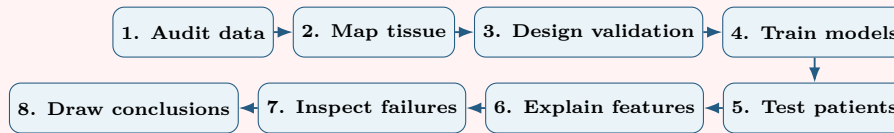
Compare: (i) a majority-class predictor; (ii) metabolic modules only; (iii) a reduced gene-expression model; and (iv) the combined representation. Use patient-grouped cross-validation for model selection and reserve at least one patient, or one full outer fold, for the final untouched evaluation. Class weights may be used when the labels are substantially imbalanced.

Teaching support and Python protocol

No prior coding experience is required. An easy-to-follow, step-by-step Python notebook will supply the data-loading, sparse-matrix, plotting and model-fitting scaffolding. Short demonstrations will introduce each technical step; students will then repeat or adapt it and remain responsible for the validation design, feature comparison, interpretation and final scientific claims.

Questions to address

The questions form one connected pipeline. Decisions made in data auditing and validation must be fixed before interpreting model performance or biological features.



- Q1. What exactly is contained in the teaching dataset?** Verify patient and section counts, class frequencies, missing values, duplicated spots, coordinate ranges and available technical covariates. Produce a patient-by-class table and determine whether every patient contributes both classes. State the precise mapping used to create the binary core/periphery label.
- Q2. How are the labels and molecular signals organised in tissue space?** For at least two representative sections, plot the supplied labels, one biologically interpretable gene, one metabolic module and simple spot-level quality measures. Identify whether apparent class differences are global, patient-specific, boundary-localised or associated with a technical variable.
- Q3. What validation design answers the scientific question?** Construct grouped folds at the patient level. Draw a diagram showing which patients enter training, hyperparameter selection and final testing. Implement all preprocessing within the corresponding training fold. As a pedagogical comparison, assess whether and by how much a naive random spot split overestimates performance.
- Q4. Which parsimonious models should be compared?** Fit the baseline hierarchy and tune the elastic-net parameters using grouped inner validation. Explain how genes were filtered or selected, how class imbalance was handled and why raw section coordinates were not used as ordinary predictors. Record runtime and the number of non-zero coefficients.

Checkpoint at the end of Wednesday

Before examining feature importance, every group should have: a documented binary label, an auditable patient-level split, one majority baseline, three regularised logistic-regression models and a table of held-out-patient predictions. High apparent accuracy under a random spot split is not an acceptable endpoint.

Questions to address – continued

- Q5. Does the classifier generalise across patients?** Report a confusion matrix, balanced accuracy, sensitivity, specificity, macro-F1, area under the ROC curve and, when class imbalance is relevant, area under the precision-recall curve. Assess probability calibration with the Brier score or a calibration plot. Give results for each held-out patient where the metric is defined, together with pooled and macro-averaged summaries; one large section must not dominate the conclusion.
- Q6. Which genes and metabolic modules contribute reproducibly?** For the linear models, examine coefficient sign and magnitude after standardisation. Complement this with held-out permutation importance. Report how often each feature is selected across patient-grouped folds and whether its direction is consistent across patients. Distinguish a stable predictive association from a causal mechanism or therapeutic target.
- Q7. Where, and why, does the model fail?** Map predicted core probability, uncertainty and misclassified spots back onto the tissue. Test whether errors cluster near core-periphery boundaries or in particular sections. Determine whether important genes mainly reflect malignant-cell programmes, normal neural-lineage composition or immune and vascular components. Repeat the analysis after excluding genes used directly in the original annotation rules and interpret any performance loss.
- Q8. What can be concluded from the comparison of data representations?** Decide whether the metabolic-module model preserves, improves or reduces patient-level performance relative to gene expression, and whether the combined model adds information after accounting for regularisation. Formulate one biological interpretation supported by the data, one limitation that prevents a stronger claim and one additional dataset or experiment that would provide a more independent validation.

Suggested three-afternoon progression

Tuesday: biological introduction, data dictionary, cohort audit, spatial visualisation and baseline model. **Wednesday:** patient-grouped pipeline, feature filtering, elastic-net tuning and comparison of gene-only, module-only and combined models. **Thursday:** held-out evaluation, feature stability, marker-exclusion analysis, spatial error maps and presentation preparation. **Friday:** 6–8 minute group presentation and discussion.

Interpretation boundary

Good discrimination does not prove that the selected features cause invasion or recurrence. The task evaluates prediction of supplied spatial annotations. Its strongest result is a reproducible, patient-generalizable molecular distinction whose dependence on annotation markers, tissue composition and technical variation has been examined transparently.

Expected outputs, group organisation and optional extensions

Required outputs. Each group should produce: (i) a cohort and class-distribution table; (ii) spatial maps of labels and representative molecular features; (iii) a diagram of the patient-grouped validation design; (iv) a table comparing the baseline, module-only, gene-only and combined models; (v) a stable-feature summary across folds; (vi) a spatial prediction or error map; and (vii) a 6–8 minute Friday presentation in which all group members contribute. The presentation must explicitly cover the question addressed, the main findings from exploration of the gene and metabolic datasets, the ML model used to classify spots as core or periphery, and the genes and metabolic modules judged most relevant through an explainability analysis.

Suggested division of labour. One trainee can lead data auditing and spatial visualisation, one can lead model implementation and grouped validation, and one can lead feature interpretation and presentation design. Analytical decisions and final conclusions should be agreed jointly.

Optional advanced directions. Compare the linear model with a tree-based classifier; formulate a three- or four-class problem using cellular tumour, infiltrating tumour and leading-edge regions; use spatially blocked validation within an unseen section; add neighbourhood-averaged features or graph regularisation; aggregate spots into patient- or section-level pseudobulk profiles; or test the workflow on an independent spatial GBM cohort.

Participant package. The recommended package contains this handout, a sparse gene-expression object, a metabolic-module matrix, a spot-metadata table, section images and a data dictionary defining the binary label and annotation-marker genes. These files will already be organised and aligned, so students can begin analysis without preliminary data wrangling. The easy-to-follow, step-by-step Python notebook should demonstrate data loading, exploration and the core regularised classifier, while leaving validation, model comparison and interpretation decisions to the group.

Key bibliography

- [1] Weller M, Wen PY, Chang SM, et al. *Glioma*. Nature Reviews Disease Primers **10**, 33 (2024).
- [2] Sonpatki P, Park HJ, Xing YL, et al. *A spatially resolved human glioblastoma atlas reveals distinct cellular and molecular patterns of anatomical niches*. Nature Communications **17**, 2951 (2026). <https://www.nature.com/articles/s41467-026-69716-2>.
- [3] Puchalski RB, Shah N, Miller J, et al. *An anatomic transcriptional atlas of human glioblastoma*. Science **360**, 660–663 (2018).
- [4] Alghamdi N, Chang W, Dang P, et al. *A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data*. Genome Research **31**, 1867–1884 (2021).
Paper: <https://genome.cshlp.org/content/31/10/1867>;
Code: <https://github.com/changwn/scFEA>.
- [5] Zou H, Hastie T. *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B **67**, 301–320 (2005).
- [6] Singhal A, Zhao X, Wall P, et al. *The hallmarks of predictive oncology*. Cancer Discovery **15**, 271–285 (2025).

Scope and reporting

This Mini-Project is a proof of concept based on a restricted research dataset. It does not produce a clinical diagnostic tool, define a surgical margin or establish independent metabolic flux measurements. Reproducibility requires reporting patient splits, preprocessing choices, software versions, random seeds, hyperparameters and the exact feature lists used in every analysis.



Mini-Project: *Early stages of brain tumour development through the lens of individual-based models*

Supervisors: *Juan Jiménez-Sánchez, Miguel Perales-Patón*

Motivation

Primary brain tumours, ranging from the slow-paced low-grade gliomas to the fast-progressing glioblastomas, are best understood as evolving cell populations rather than static lesions. Early tumorigenesis begins when a single somatic cell acquires a growth-advantaging alteration, after which clonal expansion, additional genetic and epigenetic change, and interaction with the surrounding tissue progressively generate malignant heterogeneity [1]. In glioblastoma, multisampling and phylogenetic studies show that many hallmark features of the disease emerge through branching evolution, with early events laying the trunk of the tumour and later events generating regionally distinct subclones [2,3]. This matters clinically because intratumour heterogeneity is tightly linked to treatment failure, recurrence, and poor outcome in both adult and pediatric brain tumours [3-5]. Nonetheless, this process is far from being fully understood: does evolution occur in a few great steps, or does it take place slowly but steadily?

Individual-based and stochastic mathematical models are useful in this setting because the earliest phases of tumorigenesis are dominated by chance, competition, and spatial constraint at the level of single cells and small clones. Experimental tracing in early gliomagenesis shows frequent clonal extinction and strong divergence in clone sizes, indicating that deterministic average-behaviour models alone can miss biologically important dynamics [6]. By explicitly representing cells or clones as discrete entities, individual-based frameworks can encode proliferation, death, migration, mutation, and cell-to-cell or microenvironmental interactions as probabilistic events, making them well suited to capture stochastic drift, selection, and spatially structured competition [7,8]. They also help bridge scales, because newer stochastic frameworks can track subclonal genealogies and heterogeneity while still reaching clinically meaningful tumour sizes or integrating imaging data [7,9-10].

The main scientific value of these models is that they make otherwise unobservable evolutionary processes inferable and testable. Mathematical modelling can reconstruct time-dependent tumour dynamics from temporally sparse data, distinguish whether observed heterogeneity is more consistent with neutral drift, selective sweeps, branching growth, or mixed evolutionary modes, and identify how ecological constraints shape transitions between these regimes [8-9,11]. In brain tumours, such models have already generated actionable biological hypotheses, for example that combining imaging with cell-scale data improves inference of hidden tumour composition and treatment response, and that macro-scale growth patterns can retain signatures of the underlying evolutionary process [10,12]. This makes evolutionary modelling valuable not just as a descriptive tool, but as a principled framework for linking early tumour initiation, clonal diversification, and clinically relevant heterogeneity in primary brain cancer.

In this mini-project, our aim is to simulate these early stages of brain tumourigenesis by developing and *ad-hoc* individual-based model, capturing the most relevant features of tumour cell's biology, and aiming to reproduce well-known evolutionary behaviours.

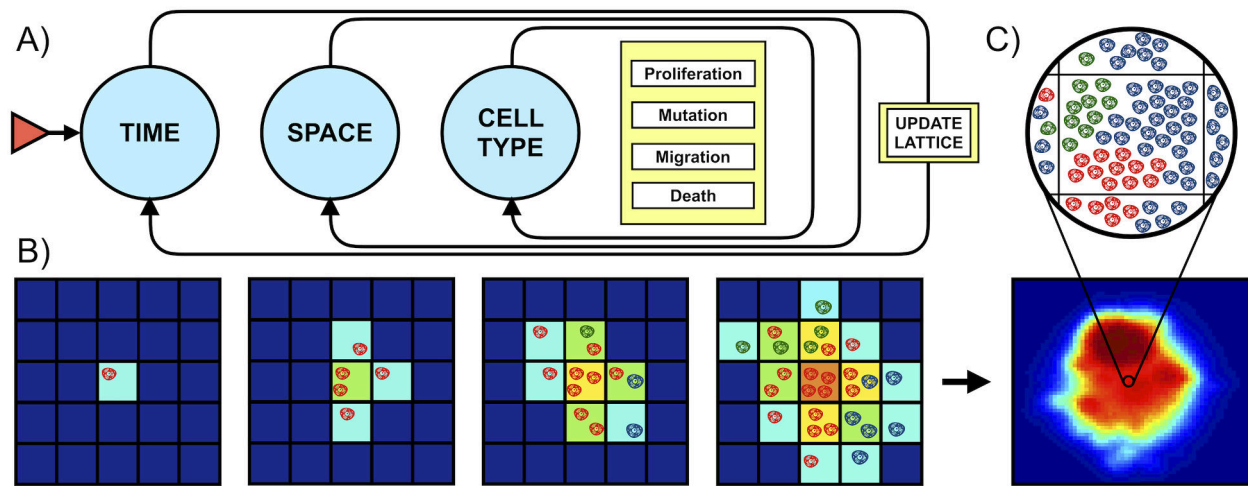


Figure 1: A schematic representation of an individual-based model, accounting for the typical workflow (including cell processes) in **A**), depicting different time steps that show how the tumour expands in **B**), and illustrating a multi-compartment example, where zoom is made over a lattice point in **C**). Figure adapted from [7].

Specific biological context of the problem

- ▶ **As simple as possible, but not simpler.** How many cell types should we consider? How many processes can cells undergo? When modelling biological phenomena, a crucial trade-off is struck between the level of complexity implemented and the ease of simulation. Adding more features to a model makes it more realistic, but may hinder our capacity to identify relationships between inputs and outputs.
- ▶ **What a cell does for a living.** A tumour cell is a boiling biochemical pool, where many different reactions are intertwined and lead to division, migration, etc. All these behaviours can be implemented, but where do we set the bar? Including all possible cell processes in our model goes against the prior statement, and may increase the computational time required for simulation.
- ▶ **A few big jumps, or many small ones?** Different evolutionary theories support different evolutionary paths. While gradualism favours a continued evolution that takes place in small steps, punctuated equilibrium endorses a more stable journey, whereby evolutionary events produce drastic changes. In an individual-based set-up, both hypothesis can be tested.
- ▶ **Trait inheritance.** Everytime a cell divides, relevant information ensuring proper cell function has to be split between daughters, and this information's fidelity must be kept during division. However, daughter cells may behave differently from one another, due to stochastic effects that influence phenotypic traits (uneven mitochondrial distribution, etc).
- ▶ **The winner takes it all.** Tumour cells compete against each other all the time: resources, space availability... This competition leads to selecting fitter clones, but fitness may be an elusive concept in evolutionary biology. Either explicitly (cells competing against each other to occupy a given lattice point in our model) or implicitly (faster-dividing clones gain a selective advantage when competing for space), competition should be implemented in our set-up.
- ▶ **Tracking evolution.** How do we *measure* evolution? In our individual-based model, there are different ways of doing so. From simply evaluating changes over time in relevant observables (division pace, etc), to quantifying diversity indexes (Shannon's entropy, Simpson's index, etc), evolution can be tracked and compared among simulations.

Mathematical modelling approach: quantities to represent

Our task is to develop an individual-based model that describes an expanding tumour cell population, representing early stages of tumourigenesis. Cells should be characterized by a phenotypic trait related to fitness, which is itself subject to change; in this way, evolution can be observed in this setup. By addressing all (or some) of the questions stated in the previous section, we should be able to simulate a growing body of cells, where relevant observables (cell number, average phenotype, etc) should be tracked.

An affordable initial approach would consist on considering an spatial 2D lattice, where tumour cells can occupy lattice points. There, cells, characterized by a phenotypic trait which determines their division rate, would only be able to divide and/or change their phenotypic trait. Despite looking simple, this set-up already arises relevant questions in the context of early brain tumourigenesis: what parameter ranges/units should we consider? Are different processes' scales similar or different? Do fast-dividing cells always *win*?

From this, we can start running batches of simulations and testing evolutionary hypotheses. More on these hypotheses is detailed below, in the **Questions to address** section. Regarding simulations, it is not just clicking the "run" button; time and space need to be discretized, a neighbourhood has to be defined, boundary conditions have to be set, and an initial configuration (as well as a stopping rule) must be defined. Updating our system can also be tricky, since synchronous vs asynchronous approaches may lead to different outcomes. All these nuances will pop up during the development of this mini-project.

Biological conditions that the model should encode

The following requirements define the modelling scenario. They do not prescribe a unique set of equations.

Process	Conditions to represent mathematically
Cell division	In the context of individual-based models, dichotomic processes (success or failure) can be simulated via Montecarlo steps, which rely on changing probabilities. These probabilities should be realistic, and capture the basic features related to the process.
Phenotypic changes	A phenotypic trait subject to changes must be associated with some cell process, in order to observe selection over time. Changes can be implemented in different ways: either seldom but big, or frequent but small. Different modeling assumptions lead to different outcomes.
Trait inheritance	From a modeling perspective, daughter cells can either be a copy of their parent, or slightly deviate from it (in phenotypic terms). Coupling division with phenotypic changes alters the evolutionary pathway in our simulations.
Daughter cell placement	A practical way of explicitly implementing cell-to-cell competition is by searching for available space to place newborn cells. Either contact inhibition or cell removal can be implemented, shaping different spatial phenotypic maps of our tumour cell population.
More complex processes	Cell division is enough to observe evolution, but more complex processes could also be implemented, adding subtler flavour to our model. Cell migration let us model invasion processes, at the cost of searching for available space in a cell's vicinity (within the lattice); cell death let us recover necrotic cores observed in aggressive tumours (such as glioblastoma), but at the cost of having to deal with dead cells in our model.
Lattice and dimension	Competition and, more generally, spatial interactions, depend on the available areas from which a cell can be attacked. Although a 2D on-lattice model is the easiest to implement, jumping to 3D and an off-lattice set-up is much more realistic (though at the expense of computational cost).
Tracking evolution	Calculating the average phenotype at each iteration is enough to measure the speed of evolution. However, more insightful information can be gathered from our simulations, such as diversity (is our population concentrated on a given phenotype, or can many different clones coexists?) or spatial distribution of changes (where is the fittest clone/cell placed?).

Modelling choices deliberately left open

All the previous conditions are to be implemented in the individual-based model, but the level of complexity to be considered is left to the participant. The choice of programming language is also up to the participant, although MATLAB/Python/Julia are preferred.

Questions to address

The central question of this mini-project is:

How do the frequency and magnitude of heritable phenotypic changes, together with local competition for space, shape tumour growth, diversity and the spatial distribution of proliferative traits?

To answer it, a model has to be built. Developing an individual-based model *from scratch* is an iterative process. Down below we list different concerns, in such a way that each question is related to a modeling assumption or decision:

- Q1. What is the minimal set of code required to simulate a tumour cell population?** Consider as few processes as possible, in order to build the minimal set of operations required to recover a growing tumour.
- Q2. How does the average phenotype change over time?** After developing the model, devise a way to track the averaged phenotypic state (among all cells) to observe whether it changes over time. Does the choice of parameters influence the speed of evolutionary change?
- Q3. Tracking evolution in a deeper way.** Devise a way to calculate diversity indexes (Shannon entropy, Simpson index) and track their evolution over time. Devise also a way to calculate the deviation of the fittest cell from the tumour center, to assess whether evolution takes place at the tumour center or at the tumour boundary.
- Q4. How do different evolutionary hypothesis shape the spatio-temporal distribution of cells?** Test two different scenarios: a gradualist one, where phenotypic changes are small but frequent, against a saltationist one, where changes are rare but drastic. Discuss which one could be more realistic in the context of brain tumours.
- Q5. Does cell competition influence the shape of evolution?** Compare the effect of letting cells go quiescent in the absence of available space (no division can occur if there is no room for it) against the effect of implementing explicit competition (newborn cells can remove pre-existing cells from their vicinity to make room for themselves).
- Q6. Which conclusions are robust to uncertainty in model structure and parameters?** Perturb influential parameters, and determine whether the main features of evolution are preserved.

Suggested three-afternoon progression

Tuesday: model development from scratch. **Wednesday:** testing of different scenarios through simulation. **Thursday:** analysis of results, preparation of presentation. **Friday:** final group presentation checks.

Key bibliography

- [1] Zhang, Shaosen, et al. *Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets*. Signal Transduction and Targeted Therapy **9**, (2024).
- [2] Spiteri, I., et al. *Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics*. Proceedings of the National Academy of Sciences, **110**, 4009-4014 (2013).
- [3] Mathur, Radhika, et al. *Glioblastoma evolution and heterogeneity from a 3D whole-tumor perspective*. Cell, **187**, 446-463 (2024).
- [4] Nguyen, M., and Pablo G. Cámara. *Abstract B019: Mapping clonal architecture and evolution in pediatric brain cancers*. Cancer Research, (2026).
- [5] Ruffle, James K., et al. *Brain tumour genetic network signatures of survival*. Brain, **146**, 4736-4754 (2023).
- [6] Ceresa, Davide, et al. *Early clonal extinction in glioblastoma progression revealed by genetic barcoding*. Cancer Cell, (2023).
- [7] Jiménez-Sánchez, Juan, et al. *A mesoscopic simulator to uncover heterogeneity and evolutionary dynamics in tumors*. PLoS Computational Biology, **17**, (2020).
- [8] Volpatto, D., et al. *A new cancer progression model: From synthetic tumors to real data and back*. PLoS Computational Biology, **22**, e1013991 (2026).
- [9] Streck, Adam, et al. *SMITH: spatially constrained stochastic model for simulation of intra-tumour heterogeneity*. Bioinformatics, **39**, (2022).
- [10] Gallaher, J., et al. *From cells to tissue: How cell scale heterogeneity impacts glioblastoma growth and treatment response*. PLoS Computational Biology, **16**, (2019).
- [11] Colson, Chloé, et al. *Mathematical modelling of cancer cell evolution and plasticity*. Current Opinion in Cell Biology, **95**, 102558 (2025).
- [12] Ocaña-Tienda, B., et al. *Growth exponents reflect evolutionary processes and treatment response in brain metastases*. NPJ Systems Biology and Applications, **9**, (2023).



Mini-Project: *Designing a Model to Distinguish Mechanisms of Therapy Resistance Acquisition through Phenotypic Measurements Across Time and Radiation Doses*

Supervisors: *José Antonio Romero Rosales & David G. Aragonés*

Motivation

Glioblastoma (GBM) is the most common primary brain tumor in adults and remains an incurable disease. The causes of this incurability are numerous and highly interconnected. Despite aggressive multimodal treatment, median survival remains approximately 18 months, with tumor recurrence typically occurring within 6 to 9 months. One of the principal factors underlying this poor prognosis is the remarkable plasticity and heterogeneity of GBM tumors [1,2].

Several forms of heterogeneity have been described in glioblastoma; however, the most relevant to this project is intratumoral heterogeneity, which refers to the coexistence of distinct functional cell populations within a single tumor. Transitions between these cellular states are thought to contribute to therapeutic resistance, tumor regrowth, and disease progression.

Vessel co-option is particularly noteworthy as a process that encompasses several characteristics associated with glioblastoma recurrence. By exploiting pre-existing blood vessels, tumor cells can infiltrate surrounding brain tissue more effectively, creating regions in which radiotherapy is less effective. In addition, vessel co-option provides access to microenvironments rich in protective factors that promote tumor survival, growth, therapeutic resistance, and dissemination.

In 2024, a novel glioblastoma functional state was described [3]. This state integrated several features commonly associated with recurrence and was termed VC-Resist (Vessel-Co-opting Resistant state). VC-Resist cells occupy an intermediate position along the proneural-mesenchymal axis and display a partially reversible phenotype.

Although this state was extensively characterized experimentally, the mechanisms responsible for its enrichment following radiation therapy remained unclear. In particular, it was not possible to determine whether its post-treatment expansion resulted primarily from Darwinian selection of pre-existing resistant cells or from therapy-induced cellular reprogramming. Mathematical modeling can provide valuable insights into this question. By constructing computational representations of the relevant biological experiments, it becomes possible to systematically vary the contributions of the mechanisms involved and evaluate which combinations of parameters most accurately reproduce the observed experimental outcomes. Such an approach may help quantify the relative importance of selection and reprogramming in driving the emergence of the VC-Resist state following treatment.

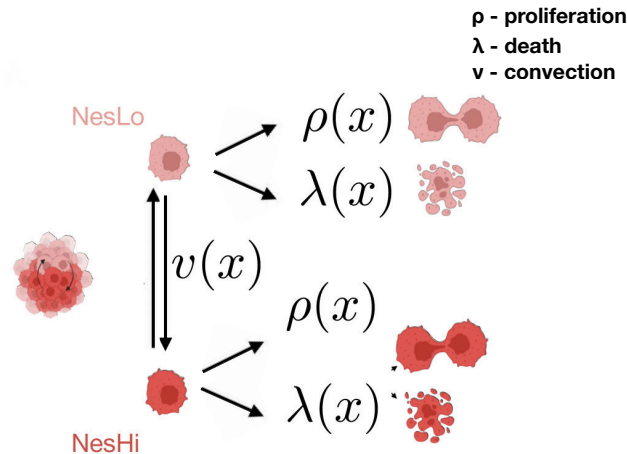


Figure 1: Diagram showing the general flow of cells from vulnerable to resistant cells and the different mechanisms relevant in this evolution.

Specific biological context of the problem

- **The phenotypic system was shown to be rich in a particular filament called Nestin.** This filament was found to be a good marker of the aforementioned VC-Resist state and is already known to be a senescence marker in other tumors. Meaning it could be divided into a NesHI/NesLO states. Several experiments were performed using MGG4 GBM cells, which were subjected to radiation and observed during seven days. The aforementioned filament was independently measured through flow-cytometry meaning our studied cells emitted a particular radiation frequency whenever Nestin was present.
- **Two experiments that are particularly relevant for its modeling are those related to cell proliferation and cell death.** Through area-under-the-curve analysis of Sytox-positive (Sytox+) cells, it was determined that cells with high Nestin expression died less frequently than cells with low Nestin expression when exposed to similar radiation doses. Sytox is a dye that cannot penetrate intact cell membranes and binds to nuclear DNA only when membrane integrity is compromised, making it a reliable indicator of cell death. A similar approach was used to assess cellular proliferation. Cells were labeled with the CellTrace marker, which becomes progressively diluted with each cell division. This enabled the quantification of the proliferative capacities of NesHI and NesLO cells.
- **Reprogramming was already established as relevant experimentally** In order to tests this researchers performed individual cell analysis in which they could observe how after radiation individual surviving MGG4 NesLO cells showing Nestin reporter activation. Furthermore the NesHI subpopulation was also shown to increase after interaction with endothelial or blood vessel-conditioned media, given that this experiment would not induced the death of NesLO cells, this serves as evidence that reprogramming may be an important factor in the appearance of this state.
- **Therapeutic relevance of cell plasticity** Glioblastoma cell plasticity is a current avenue of research currently being explored as a target of several therapeutic strategies such as targeted therapeutics and immunotherapy [4]. Representation part of the cancer's evolutionary advantages and combining intrinsic epigenetic and transcriptional reprogramming and extrinsic TME signals, its continuous understanding being key in the refinement of treatments against this tumor.
- **Other cell details.** These cells are characterized by senescent-like and stem-like features, a slow-cycling phenotype, enrichment in the G2/M phase, resistance to DNA-damaging therapies, and a strong vessel-co-opting capacity. They also exhibit upregulation of FGFR1, activation of YAP1 signaling, and enhanced DNA damage repair (DDR) pathways. This cellular state lies along the proneural-to-mesenchymal transition axis, closely associated with AC/MES-like transcriptional programs, and is mechanistically linked to FGFR1-YAP1 activation. Importantly, endothelial- or blood vessel-conditioned media can induce the transition from NesLO to NesHI cells, whereas the FGF-trap compound NSC12 blocks both conditioned media- and irradiation-induced reprogramming, although it does not revert the already established NesHI state.

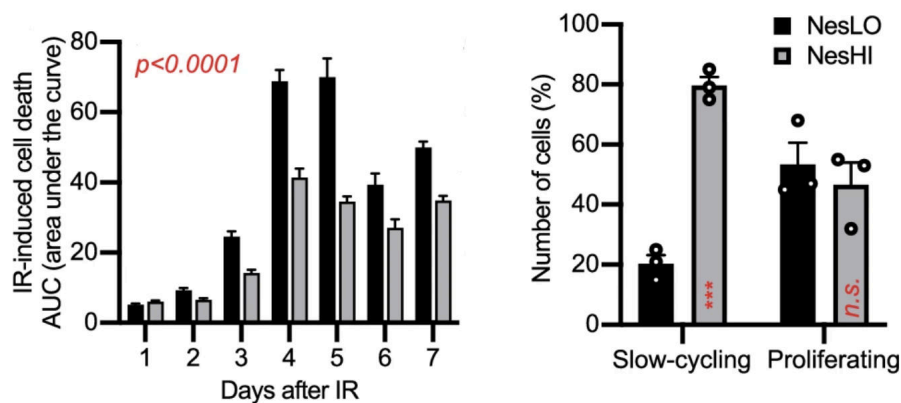


Figure 2: Experiment results relevant to cell proliferation and death.[3]

Mathematical modelling approach: quantities to represent

Your main modeling task will consist of formulating a partial differential equation (PDE) model capable of reproducing the time evolution of the probability distributions of the chosen variable. This equation should consist of three independent components corresponding to proliferation, death, and reprogramming:

$$\frac{\partial u}{\partial t} = \rho(x, t)u(x, t) + \lambda(x, t)u(x, t) + v(x, t)\frac{\partial u}{\partial x}.$$

Here, x represents the base-10 logarithm of the selected Nestin expression value, while t denotes time. The functions ρ , λ , and v represent the proliferation, death, and convection terms required to reproduce the observed evolution of the experimental data.

The precise mathematical form of these biological processes, as well as their dependence on Nestin expression and time, will be left to the students' analysis and interpretation. Likewise, any constraints imposed on the model parameters are to be defined and justified by the students. They are also free to decide which parameters should be estimated from the data and which may be fixed using values available in the literature, provided that a reasonable justification is given.

The choice of fitting and optimization methodology will also be delegated to the students. The primary objective is to obtain a final system of partial differential equations whose solutions are capable of reproducing the trends observed in the experimental distributions over time.[5,6]

Dataset file for the group

The file that should be used is found in: File for Net4brain GBM-Nestin practice. In this file you will find 42 smoothed curves that correspond to the probability profiles of each of our radiation doses [0Gy, 2Gy, 5Gy, 8Gy, 10Gy, 12Gy] and for the 7 days following radiation. These profiles are directly taken from the histograms representing the $\log_{10}$ Nestin-marker distribution of each group of cells and were smoothed through the `ksdensity` Matlab (2025b) function.

Biological conditions that the model should encode

The following requirements define the modelling scenario. They do not prescribe a unique set of equations.

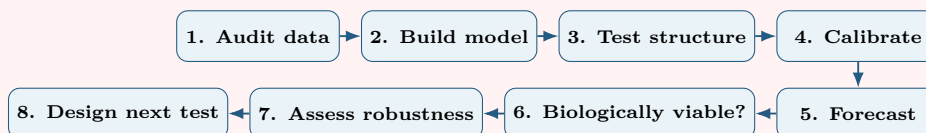
General conditions	The model solutions should be smooth, positive, and normalizable, as they represent probability distributions.
Proliferation	The model should incorporate a proliferation component. The proliferation rate must remain within biologically realistic bounds and, as suggested by the Nestin-proliferation profile, vary as a function of Nestin expression.
Death	The model should incorporate a death component. The death rate must remain within biologically realistic bounds and, as indicated by the Nestin-proliferation profile, depend on Nestin expression. Furthermore, it should increase with radiation dose, in agreement with the experimental observations.
Combined dynamics	The interplay between proliferation and death should result in a net reduction in tumor volume at sufficiently high radiation doses.
Convection	The model should include a reprogramming-convection term. This term may be either constant or dependent on Nestin expression and should capture biologically plausible reprogramming dynamics across different radiation doses.
Extrapolation	The model should produce biologically realistic predictions beyond the final experimentally recorded time point.
Identifiability	The model should remain sufficiently simple to admit an analytical solution while allowing efficient parameter estimation through error minimization against the available experimental data.
Solvability	While not a necessary condition it is highly recommended for the sake of efficiency and parametrization that proposed system has an analytical solution or is very efficiently solvable numerically.

Modelling choices deliberately left open

The relationship between the three mechanisms and Nestin expression is deliberately left open for the students to define. Likewise, any potential time dependence in these mechanisms may be proposed and incorporated by the students, provided it is appropriately justified.

Questions to address

The questions form a connected workflow. The outcome of each stage should constrain the next stage rather than being treated as an independent exercise.



- Q1. What information is contained in the dataset, and which qualitative patterns must the model explain?** Verify the number of data points, identify the main features that the model must reproduce, and determine how the model can be solved and fitted efficiently.
- Q2. What is the simplest solvable model that can capture the known characteristics?** Examine the experimental curves and estimate how they may be influenced by the different mechanisms under study. Reflect on the biological significance of each of the proposed constraints and how they may be incorporated into the model.
- Q3. Does the proposed system reproduce the expected behavior?** Fit the model to the data and obtain the simplest set of results. Assess the degree of uncertainty captured by the model and determine whether further improvements are necessary. Consider which parameters should be estimated from the data and which should be obtained from the literature.
- Q4. Is your solution realistic and biologically consistent?** Extrapolate the model beyond the experimental time window and reflect on any limitations, unrealistic predictions, or unexpected behaviors that emerge at later times.
- Q5. How does the model answer the original question?** Using the proposed framework, identify which factors appear to be the most relevant. Consider whether the conclusions may be influenced by simplifying assumptions or arbitrary decisions made during the modeling process.
- Q6. Do you believe this is the most useful model?** Can you identify any extensions that could make the model more informative, or any simplifications that could make it more elegant? Consider whether sufficient information is available for such modifications or whether additional evidence from the literature would be required.
- Q7. Which aspects do you believe limit your model?** Could the model be extended to other systems or cancer types? If so, what additional information or data would be required?
- Q8. What additional information do you believe would improve the model?** Describe the types of data and experiments that could enhance the specificity, predictive power, and biological validity of the model.

Suggested three-afternoon progression

Tuesday: Visualization of data and adjustment of control parameters. **Wednesday:** Extension of model to radiation values and extrapolation. **Thursday:** Examination of model and estimation of final biological question. **Friday:** Final group presentation checks.

Dataset, expected outputs and optional extensions

The dataset does not contain direct measurements of Nestin or complete information on all sampled cells. Instead, it provides sampled points from smooth curves corresponding to different radiation doses and time points.

Required outputs. A complete group submission should contain: (i) a brief explanation of the dataset and its relationship to the proposed model, including any parameters obtained from the literature and their justification; (ii) the complete formulation of the PDE system, together with an analysis of how each term influences the simulation; (iii) the analytical solution or numerical scheme used to simulate the model; (iv) graphs comparing the simulation results with the provided data and figures; (v) a table containing the fitted parameters and/or forecast errors; (vi) a final assessment of how to quantify the relative importance of each mechanism in the system; and (vii) a 6-8 minute presentation on Friday in which all group members contribute.

Optional advanced directions. Attempt to extend the model by incorporating partial reversibility and making the complete system time dependent. Find a sufficiently simple formulation that either preserves the possibility of an analytical solution or allows the construction of a numerical scheme capable of reproducing the observed behavior to a reasonable extent.

Key bibliography

- [1] Nahm, J., Sinha, M., Schumann, E. H., Vu, M., Hsu, S. H., Zhu, J.-J. . *Overall survival in patients with recurrent glioblastomas with combination chemotherapy and tumor treating fields (TTF)*. Journal of Clinical Oncology, **41** (16-suppl), e14057–e14057 (2023).
- [2] Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., Bernstein, B. E. . *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma*. Science, **344** (6190), 1396–1401 (2014).
- [3] Pichol-Thievent, C., Anezo, O., Pettiwala, A.M. et al. *VC-resist glioblastoma cell state: vessel co-option as a key driver of chemoradiation resistance*. Nat Commun **15**, 3602 (2024)..
- [4] Xiao T., Wei Z., Wu M., *Glioblastoma cell plasticity: A new paradigm in glioblastoma therapeutic resistance*, Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, **1881**, Issue 2, 189557 (2026).
- [5] Álvarez-Arenas A., et al, *Interplay of Darwinian Selection, Lamarckian Induction and Microvesicle Transfer on Drug Resistance in Cancer*, Scientific Reports, **9**, Issue 1, 189557 p. 9332 (2019).
- [6] Guilberteau, J., Jain, P., Jolly, M.K. et al. *IA n integrative phenotype-structured partial differential equation model for the population dynamics of epithelial-mesenchymal transition*. npj Syst Biol Appl **11**, 24 (2025).



Mini-Project: *Exploration of the FGFR3-TACC3 fusion gene in glioblastoma*

Supervisor: *Milana Frenkel-Morgenstern*

Motivation

Glioblastoma (GBM) is an aggressive primary brain tumour. Its malignant cells infiltrate the surrounding brain, and its molecular alterations differ between patients. Understanding a specific tumour-driving alteration can therefore be more informative than treating all GBMs as one biological system. This Mini-Project focuses on FGFR3-TACC3, a recurrent fusion found in a subset of glioblastomas and other IDH-wild-type gliomas [1–3]. Here, IDH-wild-type means that the tumour does not carry the characteristic mutations in the isocitrate dehydrogenase genes used to distinguish major groups of adult diffuse glioma. The project concerns primary brain cancer, not a fusion inferred from another cancer type.

A gene fusion joins parts of two genes. When the resulting RNA has a suitable coding sequence, it can produce a single protein containing regions from both parents. In FGFR3-TACC3, a portion of the fibroblast growth factor receptor 3 is joined to a portion of transforming acidic coiled-coil-containing protein 3. The altered combination can support persistent kinase activity. Experimental studies have shown that the fusion can drive glioma formation and perturb chromosome segregation [1]. This provides a concrete link between a sequence change and tumour biology.

A database entry, however, is not by itself a complete biological explanation. The record may contain only a short fragment around the fusion junction. The reading frame must be checked. The encoded protein may retain some parental domains and lose others. A predicted three-dimensional structure then adds another layer of interpretation, but also introduces uncertainty. The students will connect these stages instead of treating sequence retrieval, translation and structural visualisation as unrelated exercises.

The aim is to assemble a reproducible, evidence-based account of how the fusion changes protein architecture and why that matters in glioblastoma. The result should identify testable molecular hypotheses. It is not expected to establish a clinical biomarker or recommend a treatment.

Central scientific question: Which sequence and structural features of FGFR3-TACC3 support its role as an oncogenic fusion in glioblastoma, and which conclusions still require experimental evidence?

Learning objectives and scope

By the end of the project, the group should be able to validate a fusion junction, justify a protein-coding reading frame, compare retained domains, and interpret a structure together with its confidence information. Each conclusion should be linked to a sequence record, an analysis output or a cited experiment.

The five tasks retain the two parts of the original draft: **fusion-transcript analysis**, followed by **structural bioinformatics**. The group retrieves and compares at least six sequence records, subject to the provenance checks in Task 1. Detailed structural work concentrates on one well-supported complete fusion protein. Coding and additional structural comparisons are optional extensions, not prerequisites for completing the core analysis.



Working principle. A later analysis is only as reliable as the sequence and annotation decisions on which it depends. Resolve uncertain sequence identity before interpreting a structure.

Specific biological context of the problem

- ▶ **The normal FGFR3 protein.** FGFR3 is a receptor tyrosine kinase. Its extracellular region can bind fibroblast growth factors, its transmembrane segment crosses a membrane, and its intracellular kinase domain transfers phosphate groups from ATP to protein tyrosines. These phosphorylation events transmit signals inside the cell. Distinguishing these regions helps students ask which normal receptor functions could remain in a particular fusion. Parental annotations should be checked against the exact reference sequence used, not transferred solely by gene name [1-3].
- ▶ **The normal TACC3 protein.** TACC3 participates in the organisation of the mitotic spindle, the microtubule apparatus that separates chromosomes during cell division. Its C-terminal TACC region contains coiled-coil structure. A coiled coil consists of interacting helices and can contribute to protein association. It is not simply a synonym for any long alpha helix. Association can bring kinase regions together and favour phosphorylation between them. Experiments on FGFR3-TACC3 show that the retained TACC3 coiled-coil region is important for oncogenic activation [1,5].
- ▶ **From a DNA rearrangement to a fusion transcript.** FGFR3 and TACC3 lie close together on chromosome 4. Tandem duplication, in which a DNA segment is copied adjacent to itself, has been demonstrated as a mechanism generating the fusion. RNA splicing then determines the mature transcript. The *genomic breakpoint* is the DNA rearrangement site. The *transcript junction* is the meeting point of the parental sequences in mature RNA. They are related, but are not interchangeable coordinates [2]. This project can map a transcript junction from sequence alignment; a precise DNA breakpoint requires independent genomic evidence.
- ▶ **Why the coding frame matters.** A ribosome reads the coding sequence in successive triplets. An in-frame junction preserves a continuous protein-coding sequence across the parental boundary. A frame shift changes the downstream codons and may introduce a premature stop. Established oncogenic variants retain an FGFR3 kinase region joined to a TACC3-derived C-terminal region, but junctions and retained lengths vary [1,3]. Exon numbers also depend on the parental transcript reference. Six database records must not automatically be described as six distinct fusion variants.
- ▶ **How the fusion can promote tumour behaviour.** Combining kinase activity with TACC3-derived association and localisation features can alter protein regulation. Persistent kinase activity, spindle abnormalities and chromosome-number changes have experimental support [1,5]. Other studies identified loss of an FGFR3 RNA regulatory region that normally responds to miR-99a, a small regulatory RNA [2]. A further study connected the fusion to increased mitochondrial energy production. It identified phosphorylation of PIN4, a protein substrate of the fusion, as an intermediate step [4]. These are findings from particular experiments. They are not properties that can all be established from a static structure.
- ▶ **What structural analysis can contribute.** Domain annotation identifies the molecular functions plausibly retained by a sequence. Structure prediction asks how that sequence might fold. A locally well-predicted kinase domain does not establish its activity, the orientation of distant domains, or the biological oligomeric state. An uncertain junction region cannot be treated as a demonstrated new binding pocket [6-7]. Similarity to a parental domain supports domain-level interpretation, not the claim that the complete fusion is an ordinary parental protein.
- ▶ **Why this matters for brain cancer.** Molecular characterisation has motivated FGFR-inhibition studies in fusion-positive glioma [3]. This gives the project a therapeutic rationale, but the trainee analysis does not establish drug sensitivity or intracranial efficacy. Such claims require functional testing and evidence concerning drug exposure in the brain. The final discussion should distinguish a plausible target, an experimentally supported mechanism and a clinically useful intervention.

Evidence to keep separate

Observed or documented: the retrieved sequence and its source annotation. **Computed:** an alignment, translation, domain assignment or predicted structure. **Experimentally supported:** a mechanism demonstrated in a cited study. **Hypothesised:** a functional consequence proposed by the group. An attractive molecular image does not move a hypothesis into the experimental category.

Computational approach and starting materials

Use the course tools, with the links on page 6 as starting points. FASTA is a text format with an identifier line followed by a sequence. Work from public, traceable records rather than creating hypothetical fusion sequences. Keep the original downloads unchanged. Store accession versions, retrieval dates, organism, tumour source, sequence type and completeness in one metadata table. Use a consistent coordinate convention: for example, 1-based, inclusive nucleotide and residue intervals, with the transcript junction recorded as lying between two specified nucleotides.

Preparation to agree with the supervisor. The draft does not supply a sequence-accession list, FASTA files or an AlphaFold model. Before the sessions, the supervisor should confirm a retrievable six-record panel and access to a prediction workflow. A precomputed model for the exact approved protein, with its input and confidence files, is the backup if computation is delayed. These resources are preparation requirements, not data claimed to accompany this handout.

Questions to address: Part I – Fusion Transcript Retrieval and Analysis

1. Retrieval of chimeric RNA sequences and validation of the junction

Question. Which records genuinely support an FGFR3–TACC3 transcript, and what does each record cover?

Retrieve **at least six** chimeric RNA or cDNA sequence records in FASTA format from NCBI or another public source. Prioritise glioblastoma or other glioma-derived material. A record from another tumour type may be used only as a clearly labelled comparator with supervisor agreement. If fewer than six suitable records are retrievable, document the search and agree a reduced panel with the supervisor. Do not invent sequences to meet the numerical target.

- Record the accession and version, source link, sequence length and sample context. Classify the record as a complete transcript, complete coding sequence or partial junction-spanning sequence. Identify engineered constructs separately; they are not additional patient-derived variants.
- Retrieve the appropriate parental transcript references. Align each chimeric sequence to both parents. Report the aligned intervals, sequence identity and coverage on each side. Identity measures agreement within the alignment; coverage describes how much of the query or reference is aligned. State the denominator. Check the order of the FGFR3-derived and TACC3-derived regions.
- Highlight the junction in an alignment and a short annotated sequence window. Give the parental exon labels only when their reference transcripts are stated. If an insertion or shared sequence prevents a unique junction position, report the ambiguity instead of forcing an exact coordinate.
- Compare the six records. Separate duplicated submissions from distinct nucleotide sequences and distinct junction classes. Explain which records can support full-protein reconstruction and which only support the junction.

Expected output. A traceable FASTA collection, a six-record metadata table and a junction map. The table must state how many distinct sequences and junction classes are actually represented.

2. Six-frame translation and selection of a plausible ORF

Question. Which translation is compatible with the transcript orientation and a continuous fusion coding sequence?

An open reading frame (ORF) is a candidate coding interval without an internal in-frame stop. Translate each record in all six frames: three offsets in the supplied sequence and three in its reverse complement. For software accepting DNA notation, convert U to T in a working copy and record that conversion. Use the standard genetic code and preserve stop-codon symbols in the diagnostic output [10].

Explain why the six-frame search is useful when orientation or coding annotation is uncertain. A mature RNA with known 5'-to-3' orientation has only three forward frame offsets; reverse-complement translations are orientation controls, not six equally plausible products of that RNA. The intended full-length fusion generally has one annotated main coding frame, but sequence length alone neither proves its translation nor excludes alternative initiation.

For each record, list candidate ORFs and their coordinates. Test continuity through the junction, agreement with the coding annotation, and retention of the expected parental regions. A long ORF that fails these checks should not be selected automatically. For a partial record, report a partial translation rather than inventing a start codon or missing residues.

Expected output. Six-frame results and a short, evidence-based ORF decision for each record. Explain any unresolved case before proceeding to Task 3.

3. Identification and comparison of FGFR3-TACC3 fusion proteins

Question. What protein is supported by each transcript, and which complete sequence should be used for structure prediction?

Use **EMBOSS Transeq** or equivalent software to verify the selected translations. Retain the frame labels and options: reverse-frame conventions can differ between programs [10]. Export a protein FASTA for each supported coding sequence. Label fragments explicitly and keep the parent sequence accession in every header.

Check the annotated initiation site. An AUG codon in RNA, represented as ATG in cDNA, is a candidate start, but not every internal AUG is an initiation site. Inspect the surrounding sequence and the curated coding annotation. Check the terminal stop, internal stops and any codon spanning the fusion boundary. A junction fragment without the 5' end cannot establish translation initiation. Search NCBI Protein, UniProt and the cited literature for the **longest documented complete fusion protein found within your stated search**. Report its accession or traceable source, length, completeness and search date. Do not call the longest ORF in a partial record the “longest known protein.” If no independently deposited full-protein record is found, distinguish your transcript-derived translation from an independently annotated protein.

Map the FGFR3-derived and TACC3-derived residues onto each complete protein. Annotate the kinase, transmembrane, extracellular and coiled-coil regions when present. Record absent or truncated features instead of assuming that every variant has identical architecture. Select one complete, well-supported sequence for Task 4; the longest sequence is not automatically the best-supported choice.

Expected output. Protein FASTA files, a protein/domain comparison table and a justified choice of one sequence for structural analysis.

Questions to address: Part II – Structural Bioinformatics

4. AlphaFold structure prediction and confidence-aware interpretation

Question. Which parts of the predicted fusion structure support biological interpretation, and which remain uncertain?

Predict the structure of the complete protein selected in Task 3 using an **AlphaFold-based workflow**. AlphaFold2 through ColabFold is one option [6-7]. Record the software or server version, exact input sequence, chain composition and available run settings. Do not submit a concatenation of two complete parental proteins unless that is the verified fusion sequence. A database model of wild-type FGFR3 is not a model of the fusion.

Inspect the following outputs:

- Domain architecture.** Show a labelled cartoon with parental origins, the fusion junction and annotated domains. State whether the model is full length or a defined fragment.
- Local confidence.** Inspect pLDDT, the predicted local distance difference test score, when provided. It estimates confidence in local structure on a 0-100 scale; it does not measure kinase activity. Plot it along the residue sequence and inspect each domain and the junction separately.
- Relative domain placement.** Inspect predicted aligned error (PAE), when available. It estimates positional uncertainty after aligning on another part of the protein. High inter-domain PAE means that convincing individual domains may still have uncertain relative orientations.
- Comparison with known proteins.** Find experimentally determined parental or homologous domains through the Protein Data Bank. Record sequence identity, coverage, chain and residue range. Explain which fusion region is supported; a match to one domain does not validate the entire fusion.

Write **approximately one page** interpreting these findings. Connect the retained architecture to the biology on page 2. Separate sequence-supported conclusions from tentative structural hypotheses. A monomer prediction does not demonstrate a coiled-coil dimer or its biological assembly.

Expected output. The model coordinates and input FASTA, available confidence files, two annotated structural views, a confidence profile and the one-page interpretation. If a live run fails, analyse the supervisor's sequence-matched backup and state that provenance explicitly.

5. Secondary structure and structural classification

Question. How do local structural elements and domain families help explain the fusion protein's molecular capabilities?

Identify alpha helices, beta sheets and loops in the predicted structure. Use a stated assignment method, such as DSSP or the viewer's secondary-structure algorithm. These assignments describe the supplied coordinates; they are not independent experimental measurements. Mark uncertain regions using the confidence output. A loop is not necessarily disordered, and a low-confidence region is not proof of intrinsic disorder.

Report the fraction of assigned residues in each secondary-structure category, both for the whole model and for its principal domains. State the residue ranges and denominator. If low-confidence residues are excluded, report how many were excluded. Consider whether a high overall helical fraction simply reflects the retained TACC3 coiled-coil region rather than a global change in protein function.

Use **SCOP** to investigate structural classifications of matched parental or homologous domains.

CATH is optional, as in the original draft [8-9]. Record the classification resource/version, entry identifier, domain boundaries and the evidence linking that domain to your fusion. Distinguish a database assignment to an experimental structure from an assignment inferred by similarity. Do not assign one SCOP fold to the entire multidomain fusion. If a region has no supported classification, report it as unassigned.

Conclude by answering three connected biological questions: What catalytic capability is plausibly retained? What role could the TACC3-derived region have in association or localisation? Which additional experiment would test a claim that the sequence and structure cannot establish? Relate the answers to glioblastoma rather than to cancer in general.

Expected output. An annotated secondary-structure view, a domain-classification table and a short synthesis. Include one supported conclusion, one plausible hypothesis and one explicit limitation.

Interpreting prediction confidence

For AlphaFold2, pLDDT values above 90 usually indicate very high local confidence, 70–90 indicate confident local structure, 50–70 indicate lower confidence, and values below 50 indicate very low confidence. Use these as interpretation guides, not acceptance thresholds for biological function. Read pLDDT together with PAE, sequence coverage and domain annotation. Preserve uncertain regions in the audit even when hiding them in a figure. The linked EMBL-EBI confidence guide explains these distinctions.

Expected outputs, reproducibility and optional extensions

Core submission. Prepare a short PDF report organised by Tasks 1-5, with the approximately one-page discussion requested in Task 4. Submit the sequence and metadata tables, original and translated FASTA files, alignments, structural coordinates, available confidence files and any scripts or viewer commands used. Include labelled screenshots documenting the principal analyses, as requested in the original draft. Export plots and molecular figures directly where possible; screenshots do not replace reusable numerical outputs.

Friday presentation. Explain the question, the sequence choices, the principal results and their limitations. Use a junction/domain comparison, a structure with confidence information and a final biological evidence summary. Each group member should contribute. A proposed duration is 6-8 minutes, subject to the School's final allocation.

Optional extensions, after the core tasks. Automate translation and ORF checks in Python; compare a second verified fusion variant; align matched kinase domains and report the aligned residues and root-mean-square deviation (RMSD); or examine a carefully specified oligomeric model. These extensions must not displace the sequence-provenance or confidence checks. No molecular-dynamics calculation, drug docking or clinical prediction is required.

Working sequence and supervision

The main sessions run on **Tuesday, Wednesday and Thursday, 15:30-19:00**. The timetable also includes project time on Thursday, 11:30-13:30. Friday, 09:00-10:00, is reserved for final preparation before the presentations.

Session	Core work and checkpoint
Tuesday	Brief biological introduction; Tasks 1-2; begin Task 3. Confirm sequence provenance, junctions and unresolved ORFs before structural work.
Wednesday	Complete Task 3; launch Task 4 early. Inspect a precomputed backup while a new run is pending. Check input identity, domains and confidence outputs.
Thursday	Complete Tasks 4-5; compare evidence, refine figures and prepare the presentation. Attempt an extension only after the core outputs are secure.
Friday	Final checks, group presentation and discussion of limitations.

Supervision. Start with a 15-20-minute introduction and use short targeted explanations when needed. Divide sequence, structural and evidence-checking work among the trainees, but review the central decisions together. A prechecked reference panel and model should support progress without supplying the students' interpretation in advance.

Public resources and practical access

Sequences and alignment	NCBI Nucleotide; NCBI Protein; NCBI BLAST; ENA.
Parental annotation	UniProt; InterPro. Select the exact parental isoform before transferring residue coordinates.
Translation	EMBOSS Transeq documentation; use a tested local installation or equivalent course tool.
Prediction and confidence	ColabFold; AlphaFold DB for reference proteins; EMBL-EBI confidence guide.
Structure and classification	RCSB PDB; SCOP; CATH; the course molecular viewer.

Check access, account requirements and compute availability before Tuesday. Archive essential inputs locally. A failed web service or unavailable GPU must not prevent the group from completing the interpretation tasks.

Key bibliography

- [1] Singh, D. et al. *Transforming fusions of FGFR and TACC genes in human glioblastoma*. Science **337**, 1231–1235 (2012). [10.1126/science.1220834](https://doi.org/10.1126/science.1220834).
- [2] Parker, B.C. et al. *The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma*. J. Clin. Invest. **123**, 855–865 (2013). [10.1172/JCI67144](https://doi.org/10.1172/JCI67144).
- [3] Di Stefano, A.L. et al. *Detection, characterization, and inhibition of FGFR-TACC fusions in IDH wild-type glioma*. Clin. Cancer Res. **21**, 3307–3317 (2015). [10.1158/1078-0432.CCR-14-2199](https://doi.org/10.1158/1078-0432.CCR-14-2199).
- [4] Frattini, V. et al. *A metabolic function of FGFR3-TACC3 gene fusions in cancer*. Nature **553**, 222–227 (2018). [10.1038/nature25171](https://doi.org/10.1038/nature25171).
- [5] Wang, C.G. et al. *Oncogenic driver FGFR3-TACC3 requires five coiled-coil heptads for activation and disulfide bond formation for stability*. Oncotarget **14**, 133–145 (2023). [10.18632/oncotarget.28359](https://doi.org/10.18632/oncotarget.28359).
- [6] Jumper, J. et al. *Highly accurate protein structure prediction with AlphaFold*. Nature **596**, 583–589 (2021). [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).
- [7] Mirdita, M. et al. *ColabFold: making protein folding accessible to all*. Nat. Methods **19**, 679–682 (2022). [10.1038/s41592-022-01488-1](https://doi.org/10.1038/s41592-022-01488-1).
- [8] Andreeva, A., Kulesha, E., Gough, J. & Murzin, A.G. *The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures*. Nucleic Acids Res. **48**, D376–D382 (2020). [10.1093/nar/gkz1064](https://doi.org/10.1093/nar/gkz1064).
- [9] Sillitoe, I. et al. *CATH: increased structural coverage of functional space*. Nucleic Acids Res. **49**, D266–D273 (2021). [10.1093/nar/gkaa1079](https://doi.org/10.1093/nar/gkaa1079).
- [10] Rice, P., Longden, I. & Bleasby, A. *EMBOSS: the European Molecular Biology Open Software Suite*. Trends Genet. **16**, 276–277 (2000). [10.1016/S0168-9525\(00\)00204-2](https://doi.org/10.1016/S0168-9525(00)00204-2).



15-18 September 2026, Palacio de los Condes de Valdeparaíso, University of Castilla-La Mancha, Almagro (Spain)

Mini-Project: *Modelling Brain-Metastasis Growth and RT Response*

Supervisor: *Gabriel F. Calvo*

Motivation

Brain metastases (BMs) are the most frequent intracranial malignancies in adults, affecting up to 30% of all patients with solid tumours, showing particularly high incidence rates in lung cancer, melanoma, and breast cancer [1-2]. They are a major cause of neurological morbidity, markedly impairing quality of life, and carrying a poor prognosis despite therapeutic advances. BMs are biologically distinct secondary tumours shaped jointly by the evolving primary tumour (PT) and by selection within the brain microenvironment, creating organ-specific vulnerabilities and resistance mechanisms that cannot be inferred from the extracranial disease alone.

The metastatic cascade (Fig. 1) begins when cancer cells detach from the PT, enter the circulation and reach the central nervous system. Only a small fraction cross the vascular barrier, evade elimination and establish a metastatic niche. Brain colonisation involves vascular co-option, blood-tumour-barrier remodelling and reciprocal signalling with neural, glial, vascular and immune cells [3-5]. Lesions may remain few (oligometastases) and spatially confined, or numerous colonies may emerge over time. These states depend on metastatic seeding, lesion-specific growth and the tumour-brain microenvironment. Molecular and spatial studies reveal marked heterogeneity, but their findings remain difficult to connect to serial human imaging and lesion-level treatment response.

Treatment includes surgery, stereotactic radiotherapy and systemic therapies. These can provide local control and symptom relief, but individual lesions may respond differently. Clinical decisions still rely substantially on population-level evidence and the interpretation of serial images. A quantitative model offers a complementary way to express assumptions, estimate changes over time and test predictions against a later scan.

This project uses real MRI-derived volumes and treatment histories from the OpenBTAI resource [6]. Its aim is not to explain every process. It is to determine what a simple mechanistic mathematical model can learn from the available measurements, and where those measurements are insufficient.

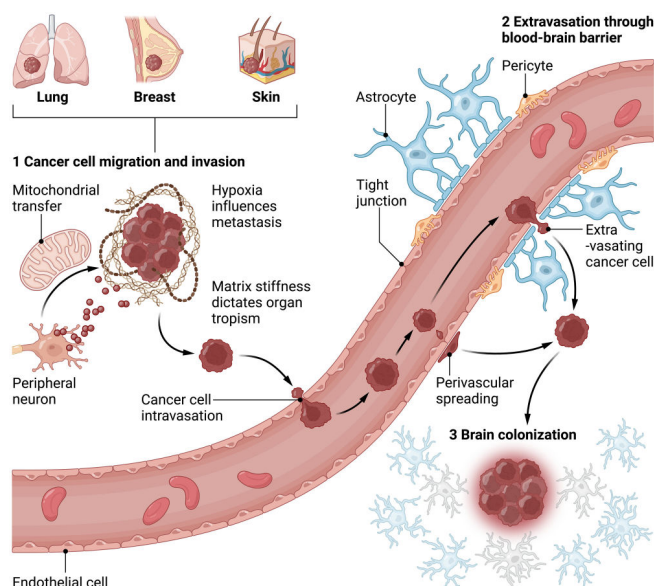


Figure 1: The metastatic cascade. Tumour cells enter the circulation, cross the blood-brain barrier and establish brain colonies. Colonies may die, remain dormant or progress. The present project begins after a lesion is visible on MRI; it does not infer its unobserved seeding history.

Specific biological context of the problem

Radiation damage is not immediate disappearance. Radiation can prevent cancer cells from sustaining further growth. Damaged cells and tissue may nevertheless remain within a visible lesion. Their subsequent death and clearance take time. A model of treatment response should distinguish loss of reproductive capacity from loss of physical volume.

Enhancement is an imaging observation. Contrast-enhanced T1-weighted MRI depicts regions in which contrast accumulates. The enhancing mask is not a direct count of viable cancer cells. A treated region can contain viable tumour, damaged tissue and inflammatory changes. The model must state which of its populations contribute to the measured signal [3,6].

“Necrosis” is the source’s segmentation label. The OpenBTAI resource distinguishes enhancing tissue from a central non-enhancing or necrotic region [6]. We retain this label in the data files. It is not proof of histologically confirmed radiation necrosis. A non-enhancing region may already be present before treatment, and its measured baseline volume must be retained.

Separate components can move in different directions. Total volume may increase while enhancing volume decreases. Formation of non-enhancing material, delayed removal and changing enhancement are possible contributors. These possibilities motivate compartments, but MRI volumes alone cannot identify which mechanism dominates.

Treatment acts at different levels. Whole-brain radiotherapy (WBRT) concerns the patient, whereas focal irradiation and surgery concern particular lesions. Systemic therapies can affect several lesions. Pool patient-wide histories across the clinical rows, but do not assign one lesion’s SRS to every other lesion in that patient.

A treatment-free scan is not necessarily a treatment-free interval. A drug course may occur between two MRIs. Previous radiation can also have persistent effects after its delivery has ended. A later scan can initialize a model, but resetting the clock does not erase treatment history.

Central scientific question

Can a simple ODE model explain the separate enhancing and central non-enhancing components of a BM and improve the prediction of a later MRI over a simple baseline?

Dataset origin and access

Ocaña-Tienda and colleagues published an imaging study of **75 patients with 260 BMs**, together with clinical information and segmentations [6]. The complete resource and the two Excel extracts supplied for this project are not interchangeable and come from the full files available in:

MOLAB dataset page: [clinical and radiomic downloads](#)

Dataset description: [Scientific Data 10, 208 \(2023\)](#)

Archived imaging and data collection: [Figshare](#)

To facilitate the analyses, you will use the two file extracts: `OpenBTAI_METS_ClinicalData.xlsx` and `OpenBTAI_Radiomics.xlsx`. Pay attention to patient ID, lesion id, and the MRI follow-ups dates in `OpenBTAI_METS_ClinicalData.xlsx` (those marked in blue have MRI-associated quantities). In `OpenBTAI_Radiomics.xlsx` you have the specific values of the MRI-associated quantities for each patient and lesion. No missing MRI volume has been simulated or interpolated.

Dataset, baseline and quantities to represent

The unit of analysis is one lesion over time. Match patient identifier, lesion identifier and MRI day. Several segment rows from one visit are not several time points. Multiple lesions from one patient are not independent patients. Lesions in the same patient may display a divergent progression as they evolve.

Count rules explicitly. Prefer a measured baseline plus four follow-ups, giving **five visits**. Also report the larger set with **four visits in total**. Four observations are a practical screening threshold, not a guarantee that two or more parameters can be estimated reliably.

Keep two clocks. Retain the original study day d . The resource's day 1 is tied to the first BM identified in a patient, not necessarily the first appearance of every later lesion [6]. For a chosen measured baseline d_* , use elapsed time $t = d - d_*$ in the model. Translate all treatment dates by the same amount. A later untreated scan may serve as the model baseline; document why the whole fitted interval is suitable. Label a post-treatment baseline as such.

Measured quantities. Represent total volume by $V_T(t)$, contrast-enhancing volume by $V_{CE}(t)$ and central non-enhancing volume by $V_N(t)$. Volumes are in cm^3 after converting the exported voxel-volume feature from mm^3 . The segmented partition is $V_T = V_{CE} + V_N$.

Possible model states. An initial model may represent the enhancing and non-enhancing components by $E(t)$ and $N(t)$. A radiation model may distinguish proliferative tissue $P(t)$ from lethally damaged tissue $D(t)$. These are model-equivalent tissue volumes, not measured cell counts. Select the necessary states and explain their relation to the observations before writing the equations.

Conditions that the model and analysis should respect

Volume provenance. Use the voxel-volume feature. Do not silently replace it with mesh volume. Distinguish an explicit necrosis ROI, a zero remainder verified by equal total and enhancing voxel counts, and a positive remainder calculated without a separate necrosis ROI.

Biological balances. Explain every source, transfer and removal term. Keep states non-negative. A radiation-damage term need not immediately reduce the sum of all compartments.

Recorded treatment only. On the original study clock, 040019 has WBRT starting at $d = 15$ and crizotinib at $d = 15$ -134. Its WBRT fraction count and end are unusable; no SRS is recorded. For 040074/L1, a course entered in the SRS fields spans $d = 29$ -51: 3.7 Gy per fraction, 10 fractions. Pemetrexed plus platinum is recorded at $d = 271$ -410. Individual fraction dates are unavailable. With baseline $d_* = 1$, subtract one from these dates for model time.

Parsimony and a genuine held-out value. Estimate only a few coefficients. Do not fit total as an independent third channel when it is the sum of CE and N. Remove the final MRI before the prediction fit. Do not choose the final model because it happened to predict that withheld value best.

Questions to address: from data to a model

Q1. Which trajectories can actually be used? Audit the patient, lesion and visit structure. Compare five visits with four visits in total, and diagnostic with later measured baselines. Check the complete treatment interval before calling a trajectory untreated. Explain how the radiomic timepoint labels match the clinical study days. Verify one volume partition and its source cells. Report eligibility with and without positive derived N remainders.

Output: a cohort table, a documented selection rule and one hand-checked record. Distinguish radiomic coverage from eligibility for untreated growth or radiation-response modelling.

Q2. What do the measurements and treatment histories mean? Plot all three observed volumes for one lesion. Place its relevant treatments on the same time axis. Identify an interval in which total and enhancing volumes change differently. Explain why neither a growing total volume nor a growing N mask alone diagnoses progression or radiation necrosis.

Output: a measured trajectory, a treatment timeline and a short statement of the observation assumptions.

Q3. How can growth be estimated without inventing an SRS effect? Calculate an effective exponential rate between two real visits. Then estimate an initial volume and a constant effective growth rate from several visits. For 040019, describe the initial decline separately and construct a minimal regrowth model anchored at the first observed follow-up. Include formation and removal of non-enhancing tissue when justified. Derive an explicit solution where possible. Distinguish net volume change from a cellular birth rate.

Output: a hand calculation, explained model terms and a baseline fit. State why treatment-exposed rates cannot automatically be called untreated proliferation rates.

Q4. How should radiation be represented? For 040074/L1, explain why proliferative and lethally damaged compartments might contribute to the same enhancing region. Preserve baseline N. Represent the recorded fractionated course without inventing fraction dates. Check positivity and the total-volume balance. Decide which immune or inflammatory states cannot be estimated separately from these data.

Output: a compartment diagram, an observation map and a numerically checked radiation-response model. As an extension, adapt the treatment input to the single-fraction case 040073/L2 with its later baseline.

A manageable core with a larger real-data extension

One lesion and one model are enough for a meaningful core analysis. For instance, if you consider 040019/L1 for the simpler formation-clearance route or 040074/L1 for the radiation route. Lesion 040019/L2 provides a comparison within the same patient.

An additional example, 040073/L2, has four explicit contrast enhancing CE-N observations at original days 732, 862, 962 and 1066, with one recorded 18-Gy fraction on day 750. Use day 732 as a later model baseline, while acknowledging prior WBRT and systemic therapy.

The coverage audit also provides every lesion with at least four measured volumes. A later starting scan can enlarge this treatment-aware pool. It does not, by itself, establish an untreated growth series. Do not attempt to fit every mechanism or every patient within three afternoons.

Questions to address: from fitting to interpretation

Q5. How are parameters fitted? Define a model-measurement residual and explain its scale. For a compartment model, fit CE and N without adding total as an independent observation. Choose and justify bounds. Start the numerical search from several guesses. Report parameter units, fitted values, the number of independent time points and at least one discrepancy that the model cannot explain.

Output: the fitting criterion, parameter table, reproducible code and an observed-versus-fitted compartment plot.

Q6. Can the model predict a later MRI? Remove the final visit before fitting. Predict its CE, N and total volumes. Calculate signed and absolute errors. Report percentage errors only when meaningful. Compare with persistence: carrying the previous measured volume forward. Distinguish an all-visit fitted curve from a prediction based only on earlier visits.

Output: a fit-forecast figure and a table comparing both methods at the same endpoint and horizon.

Q7. Which conclusions are robust? Change one modelling assumption or discrepancy measure. Alternatively, fix one coefficient at several trial values and re-estimate the others. Explain why a parameter can influence a prediction strongly yet remain poorly determined. Discuss how a radiation-effect estimate depends on delayed death, the observation map and systemic-treatment history.

Output: one robustness test with an explicit biological interpretation. Do not turn an arbitrary parameter range into a confidence interval.

Q8. What follows biologically? State what the model explains and what it cannot establish. Identify one additional real measurement or source-record check that would most improve the analysis. A model that loses to persistence is an acceptable outcome when the implementation is correct and its failure is interpreted clearly.

Output: a short conclusion that separates observed facts, model assumptions and untested explanations.

Reporting boundary

A good retrospective fit is not evidence of a uniquely identified mechanism. MRI-mask volumes do not establish a clonogenic surviving fraction, a histological diagnosis or a treatment recommendation. The final report must preserve uncertainty in treatment dates, missing ROIs and any assumed mapping between tissue and images.

Expected learning outcomes

Turn a clinical-radiomic export into an auditable lesion trajectory. Express an ODE as a biological balance. Fit a small model, test a later observation and recognise when its assumptions exceed the information in the data. Communicate a reproducible result, including a negative predictive result when that is what the data show.

Work plan, files and expected outputs

Tuesday afternoon, 15:30-19:00: brief biological introduction; data audit; measured-volume plots; first growth calculation and model formulation.

Wednesday afternoon, 15:30-19:00: implementation, mathematical checks and parameter estimation. Show the supervisor the fitted compartments and one unexplained feature.

Thursday afternoon, 15:30-19:00: withheld-MRI predictions, one robustness test and preparation of the final figures. Use any additional project time shown in the final programme. Friday morning is reserved for final preparation and presentations.

Participant files. The package provides the handout, the two accompanying graphics, a data dictionary, a model-free starter notebook and audited CSV tables. These include `teaching_three_lesions.csv`, `coverage4_longitudinal.csv`, `coverage4_reliableN_longitudinal.csv`, `patients_ranked.csv`, `lesions_audited.csv`, `single_fraction_SRS_case.csv` and `treatments_patient_and_lesion.csv`. The original workbooks remain the source of record.

Minimum deliverables. A cohort rule; one measured-data figure; explained equations and observation map; fitted parameters; a fit-forecast comparison including persistence; documented code; and a concise biological conclusion. Follow the final programme: a 15-20-minute team presentation, 10 minutes of questions and a 1-2-page technical note (optional). Every trainee should understand the principal decisions.

Optional extensions. Compare lesions within a patient. Fit descriptive two-parameter growth baselines across the larger coverage set. Add continued formation of non-enhancing tissue. Compare delayed-death assumptions. Construct a re-optimized parameter profile. Investigate whether an additional listed MRI could yield a real pretreatment segmentation. Do not create new measurements to favour a model.

Key bibliography

- [1] Valiente M, Anders C, Boire A, et al. *The evolving landscape of brain metastasis: volume II*. Trends in Cancer **12**, 116-146 (2026). doi:10.1016/j.trecan.2025.11.007.
- [2] Chafe SC, Mobilio D, Zhai K, Kieliszek AM, Venugopal C, Singh SK. *Emerging paradigms in the study of brain metastases*. Science Signaling **19**, eadj9726 (2026). doi:10.1126/scisignal.adj9726.
- [3] Steeg PS. *The blood-tumour barrier in cancer biology and therapy*. Nature Reviews Clinical Oncology **18**, 696-714 (2021). doi:10.1038/s41571-021-00529-6.
- [4] Schreurs LD, vom Stein AF, Jünger ST, et al. *The immune landscape in brain metastasis*. Neuro-Oncology **27**, 50-62 (2025). doi:10.1093/neuonc/noae219.
- [5] Dong W, Sheng J, Cui JZM, Zhao H, Wong STC. *Systems immunology insights into brain metastasis*. Trends in Immunology **45**, 903-916 (2024). doi:10.1016/j.it.2024.09.010.
- [6] Ocaña-Tienda B, Pérez-Beteta J, Villanueva-García JD, et al. *A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data*. Scientific Data **10**, 208 (2023). doi:10.1038/s41597-023-02123-0. **Essential dataset reference.**