# An Agent Based Modelling Framework for Dynamic Biological Systems and Applications to Cancer Cells, G Protein Coupled Receptors and G Proteins

**By Sam Robin Edward Benkwitz-Bedford**

A thesis submitted to the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Centre for Computational Biology
Institute of Cancer and Genomic Sciences
College of Medical and Dental Sciences
University of Birmingham
December 2021

# Abstract

Dynamic real-world biological systems are quite difficult to study because it requires us to understand the interactions within, without being able to isolate them. This hidden complexity means, when designing representative models, problems often arise in appropriate representation, abstraction and applicable comparison with real-world phenomena.

The way distinct systems evolve over time by interaction between population members and their environment can generate emergent patterns of behaviour, here we observed it indirectly via visualisation of movement. This work centres on improving our understanding of real-world complex dynamic spatial biological systems, looking at two example populations: Cancer cells and G-protein-coupled receptors (GPCRs), for support of biological exploration and hypothesis development.

A framework was developed to take sets of population tracks and digitise them in a unifying representation for observation that could also be used to design representative models. Representative visual patterns were found and replicated: strand-like movement patterns from Cancer cells and movement hot-zones in GPCR and G protein sets. We isolated and visualised movement choices in relation to position and time. Artificial neural nets (ANN) were also applied to image classification; generating similarity measures between model and biological systems. Populations could also be split with ANNs on individual track morphology to assess specific pattern subsets. We successfully developed and applied our framework by applying generalised analysis and modeling tools to gain insight into our chosen biological systems.

# Acknowledgments

I would first like to acknowledge and thank my supervisors for creating the project, Dr Shan He and in particular Professor Jean-Baptiste Cazier whose constant guidance, time, and encouragement over the last four years has been incredible; this project would not have been possible without it.

Initial project development was focused on Cancer cell data which Dr Anderson Ryan provided, I am extremely grateful to him. Also to Jacopo Credi who had completed some preliminary analysis and thus helped shape the project.

Similarly, Dr Sabine Hauert provided initial input and discussion, essential to our investigation of a possible robot model, although not eventually included it was greatly appreciated.

My thanks to Professor Davide Calebiro and his group; for access to their GPCR and G protein data and their insight. I particularly appreciated the help of Dr's Yann Lanoiselee and Zsombor Koszegi, discussion with both was always helpful and ever patient.

The support of the UoB Advanced Research Computing group and in particular Research Software group leader Dr Andrew Edmondson was very helpful. Advice, hardware and software support were all graciously provided and extremely well received.

I worked primarily within the Centre for Computational Biology (CCB), everyone, from students to staff was always friendly; I have never known such a conducive environment and all deserve thanks. I would like to particularly thank the administrative group, Alice, Jessica and

Jordan whose help I could always and frequently did rely on, Chris for his company while rubber ducking and Danesh for his guidance and patience while completing the project.

Finally, I have also had a great deal of indirect help throughout, providing me with the stability to focus and keep working. To my family: parents Jo and Nigel, and sister poppy, thank you, above and beyond. To my friends: whom I met (and suffered with) through the process Agata, Greg, John and Nadia and otherwise Adam, Angeliqa, Ash, Christopher, Pete and Isak, your support has been invaluable and its contribution to my sanity, immeasurable.

I am and will always be grateful to you all.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Background

Observing, interpreting and understanding real-world disease is a long and complex, yet necessary, challenge to ultimately change disease outcomes. Here we use a modelling approach to enable observation, facilitate interpretation and improve our understanding.

The way distinct biological systems evolve over time by interaction between population members and their environment can generate emergent patterns of behaviour. We aim to improve our understanding of real-world complex dynamic biological systems with a framework level modelling approach, looking at two example populations: cancer cells and G protein-coupled receptors (GPCRs).

Cancer, our first system is one of the leading causes of death worldwide, its effects are wide reaching, as such it is an area where increased knowledge can be very impactful. With the number of new cases per year projected to rise from 14 million in 2012 to over 22 million by 2030, the impact of both mortality and extended treatment are significant [1]. Improving methods for better treatment, early diagnosis, prevention and monitoring via investigation, are an important focus for research. Developing our understanding of cancer as a system of interactive cells in motion, could provide invaluable insights to move closer to these goals.

As a broad and heterogeneous disease, there are many approaches to investigating cancer, even in understanding cell motility. Existing studies cover a wide variety of topics often relating to metastasis, from invasion by individuals [2] or groups [3], modifying their environment [4] or being driven by it [5] to behave in an invasive manner. However, it is difficult to find work addressing general localised motion within a permissive environment; at this scale movement of

populations is usually studied as a mechanism of movement around a larger tumour structure.In this work, we focus on an individualistic approach to understanding cell population movement and interaction, with their environment and other population members. A better understanding of what drives the behaviour of unaffected cancer cells could provide useful insight for the prediction of environmental change or treatment response.

The second system is of a very different scale and nature: G protein-coupled receptors (GPCR), a large and diverse group of cell surface receptors. To communicate extracellular behavioural changes, receptors interact with proteins in the plasma membrane and react when, for example, co-localizing to send signals into a cell. GPCRs enable cells to sense and then react to a wide range of environmental changes, also facilitating communication [6, 7, 8]. Therefore, we can study GPCR behaviour over time to further understand how a cell identifies, controls and apprehends its environment.

More detailed observation of GPCR and G protein movement and interactions have been made possible by imaging improvement in recent years. However, while extensive work has been done to understand and characterise the system, many questions still remain [9]. Our development of representative models and visualisation will help to further characterize interactions such as cytoskeletal confinement [10, 9].

In this work we develop a framework and use it to digitize real-world data and then define representative models for both systems. The framework enables an analysis with a focus upon movement and pattern based meta data extraction, observation and interpretation, for knowledge development.

## 1.2 Biological systems

### 1.2.1 Biological systems and modelling

Our work can be described as the process of translating a real-world biological system into observations which lead to the definition of representative models, thus enabling analysis. Furthermore, it allows us to test *in silico* the effect of a variety of perturbations to the system such as different environmental effects or varying population distribution. We aim to define a novel scientific approach to problem solving and experimental understanding via the study of systems, implementation of change, and observation of subsequent effects. Ordinarily, we can extract some information directly from real-world or *in vitro* observation. Here we attempt to more indirectly observe systems, movement visualisation and pattern identification improves possible analysis, identifying new patterns of behaviour. Biological systems can be complex with large sets of interdependent interactions leading to emergent behaviours and effects, making accurate observation difficult. Further, many effects arise as a combination of causes, obscuring the importance of individual features or appearing random [11, 12, 13]. In defining a system, we need to assemble our available knowledge into a form that can further improve information extraction and interpretation [14, 15, 16]. Over time, we can combine observations into hypotheses for the causal understanding of these systems, a conceptual model.

Conceptual modelling is the process of defining a system, a tool for analysis and a step towards understanding it [17, 18, 19]. There are at least as many ways to analyse and interpret biological data as there are data types or systems [20, 11]. A conceptual model can help us assess and test assumptions about a biological system. Specificity and definition force logical causality to be assessed, interpreting interaction by incrementally testing our assumptions. We attempt to describe a system as the interaction of many parameters with varying importance. When referring to a motile population, each entity interacts with each other and their environment to drive movement choices often by definition. Other processes such as behavioural change, death or replication can also occur. Here, we are analysing the way populations of individual entities move and interact over time. We will need to define interacting parameters for population members, motility, time and environmental effects.

Testing our conceptual model allows us to observe the referential logical integrity of interaction, such tests help us explain real-world causes and effects and better understand a system [21, 22].

Further, to observe the subsequent effects, we can introduce factors such as environmental extremes that may be difficult to create in real-world or *in vitro* experiments. This exploration is often a valuable step towards further model improvement and knowledge development. An accurate and predictive model can help direct attempts to engineer systems, ultimately exercising some level of control for positive real-world change [23, 24, 25, 26].

For example, a predictive model of cancer cell movement might indicate that such cells can't move as freely if physically blocked within their environment. We can hypothesise that increasing environmental density would stop motility and subsequent deleterious effects, and then simulate the scenario with the model to quantify the effects. If the effect of the tested intervention is successful, experiments can then be performed, eventually it may lead to a new treatment. However, often and in this case, real-world systems are not as simple as we may like. Our conceptual model lacks a representation of the way such treatment would affect normal healthy bodily function; ultimately such real-world experiments may not only fail but also be unethical and have a high human or animal cost. A model can be a powerful way of utilising what we know, but always with the understanding that there remains much that we don't know.

So, to entirely represent all possible scenarios and reactions within a system, our model needs to be constructed with fully representative knowledge. If we seek to understand a system with models, this may look like a cyclical or infinite regress problem: we need representative knowledge to correctly model, and a model to develop our knowledge. We therefore need to differentiate between explanation and prediction, drivers of model definition and operation. We can use models to test, illustrate and explain our hypothesis or predict real-world outcomes for situation modification. There is some overlap between each purpose but the accuracy of representation and therefore required knowledge needs to be tuned to each use case. Every model is limited, computationally or physically [27, 28]. Limitation often refers to the scope of included parameters; models are only effective in their specialised scope, 'all models are wrong, but some are useful' -George Box [29]. For conceptual modelling, scope, accuracy and complexity are often limited by the human ability to maintain an understanding of multiple interacting parameters. We need a tool to formalize, define, assess, validate and test the interactions of complex systems [30, 31, 32, 33].

### 1.2.2 The cancer cell system

We have described cancer as an important area for research, improving methods of treatment, early diagnosis, prevention and monitoring. Therefore, improving understanding via modelling is particularly valuable. Further, occurring in a wide range of places in the body as well as tissue types, cancer is a complex and difficult subject for effective modelling because of its multi-scale heterogeneity. Cancer can differ greatly from patient to patient and at all scales, there are a great many differentiated and diverse forms. As such, when observing cell level cancerous micro-environments, we need to remain aware of subsequent population heterogeneity [34, 35, 36, 37, 38]. This breadth of possible behaviours necessitates a generalizable approach that can be applied in a flexible way. In the most general terms cancerous tumours emerge from malfunctions in the process of cell replacement and the resultant life cycle of an affected cell [39]. As replication continues these cells are influenced by, and change, their local micro-environment, an important two-way interaction. If the cancer establishes and spreads, changes over time can become detrimental on a larger body-wide scale leading to a variety of negative symptoms, eventually becoming fatal. We sought to integrate and detail an alternative *in silico* approach to the digitisation, interpretation and exploitation of data gathered from cancerous cell observation.

Cell movement can be indicative of a wide range of underlying causes and interactions, population or environmentally driven. Because of this two-way interaction it is important to consider the micro-environment as a complex heterogeneous group of cancerous and non-cancerous cells, but also as a spatially complex set of geographical co-localizations and localities. The relationship can also be co-operative with interaction between non-malignant cells supporting tumour growth [40]. Micro-environmental interactions of cancerous and non-cancerous cells often drive growth rates and morphology [41, 42]. For example, cancerous tumours often form necrotic cores due to nutrient competition and eventual hypoxia, eventually tumours can address this via development of angiogenesis [43, 44]. Response to nutrient starvation can be observed to cause movement, chemotaxis driven by nutrient gradients within the micro-environment, the process also functions as a form of immune evasion similar to that of suppression [45, 46].

To analyse cancer cell population interaction and movement within our model framework we need real-world or *in vitro* observed data for tracking. We were given access to non-small lung cancer cell videos of which 5 were of sufficient length and population density to produce inter-

esting positional data sets post tracking [47]. The first four were 12 second 84-time increment data sets and the fifth longer 54 second video processed into 181-time increments with attendant tracks. Previously, work had been performed to track and analyse the cell movement but without modelling as a next step [48]. We therefore benefit from initial applicable pattern information to inform and direct investigation, cell movement had been observed to follow paths. Cells were re-tracked to ensure uniform application and input as a real data source, digitised videos became maps of movement, turn preferences and population growth over time for identification of representative patterns.

### 1.2.3 The G protein Coupled Receptor system

We can improve our understanding of fundamental cell processes, pharmacological targets, and drug performance by developing and testing models of GPCR interaction and behaviour. Thus far, obtaining and analysing tracks with new observation techniques has revealed deep GPCR domain complexity [6]. Therefore, application of further multidisciplinary approaches to analysis and observation should drive progress in the field. Important new insights might be gained by introducing new combined approaches to data visualisation, modelling and analysis. Regarding pharmacological improvement, GPCRs often influence and translate the biological effects of hormones and neurotransmitters, in doing so they represent reasonable targets. While a large portion of currently marketed drugs target GPCR (30-35%) [49, 50] only a fraction of targetable GPCRs are exploited [6]. We could further define targets, interaction, co-localisation and behaviour of GPCRs still not entirely understood by observing changes in local behaviour and diffusion [51, 7].

GPCRs enable cells to sense and then react to a wide range of environmental changes, also facilitating communication [6, 7, 8]. Receptors in the cell membrane can co-localize and then interact with G proteins; membrane associated protein molecules involved in signal transducing. A signal is transmitted to the interior of a cell, the exact scenario for interaction being ambiguous and the subject of other research [51, 52, 53]. Therefore, to better explain the process *'the 2A-adrenergic receptor (2A-AR), a prototypical family-A GPCR that couples strongly with the inhibitory G protein (Gi)'* was chosen [52, 54]; by observing a typical GPCR, the general behaviour of other GPCRS can also be understood. Movement was recorded and extrapolated

via tracking to generate paired sets of receptors and proteins that we can digitise and model.

The available populations were previously classified into behavioural states based upon diffusion properties in the cell membrane for both receptors and G proteins: four groups ranging from virtually immobile to fast diffusing. State change to more immobile patterns was particularly prevalent, it was suggested that the slower states were caused by compartmental trapping; semi permissive boundaries retaining entities much like a net. Protein and GPCR co-localization lasted around a second. Under ordinary conditions, proteins and lipids within the cell membrane have been observed to undergo hop-diffusion, jumping between membrane compartments within which they are trapped [6, 55, 56, 57]. A 'fence and picket' model of the plasma membrane suggests that it is sub divided by actin-based skeleton 'fences' and trans membrane protein 'pickets'. The model was used to explain compartmentalization of nanodomains and barriers affecting free diffusion [56, 58]. Such a model might enhance co-localisation and in turn biochemical reactions such as protein receptor binding by confinement of GPCRs within these spaces [59, 6].

## 1.3 Computational modelling

Conceptual models can be improved iteratively by placing them within a computer interpretable ruleset, such as programmatic logic [60, 21, 22, 61]. We formalise and enable automatic logical testing by taking a step from conceptual model to computer model, empowering our process of system analysis. Further, tools applicable to several systems/models can be created by using a generalizable approach and common metrics such as population movement. We propose to further understand biological data with computer vision, natural computation and pattern-based comparison of population movement for novel new insight [62, 63, 64].

Computer modelling in general aims to implement relationships and interactions that represent an environment at different points in time *in silico* [44, 30, 65, 16]. Once a basic implementation of a model is completed, it can be manipulated, extended and operated to explore hypotheses and observe the emergence of complex phenomena that might be otherwise inaccessible. As with our conceptual models, the central hypothesis is that if a model is built using known mechanisms, by adding a theoretical effector and being judged valid, via comparison to existing real-world

or *in vitro* data, it supports the existence of that theory, or a similar one [66]. The availability of computational power continues to improve year on year; therefore, model complexity can improve with it. The implementation of additional effects such as entity to entity interaction and environmental phenomena, along with more powerful automated data extraction tools, becomes increasingly relevant. A formalised computational environment enables our models to become more complex and informative with representative validation, through step-wise improvements.

We assume that the biological systems we want to study consist of interacting populations; these interactions translate to behavioural change and can be observed in movement patterns. When attempting to understand a system or environment we therefore try to develop a theory of causal interaction and create a representative model that can be operated and examined further. We can then expand our understanding with simulations run to support parts of the causal conceptual hypothesis and drive a cycle of improved understanding. However, system complexity can lead to difficulties in parameter management at greater scales [67, 68]. Unmanageable conceptualisation, as a model becomes more complex it becomes easier to misinterpret or overlook mistakes.

The transfer to computational model formalisation within an *in-silico* approach can help compartmentalise a problem into a series of more manageable sub problems in the form of narrow focus models. Converting a conceptual model to a more formalised version forces conversion through a mathematical or programmatic interpretation layer. Therefore, fundamental issues and conversion difficulties leading to oddities in development and output can highlight logical incompatibility. Any single model is likely to have issues, being able to rapidly alter scope via input parameters would be advantageous for initial and ongoing investigation. Developing a framework of model generating tools may take more time initially but allows us the flexibility to rapidly explore the systems we seek to understand. Further, by focusing upon population movement we can be flexible about the target system as well.

### 1.3.1 Modelling background

The applicability of an approach is often simply defined as whether they can represent a system. However, ease of application and conceptualisation are important concepts, if we cannot understand the model definition, observing and extracting representative information becomes very difficult. Any model is made more complex with the addition of more parameters and interactions. Our approach to understanding cancer cell and GPCR and G protein movement has an emphasis on spatial relationships over time. Models with multi-dimensional movement in free space can become parametrically complex very quickly. A more programmatic definition should therefore be appropriate to breaking down each target pattern into manageable representative model definitions.

We must carefully define parameters and their relationships throughout design and application from the ground up. Model definition is usually driven by variable issues such as problem space, representation specificity, available data, individual expertise and exercise objective or output [12, 13, 69]. In defining an environment, relationships and interactions between entities and environment need to be strong and specific enough to represent their real-world counterparts and generate comparable patterns. Often seemingly less impactful, variables can still have important indirect effects. Model definition dictates applicability, usefulness, validity and even overall viability. Models, vision and comparison approaches all rely upon proper problem and variable definition to generate good results in reasonable time. Definition also needs to be generalised, conceptually and computationally to be understandable and widely applicable.

#### Differing approaches

Computer modelling is a term that can apply to many of a large and varied field of approaches, each with their own strengths and weaknesses, methods of definition and levels of abstraction. Mathematical models employ a wide range of approaches and sub methods. Fundamentally, many computational models are still based on mathematics, often a programmatic level of abstraction with a mixed rule set. For example, ordinary differential equations (ODE) can be used to define the relationship between a variable and its derivative [70, 14]. Nutrient availability and tumour size over time might be tied to represent the relationship between nutrition and proliferation rate. Such an oversimplification then requires additional rules and associated parameters,

such as maximum number of cell can divisions in a time span or a spatial component dictating if there is space to expand into. A model may therefore consist of many different competing equations, non mutually exclusive and therefore inter-dependant. Developing a definition and understanding in purely mathematical terms can quickly become conceptually difficult. Purely mathematical models can be encapsulated within the definition of computer models, but a more abstract logic driven, or mixed approach can also be used.

Procedural instruction based sequential languages have classically been closer to machine code and benefit from compute speed. Object oriented languages tend to benefit from a baseline compartmentalised approach, everything being an object that acts upon another object. While selection is often driven by user familiarity, problem scale and the abundance of relevant available and applicable libraries is also important. We aim to be able to develop models at multiple levels of complexity in a flexible manner, an object-oriented language offers a fundamentally bottom up approach to hyperparameter compartmentalization.

#### Model definition

Due to so many, and often conflicting, requirements, it is useful to again remember that "all models are wrong, but some are useful" [29]. Some common objectives can compete without being mutually exclusive, quantitative versus qualitative analysis for example.

Given that it can be difficult to properly define model parameters for representation of real-world phenomena, the identification of target behaviours can require extensive initial investigation alone. Therefore, we need to take an ongoing iterative development approach, with successive rule/parameter addition as part of identifying appropriate model construction. We can start with a simplest case model and add parameters, identifying hyperparameters as we test and iterate our own conceptual hypothesis; results can also be confirmed and hypotheses supported by iteration with new models. Appropriate definition is then driven by available information. Again, we need to place emphasis upon the construction of tools for model interpretation and metric patterns as well.

#### Pattern emergence

Behavioural emergence often observed in machine learning or modelling refers to the confluence of interactions and circumstance resulting in indirectly defined population reaction or effect [71, 72]. We should be aware of a counter-intuitive relationship between serendipitous behavioural emergence and selection of appropriate parameters: we have already stated that a significant number of parameters and dependant interactions is often necessary and causes model complexity. However, counter-intuitively, model breadth can have the opposite effect and reduce cases of behavioural emergence. Higher complexity without appropriate definition reduces the power of each interaction, what may have occurred quickly with a few well-placed interactions becomes difficult to create. There are many possible reasons such as missed time sequences, appropriate definition should therefore seek to allow for nuanced variance and more generalised implementation where needed.

By remaining flexible and selecting a paradigm that allows access to a wide range of representative parameters we can iteratively address risk through appropriate conceptual selection.

### 1.3.2 Appropriate conceptual selection

When considering appropriate modelling approaches, we also need to define the conceptual use case. We primarily aim to generate models that can explain real-world or *in vitro* observed patterns but with the flexibility to advance and generate useful predictions about real-world systems if the opportunity allows. We also hope to explore alternative novel hypotheses when possible. A flexible framework for definition and analysis allows us to aim for inclusion of multiple conceptual uses with priority defined by our use case.

An **explanative** model may commonly be employed to show and clarify a specific interaction or emergent trend. Rather than representing an entire environment it may be prudent to identify only the expected actors and implement their representations. If this does not create output like that observed in the compared system, part of the causal process has been missed. Similarly, too many effectors can be applied, an explanative model might be narrowed incrementally to identify the minimum requirements for pattern emergence.

**Explorative** models might assemble representations of all known cancer cell interactions within a micro-environment. Observation of an explorative simulation might correlate with real-world data such as growth rates, tumour shape or mutation profiles. High correlation between real and simulated results supports the hypothesised mechanism. If we observe divergence between real-world and an explorative model it needs to be qualified and quantified and can then hint at missing causal relationships or actors, and therefore avenues for more in depth research.

All models have a predictive element, a simulation being the predicted result of an input rule set. **Predictive** models however expect a system to be sufficiently understood to provide value in the applicability of their results to real-world circumstances. An accurate predictive model might be able to predict patient specific tumour progression or likely metastasis sites. Model accuracy improves application in almost all cases but can be less important based upon the objective. However, a specific predictive models' value is often directly dependant on its accuracy, requiring a high degree of predictive validity. For example, surgery can be a very invasive and dangerous intervention but is often essential to diagnosis or treatment. We may use a model to attempt to replace a diagnosis element or inform the surgical necessity. If the model fails the human cost is extreme, therefore, a model in this case would require accuracy and validity levels that almost never give false positives.

### 1.3.3 Common issues

The identification of patterns in real-world, *in vitro* or model derived data is intended to represent and explain underlying system interactions. These patterns are often created by or directly represent the bias of population members within a system. Artificial bias can be introduced as a way of highlighting certain interactions or accidentally, accidental artificial bias leading to misunderstood data and inaccurate assumptions. Therefore, unrecognised artificial bias poses a significant problem at all levels of pattern comparison and validation of model representations [64, 11]. In a simplistic case, the aim of a model to represent a causal system can be compromised by biased selection of dependency. By declaring a dependency irrelevant, an artificial bias for the remaining relationships is introduced. We need to remain aware that observed

bias can be driven by both conscious and unconscious design decisions as well as real movement patterns. It can become very difficult to differentiate realistic model definition and preferred definition; favouring a model that artificially shows us what we wish to see rather than being representative and resulting in our observed outcome.

Artificial bias can obscure comparative results and experimental validity when conveying data patterns between different groups of people or parts of a pipeline. We need to identify salient patterns that are also real, particularly where automation may have continued unsupervised and introduced such issues. The real-world data acquisition step can itself introduce bias in a near infinite number of ways, population sizes, pre-selection or even sample availability. Larger data sets help reduce the effects of variance but collection methods themselves can introduce bias, affecting an entire data set. Even physical phenomena like light penetration depth can create bias issues for sensitive equipment.

Often, by the trajectory interpretation stage, artificial bias has already been introduced. Therefore, we need to collect information where available and support it with hypotheses or similarly reported patterns elsewhere in the literature. However, a broad and effective analysis suite can help pattern integrity even after data acquisition. With multiple overlapping patterns it becomes less likely that a bias is unregistered even if caused by input data. We can maintain awareness, employ a wide range of metrics to spot artificially biased patterns and feed issues back to the data origin group where possible.

#### Over and under-fitting

One of the major risks of model definition is over-fitting [73]. When developing a model to reflect a set of real-world data, representation can become so close as to preclude the possibility of variance. An over-fitted model would be so close to the input data that it is not applicable to other similar data sources or able to accurately make predictions. Under-fitting is the opposite circumstance, a model is under-fit if it doesn't represent a comparable data set adequately. When applied to reproducibility a model might be under-fit if only a single high variance operation produces comparable results. A model is likely over-fit if every repeated output creates exactly the same patterns despite high variance.

Over and under-fitting can be caused by input derived bias or a lack thereof. Too strong a bias might skew model definition and over-fit it to our input set, without a bias we may find no comparable pattern for representation. In the context of artificial bias we may even wish to add a way of highlighting patterns for fitting or be aware of the possibility of accidentally over representing one. Appropriate parameter and metric selection in combination with variance parameters are also important at the definition stage to help handle this. Stochasticity helps define variance tolerance levels when combined with repeatable operation. No single iteration of a model should produce results identical to real runs but rather reproduce our target pattern despite high parameter variance. These are inherent issues in modelling and should be designed around, again, where possible we should also design to identify when it occurs. Another reason for a wide range of metric comparative points.

Over-fitting can often be a problem of limited metric or real data availability and under-fitting unaccounted for variables. By creating a model framework that allows for streamlined addition of more data and comparative metrics, both issues can be addressed. Validation can then be used via automatic comparison across many data sources with stochastic model origin providing enough variance to highlight trends for extrapolation.

#### Validation

The definition of good results in reasonable time is open to interpretation. In attempting to generate informative results we apply a requirement for validation. If not validated, comparisons, models and analysis are purely putative, possibly still providing value as conceptual tests, hypothesis generation or explorative investigation but to a lesser degree. In other words, we need to be able to show a reasonable similarity of patterns at several levels to those produced in the real-world. The level of validation importance also varies between explorative, explanative and predictive models. Since we want to be able to generate results of all three types, we need to employ a powerful and variable validation approach.

Validation of a model can be highly complex [11, 64, 74, 75, 76]. Validation complexity comes

from the comparison and evaluation of importance for various inexact metrics, the breadth of potential factors and the difficulty of definition. The possibility of eventual impact upon real-world data gathering or patient treatment dictates a greater need for accuracy. Equally the quality of any insight gained in the exercise depends upon its integrity, also important in use as support for further investigation. Proving that the model is accurate and therefore valuable as an explanative, explorative or predictive tool can be eased by design decisions. Creating comparisons between model results and real-world observations should be a focus.

Pattern Oriented Modelling (POM) combines all available data analysis angles by definition: it takes a multi-level pattern-oriented view. We combine quantitative and qualitative measurements, population size and movement metrics along with visual representations of positions over time. POM as discussed by Grimm et al [11] describes a pattern-based approach to validation and model construction. The central assumption being that a validation framework can be built around a model by extracting and comparing a wide variety of data patterns from different parts of a model. An example given in the paper refers to forest growth, taking patterns from different levels it suggests comparison of tree top cloud level model representations, root level growth and the results of ground level tree fall with real-world data. It is suggested that developing a model with this in mind and validating with similar comparison should lead to more complete and robust representation [64].

Ideally, a higher number of possible comparison points at multiple levels between model output and real-world data improves comparison and validity. However, analysis of all possible metrics can become resource intensive and not always informative. In the case of a singular model this becomes a prohibitive concern for personal and computational resources. We can use a wide range of metrics to allow for previously unforeseen correlation and analysis. While ordinarily very costly, development overhead reduces significantly if placed within a multi model framework.

Validation is important and impactful even at the point of definition. Defining a problem that can't be validated can lead to changing results to fit the evaluation style and introducing artificial bias. We can address the issue by awareness of bias introduction and maintained flexibility to address concerns by using multiple levels of pattern comparison.

Model validation can also serve as a backstop measure for identifying and controlling artificial bias. If real-world and model output don't align we may initially assume the model needs adjustment, however if the problem persists then bias may have been introduced accidentally by design decisions. It is important to note that validation metrics are also a source of artificial bias introduction. Bias due to limited possible validation patterns or the judgement selection of 'appropriate' validation.

**Definition**

Any decision in the creation and implementation of a model can lead to introduction of artificial bias. However, no model is perfect, in attempting to create a useful representation, incidental artificial bias should be expected but minimised when possible and reasonable. Essentially, we can re-shape the discussion around the idea of patterns. We want patterns that are representative and mean something that can be useful in understanding the real-world. The more patterns we have the more we can compare and understand, we can increase pattern availability via analysis tools or model scope.

A lack of direct singular purpose in a more holistic explorative model can mitigate some effect of the bias introduced by preferred outcomes: we introduce parameters and observe results without a purpose beyond that observation, but it is not a panacea [77, 78]. In some cases, artificial bias should even be purposefully introduced: safety critical predictive models may introduce an artificial bias to ensure that if a model is unsure of outcome it will always predict negatively to protect the system from failure. If we accept that some bias will always be present, examining each part of the process for inherent bias can become a scalable issue. By developing tools to address it quickly and effectively we can use shared libraries to offset some of the resource cost by repeated and automatic application.

### 1.3.4 Modelling approaches

The field of available computational modelling techniques is extremely diverse; each unique problem has a unique optimal search solution or approach. However, practically, approaches tend to fall into categories and groups with problem specific changes. We can narrow our selection by focusing mainly on approaches to spatial representation and individual entity definition, population wide movement problems. The plurality and breadth of possible approaches can make comparison of results difficult. To compare results with other similar studies we often need to apply similar output approaches or at least be aware of the effects of our design choices. Each approach also tends to employ different levels of representative specificity, for example we could treat a cell as an individual or a sub system with thousands of internal relationships. We want to be able to encapsulate all the important interactions in our chosen system but the more complex and broad a definition we choose the greater the cost in computational overhead, development time and implicit conceptual compatibility.

Since approach diversity juxtaposes with the definition of loosely defined central archetypes, few models stick rigorously to any single definition. There is also the option of hybridisation, integrating design choices from multiple approaches to best effect, mediating computational overhead, conceptualisation and accurate representation. It is however important to note that different model approaches can represent the same phenomena with differing approximations. Cellular Automata (CA) tend to use a grid based lattice where each grid space is an entity and interaction is derived from overarching rules and adjacency interaction. Agent Based Models (ABM) tend to have free movement and entities behave according to internal individual logic. So, for example, a CA usually has lower computational overhead but difficulty with individualistic representation making it more appropriate for larger numbers but less complex interacting cells. An ABM can be implemented to more closely investigate the smaller scale spatial relationship between individuals.

#### Cellular Automata (CA)

Masoudi-Nejad et al [14] define CA as characterised by a group of 'cells'(referred to herein as lattice sites) with a finite number of states updating simultaneously in accordance with their local environment (Figure 1.1). Each is a product of its and the surrounding cells states at the

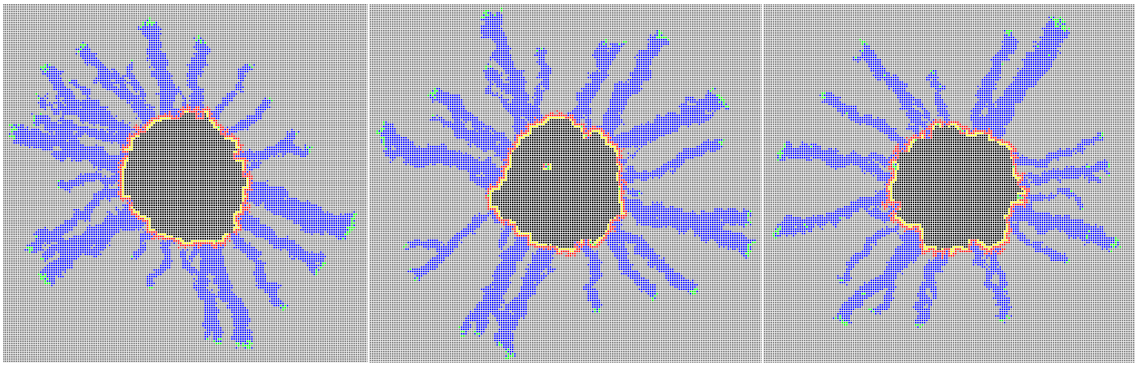


Figure 1.1: [CA Image taken from previous personal work on tumour morphology]A famous fundamental example of a CA is Conway's game of life, it takes an extremely simple rule set and uses it to create a deterministic but constantly changing simulation state from an initial pattern input [79].

previous time step in accordance with the general expression:

$$s_i(t+1) = \Phi(s_j(t); j \in N_i) \tag{1.1}$$

where $s_i$ is the state of a cell at time $t$, $N_i$ its neighbourhood and $j$ the cells in that neighbourhood.

CA modelling generally utilises grids of lattice sites, each with a specific state and behaviours dictated therein. For each cycle, lattice sites interact with other local sites in accordance with their behaviours and develop in a synchronous step by step manner [80, 81, 82]. The advantages of such an approach are both speed of development and an overall computationally low overhead [14]. A CA can also utilise more representative grid morphology with techniques such as Voroni tessellation [80, 83].

### Agent Based Modelling (ABM)

Zhang et al [30] defines AB modelling as a '*computational technique that simulates the (inter) actions of autonomous individuals ('agents') within a complex system*'(Figure 1.2). Within this simulation each agent contains a set of rules dictating its actions and allowing it to make decisions. While the ABM approach has been successfully and frequently employed at a cancer cell level [84, 85, 31, 86], specific individualistic cancer cell movement ABM's are still difficult to find. To our current knowledge, there are no GPCR and G protein movement ABMs.

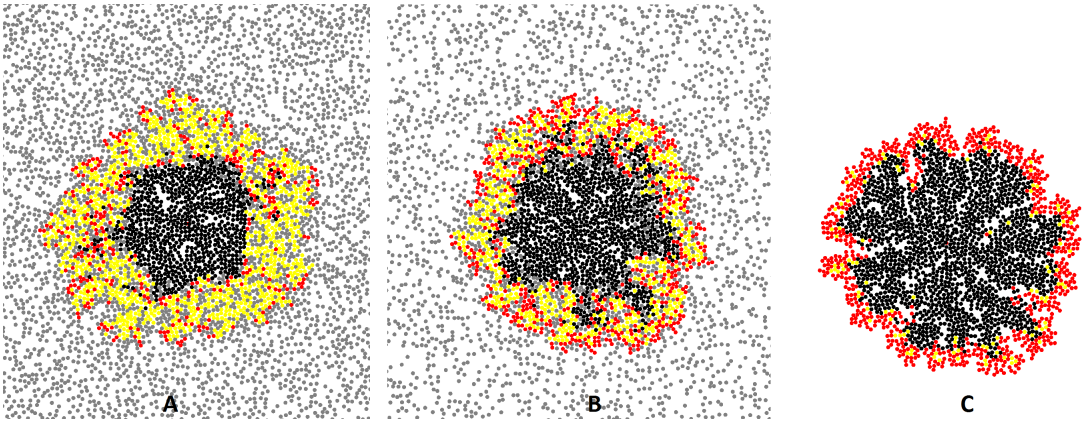



Figure 1.2: Three images (left to right: high, medium and low density) showing the results of an ABM designed to look into the effects of non uniform force application at vary densities of blocking cells. The addition of force allows cancerous cells to push the grey neutral cells out of the way or into blocking formations.[Adapted from previous personal work on tumour morphology]

An ABM represents agents as entities within a free moving simulation with behaviours being dictated by simple internal rules and their observable local environment [30, 65, 16]. The ABM approach therefore provides both a greater granularity in spatial interactions and entity representation. Entities can occupy infinite unique points in space within their defined boundary and their definition is more individual.

### Robotic modelling

A robotic model while not common would enable several unique opportunities and capabilities. Rubenstein et al [87] have developed a swarm of miniature robots called 'Kilobots'(Figure 1.3). Their goals were that such robots should be low cost and easy to program. Work done thus far by the group has taken advantage of the 1024 kilobot swarm to investigate self-assembly and movement of large swarms [88]. By modelling approximations of cancerous cellular behaviour upon such robots' similar swarm behaviour might be analysed. Additionally, the robotic environment might allow easy inclusion of secondary factors that are harder to approximate *in silico*. For example, infra-red light can be used to communicate with the bots, this might be used to create light gradients approximating chemotoaxis conditions. High variance in locomotive and sensing abilities would also represent the heterogeneity of a cancerous micro-environment. Similarly, there is the possibility of uncontrolled secondary interactions, noise from physical environmental

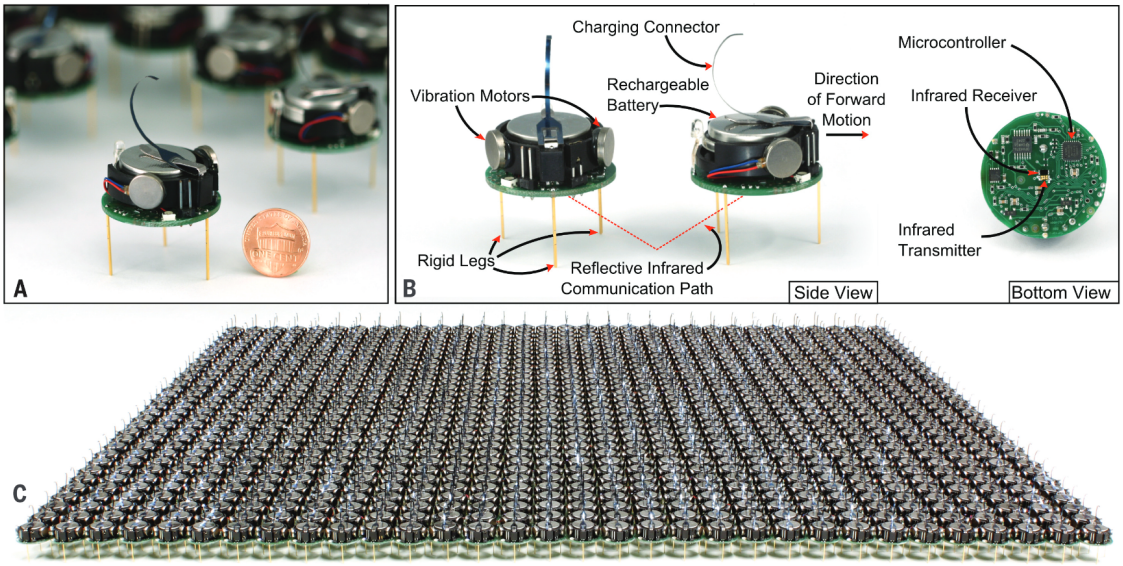



Figure 1.3: Swarm Kilobots from the work of Rubenstein et al [87] image A: Kilobot scale, B: basic capabilities and C: a 1024 bot swarm large enough for a real-world robotic model

conditions brings the approach one step closer to reality. Even immobilising circumstances like robot break down might represent sudden cell malfunction and swarm reaction. The nature of physical robotic representation would further aid in conceptualisation and allow dynamic environmental change, reaction and control.

## 1.4 Interpretation and implementation

### 1.4.1 Computer interpretation

In order to apply a computational approach to biologically derived data, we need to translate observed data into a computationally interpretable representation, i.e. a digitised format. Since we are analysing population movement and interaction, our initial input is composed of videos taken from an *in vitro* environment. Tracking of individual entities enables the identification and extrapolation of data from video input to population positions over time; into computationally understandable information [89, 62].

Many conceptual problem areas involve the interaction of large numbers of distinct individuals [90, 91, 92]. Biological micro-environments are often a product of both interaction and

environmental pressure. Over time, systems can be observed to create patterns of motion such as group alignment [71], following [92, 93] or preferential least resistance following [42, 90, 71]. Movement patterns on both an individual and population scale can therefore be analysed as a product of emergent circumstance. Movement patterns are often the result of adjacency communication as in flocks or schools, thus, structure can represent an overarching objective or reaction to environmental conditions. Interestingly, population alignment can also be achieved without thought or biological instinct, so one should be careful when describing population drivers, in some cases it only provides information of environmental pressures [91].

### 1.4.2 Implementation

To enable computer interpretation we have recorded and digitised *in vitro* cancer cell and GPCR and G protein movement, we have also identified computational modelling as an approach for its representation. Comparison between digitised *in vitro* sets and model results is important for validation and iterative hypothesis improvement. When comparing digitised *in vitro* and model derived data: we need to identify important patterns and generate them in a uniform way that doesn't introduce potentially misleading information. We also aim to improve our understanding of these systems. We maximise salient pattern identification opportunities by developing tools that produce a wide range of metrics and data points. With more gathered information we expect to observe more patterns, quantifiable and visual, differentiating between those that are more and less important. However, such a broad approach comes at a cost, gathering and analysis time increases with metric breadth. We can make such an approach practicable by automating large parts of the process within a repeatable framework and with novel methodologies.

We apply several approaches from a field known as natural computation; algorithms are derived from, or applied to, natural contexts, we draw from nature to understand it. In this case we primarily wish to use approaches such as agent-based models[11, 94, 30]. Multi objective optimisation such as genetic algorithms and neural nets may also be applied for pattern comparison at later stages [12, 95]. Modelling represents our core approach to analysis, comparison between modelled patterns and real observations.

There are a wide range of different applicable computer modelling techniques [96, 83, 14], they vary in applicability and effect based upon the chosen system. There can be difficulties in appropriate selection and application [97, 98]. In selecting any specific approach, restrictions, assumptions and generalisations are created. We not only assume that a model can accurately capture the interaction we wish to observe but that it is an effective way to do so.

### 1.4.3 Digitising and representing real data

Computers are very good at processing and communicating via numbers, but the real-world is different. Tracking can be characterized as the transformation of visual information to a numerical digitised context. Once tracked, we can then leverage the compute power to generate quantitative statistics that might be difficult to develop by human observation [62, 74, 75, 98]. Much of the process can also be run automatically, allowing us to address questions that may have previously been impossible or prohibitively time expensive. We also increase metric accuracy substantially, assuming the input is correct. Therefore, input should be checked and assessed for artificial input bias regularly.

As we have discussed, assumption of accurate input and representation can be problematic. Inversely, while computers are inherently very good at handling numbers, people are unfortunately not. We need to account for both computer and human understanding when we present data to each. Computer vision can therefore also be construed as a way for computers to communicate data only they can currently s̈eeẗo people, in our case researchers. The inverse question then becomes whether and how people can communicate useful information back. When developing a model, we use feedback from individual researchers to communicate and specify required patterns where computer vision is currently lacking. A very similar process to the abstraction of concepts to programmatic paradigms into mathematical rules when converting conceptual to computational models.

### 1.4.4 Applied Computer Vision

Here, computer vision refers to computational recognition and understanding of biological information, such as the movement of populations of cancer cells, or GPCR and G proteins over time. Our videos of population members moving over time are tracked to transfer useful information from our real-world videos to a computational context. In turn we can understand the transformed and processed data, derive possible comparable meta patterns and then identify useful patterns with the intent to validate our observations and representative models.

For example, we might seek to track individual cancer cells moving across their topographical micro-environment. Identifying each cell as they move around in a video becomes very difficult so we take a frame by frame approach. Now, presented with consecutive images of the population we can manually identify each member, a time consuming, but more reliable, process. To speed the process, we may train a computer to identify and locate cells in each frame, we now have a unique identifying number and coordinates for each frame of our data video. Given locations of cells at each time point, tracking becomes the process of linking each member's position across frames to create a representation of their probable movement behaviour automatically. We generate a data set that represents how the cells travel over time, allowing us to develop further insights such as speed, movement, and interaction trends. Our tracked data set becomes a computationally understandable representation of the patterns observed in the real-world or *in vitro* biological system.

### 1.4.5 Pattern analysis

In applied computer vision we can state that patterns exist as a form of meta data representation. These patterns can be derived from information that people cannot easily recognise through other methods. To present them, we often use tools such as graphs to show trends in numerical data. Overlaying multiple trends via population member tracks can also generate more understandable grids of movement patterns, representing spatial complexity over time. Meta representations are a further communication layer allowing us to interact and understand computational interpretation of biological systems.

### 1.4.6 Choice of approach

When selecting modelling paradigms, we have focused on appropriate and flexible environmental definitions. While most approaches can be modified to implement interactions in some form, it can be difficult to add further environmental interactions later. The application of any representation in a poorly constructed framework can cause the implementation of other interactions to become more difficult. Many paradigms can be applied in some way to problem areas regarding motion. Hybridisation of approaches enhances the ability to handle heterogeneity and can be developed over time [83, 99, 100, 101, 102, 31, 86]. Here we can effectively apply an ABM approach by addressing the practicalities of development and maintaining the principle of imperfection but usability in model definition.

While all computer models are essentially logical or mathematical, we need to separate them on a conceptual level for comparison. Applied mathematical models such as those based upon ordinary differential equations often have difficulty with human conceptualisation, local geometric representation and specificity; the more specific a model gets, the more complex the required network of interdependent equations becomes. Applied mathematical rules can however be scaled up or down for differing levels of generalisability very easily. Less computational overhead is also required for tasks designed to help people understand the process or results. Results can be harder to interpret but more appropriate for quantitative and large-scale validation.

A CA approach encapsulates relationships within a rigid grid but allows us to view each lattice site as an individual. With CAs we can conceptualise or compare with a real-world observations but are limited in interaction definition and spatial complexity. Grids are not required to be square and approaches such as Voroni tessellation can create more cell-cell like environmental structures [81, 82, 80]. However, CA is still essentially limited and imposes spatial interaction definition.

An ABM approach allows more defined spatial complexity and interaction; compartmentalisation of entities makes infinite sub population definition more conceptually approachable [44, 103, 30, 16, 84, 85]. Also, visualisation of ABM models can often be more directly comparable with reality without discretised positional definitions effecting extracted patterns. By bringing

the way we view, understand, and interpret an approach closer to real-world observation, we can spend computational resources to empower our communication with computational interpretation.

Within this context a physical robotic model represents a further step towards real-world representation. We spend more computational overhead but gain closer comparative and model definition tools. In this case the additional development and computational cost was greater than the perceived benefit beyond an ABM approach.

So, mathematical models can be faster to run and develop but harder to understand conceptually and thus compare accurately with real-world results. CA is more computationally intensive and requires greater development but is more visually and conceptually comparable. ABM then brings visualisation and representation even closer to real-world observation. A primarily ABM approach eases the communication of information into a computational state but also transfer back into a form that we can interpret. Our implementation also hybridises with CA for entity tracking and collision detection with internal entity decisions dictated by logical or more purely mathematical rules. While we have chosen an ABM archetype few effective computational implementations are so clearly defined.

## 1.5 Framework level design

### 1.5.1 A framework approach

A framework level approach with generalised holistic definition can enable representative definition and imputation. If properly defined complex relationships can be represented in a way that prioritises important interactions with minimal computational cost. The process of finding such strong effects is the process of identification and definition for *hyperparameters*. We can also augment and automate aspects of system observation and comparison between both; an overarching design that unifies visualisation and pattern extraction.

Many of the common difficulties found in computational modelling pertain to model definition [13, 12]. The choice of scope might be seen as a choice to exclude less important interactions either due to practical computational constraints or development complexity. Unfortunately, such choices can become an issue as some patterns are impossible to replicate without the added complexity [104, 71, 42, 105]. Ultimately, we may have a good understanding of a phenomena but as systems become more complex, the introduction or removal of interactions can have effects beyond our expected scope. By implementing a holistic generalised approach to the problem area with a strong modular tool set, we can allow for both narrow and broad models within a single overarching framework.

Both explanation and prediction of real-world phenomena and interactions requires the validation of models, supporting evidence that the patterns and interactions therein are representative enough to be relevant [106, 97, 98]. We can test its validity by broadening comparable patterns to identify previously obscured interactions [11, 64]. Quantifiable comparative measures are produced by our framework across all models generated and real observed data input. Unification within a single digital structure removes most comparison barriers and enables more reliable validation (Figure 1.4).

In digitising real data the addition of a bias that isn't representative can be a significant issue [107, 108, 109, 110]. If bias is introduced at the observation or tracking stage, consequent models or meta data generation will not be truly representative beyond the input technique, an unlikely subject of investigation. Artificial bias can become a significant problem at all levels of pattern comparison. Particularly affecting the validity of a final model, biased inference can lead to inherently compromised and unrepresentative results.

A robust framework level approach to model generation allows us to analyse a system with models of differing scope. By centralising tool generation we can apply a wide range of pattern analysis and extraction approaches to augment interpretation. Fundamentally, the more patterns we can extract and compare, the easier it becomes to identify bias issues and important interactions.

A sometimes-overlooked aspect of applying computational analysis is development time. Time spent developing the tools and libraries necessary for visualisation, interpretation and modelling of data is not spent upon interpretation of results. Similarly, if initial analysis can be accomplished quickly, we reduce time between rounds of investigation, experimental redesign and application. Therefore, throughout we attempt to identify areas where computational and developmental cost can be reduced with the assumption of a correlated increase in time for analysis, interpretation and experimental application. The developed framework in turn seeks to reduce model development time while also allowing effective comparative study and overall bias reduction with iteratively driven cyclical improvement.

Within our scope of practical concerns we can refer to both software development and model definition, each affects the required resource allocation of a chosen approach. A framework level design in software terms means encapsulation of many available tools in one overarching package (Figure 1.4). The more complex a framework is, the more options are available for investigation and definition. However, larger design scope leads to greater development time and can cause combinatorial issues. For example, if parts of the framework are poorly separated, a new parameter in one part may need to be accounted for in another and increase computational cost overall, even when not using those specific modules. Over time, in an expansive framework such addition can cause significant performance reduction. The slower a framework can run the fewer model iterations become practicable; a software level issue with an experimental consequence. A similar incrementally increasing process can occur with error or bias introduction; it is very important to maintain proper module separation for large projects.

Model complexity can either be limited by framework scope or dictate further expansion, if we don't have a tool to define entities replicating, we either need to add one or remove replication from the model. Even with effective modular software design models with large numbers of implemented hyperparameters will have a high computational overhead per iteration and population member. For that reason, one philosophy for model definition is to restrict it to the minimum possible interactions for target phenomena to occur. Minimum interactions leading to minimal software complexity, reducing accidental bug or bias introduction. We want to take a more holistic approach; our models are likely to consist of many hyperparameters. Therefore, a corresponding overarching framework is required. Proper planning, project conceptualisation and coding practices are all practical issues that can be addressed by restrained appropriate

problem definition.

### 1.5.2 Model and system interpretation

In our case both GPCR and cancer cell data can be digitised and modelled for interpretation as groups of population members moving around an environment. We have identified that a multi scale modelling approach would be advantageous where simple minimum representation models informs more complex examples for pattern replication and explanation. Given that we aim to address multiple distinct systems with a holistic generalised approach, an overarching framework for our tool set seemed reasonable. There are also secondary considerations and capabilities that an overarching design makes relevant and possible: handling multiple data sources, wide scale reproducibility of model results and artificial intelligence (AI) for automation along with introduction of more abstract validation approaches to name a few.

### 1.5.3 Centralised digitisation

Our framework includes tools to digitise tracking information from a variety of sources into a singular comparable format. A digitised model run can then be treated and handled by the same tools as a model to generate comparable metrics without the effects of differing representation creating artificial bias. Similarly, tools can record model or real data subsets that can also be passed to the digitisation end. Such input and output generalisation gives us a great deal of flexibility in the way we break down and observe data, real or model generated.

### 1.5.4 Reproducibility and robustness

When creating a large-scale modelling and comparison framework, robustness and reproducibility become an important factor [73]. Stochastic elements can create significant output differences in each model run [111]. Reproducibility refers to the ability of our framework to generate models with recognisably similar output but allowing for some level of representative stochastic elements. Stochastic reproducibility allows identification of an introduced variable as leading to emergent behaviour and patterns rather than as an outlying combination circumstance. In other words, it

allows us to assess and reduce over-fitting. We want to produce patterns like those observed in input data but not exactly and across each run with added stochasticity.

### 1.5.5 Representing real-world noise

Stochastic elements are also needed to represent unknowable prior environmental circumstances or unimplemented lower level relationships and interactions. Prior to observation real-world systems often exist as the sum of far-reaching previous interactions; these interactions are often unknown and need approximation. Stochasticity as a way of generating a base environment to overlay effectors upon can therefore help represent the variance of real-world conditions. When combined with reproducibility across many runs we can create a stochastic and varied basis for introduction and observation of hyperparameter effects. We can also apply different stochastic noise distributions beyond random such as Gaussian distribution to investigate different effects upon parameters such as population starting position.

At most scales of life-based motion, environmental heterogeneity is a powerful and oft repeating factor [112, 113, 80, 42, 71]. Part of that is also population heterogeneity: sub population groups and smaller individualistic variance can be identified [113, 80, 40]. Individual refinement of model populations can quickly become inefficient with large groups of entities. Therefore, stochastic state assignment for everything from positional preference to movement rate and directional choice is common in such representation. Again, reproducibility allows us to safely leverage stochastic variance, identify bias and drive automated development without over or under-fitting.

#### Handling stochasticity

Models with widely applied stochastic elements run the risk of conscious and unconscious bias, selecting only the runs that look like the input data for reporting. Unfortunately, this is difficult to remove but when required our robust tool set creates log files for all generated model runs and results at requested time points. In a similar vein all model result files carry the randomness seed to regenerate and observe the entire run. Nothing on a computer can be truly random, a

randomiser generates a seemingly random sequence of outputs from a starting position based upon obscure and unique starting conditions. A run seed represents the starting position of a randomiser, with the seed we can identically reproduce all stochastic choices in a model run. We attempt to assuage common risks but in a way that doesn't add complexity to the model design and iteration process where possible.

Results need to be reproducible and differentiated from noise. Reproducibility can allow for rapid testing of possible over-fit and under-fit hypotheses. Automatic large-scale generation with meta data identification across many model input parameters to identify and represent target patterns also becomes a reasonable possibility. The scalable and reproducible application of comparative optimisation algorithms allows limited automation of model definition, a powerful tool for exploratory analysis.

### 1.5.6 Automation

Automation in the context of an overarching framework can have several meanings. In the case of model definition, we can refer to the automation of applicable representative models, selection of input parameters, automation of validation and iteration. Alternatively, it can mean automation of comparison. Full automation of interpretive processes such as defining the similarity of two sets of results can be very computationally difficult in complex environments. Often there are complex mutually exclusive multi-objective optimisation problems and difficult quantification of visual pattern recognition. Automation can also refer to many less impactful tools automating things such as recording model states, reproducibility and representation output. All can be time prohibitive in the case of small-scale model design and development but become necessary and reasonable across larger framework level development.

The automation of comparison between model and real data assumes that we can convert behaviour into a measure the computer can interpret. Information such as population size or distance travelled is available after tracking. We can assess and use these numerical measures in automatic trend fitting at different time points, quantitatively comparing simulated to real environments. We can then use that similarity to report back to a user or inform model design

decisions automatically. It should be noted that any unsupervised approach is vulnerable to over and under fitting and care should be taken, to minimise fitting issues, automation can be kept to initial investigation or made semi-supervised. However, it can be difficult to quantify visual similarity of patterns, if we can achieve that then our fitting automation can become much more effective. During the PhD thesis, we attempted both crowd-sourced and neural net similarity measures, the latter was best applied and is later reported.

#### Neural nets

Neural nets are an example of existing artificial intelligence approach that can be utilised to obtain repeatable and quantitative similarity comparisons from meta images. Either by classifying different sub behaviours, or directly comparing overall image similarity, neural networks are a powerful tool for extracting comparative points [114, 115]. Artificial neural nets are designed after our understanding of biological neural nets and can be trained to similarly interpret visual information. We give a net model generated image patterns and train it to recognise them, classifying similarity, a trained net is then given the real-world generated images and asked for a similarity measure. While a completely unobserved process might be daunting, some application of these automation approaches can help us generate initially representative models. Additional metrics also add to the layers of patterns for our overarching POM approach.

## 1.6 Aims

With this work, we aim to further understand complex biological systems with entity movement over time through the development of a novel framework. We focused on identifying patterns in cancer cell and GPCR and G protein movement. By doing so we aim to gain knowledge about the underlying interactions that define their behavioural trends over time. We expect populations to be partially defined by their environmental conditions, and in turn impact it. Representative modelling and digitisation of real-world *in vitro* data videos within an overarching framework will enable analysis. A pattern-oriented modelling approach will be combined with agent-based representation to enable effective model definition at multiple levels, simple targeted and explorative holistic representations. Further, we will eventually include novel methods of model generated and *in vitro* data comparison to generate greater observation depth. Our framework

should demonstrate that it is possible to primarily inform further understanding of biological systems, but also the process of method application within a computational modelling context.

Therefore, we can summarize our aims as follow:

1. Identify, design and develop our modelling framework for biological systems, first focusing on cancer at a cell level with individual free movement (Chapter 2)
2. Being able to examine the spatial relationships between GPCR and G protein to understand how their environment affects patterns of movement and behaviour (Chapter 3)
3. Improve our framework to enable the characterisation of directional trends within micro-environmental patterns, their construction and morphology over time and apply to both systems(Chapter 4)
4. Finally, use artificial neural networks to develop novel workflow; classifying identifiable visual patterns for model/system comparison and filtering populations into sub sets on track morphology (Chapter 5)

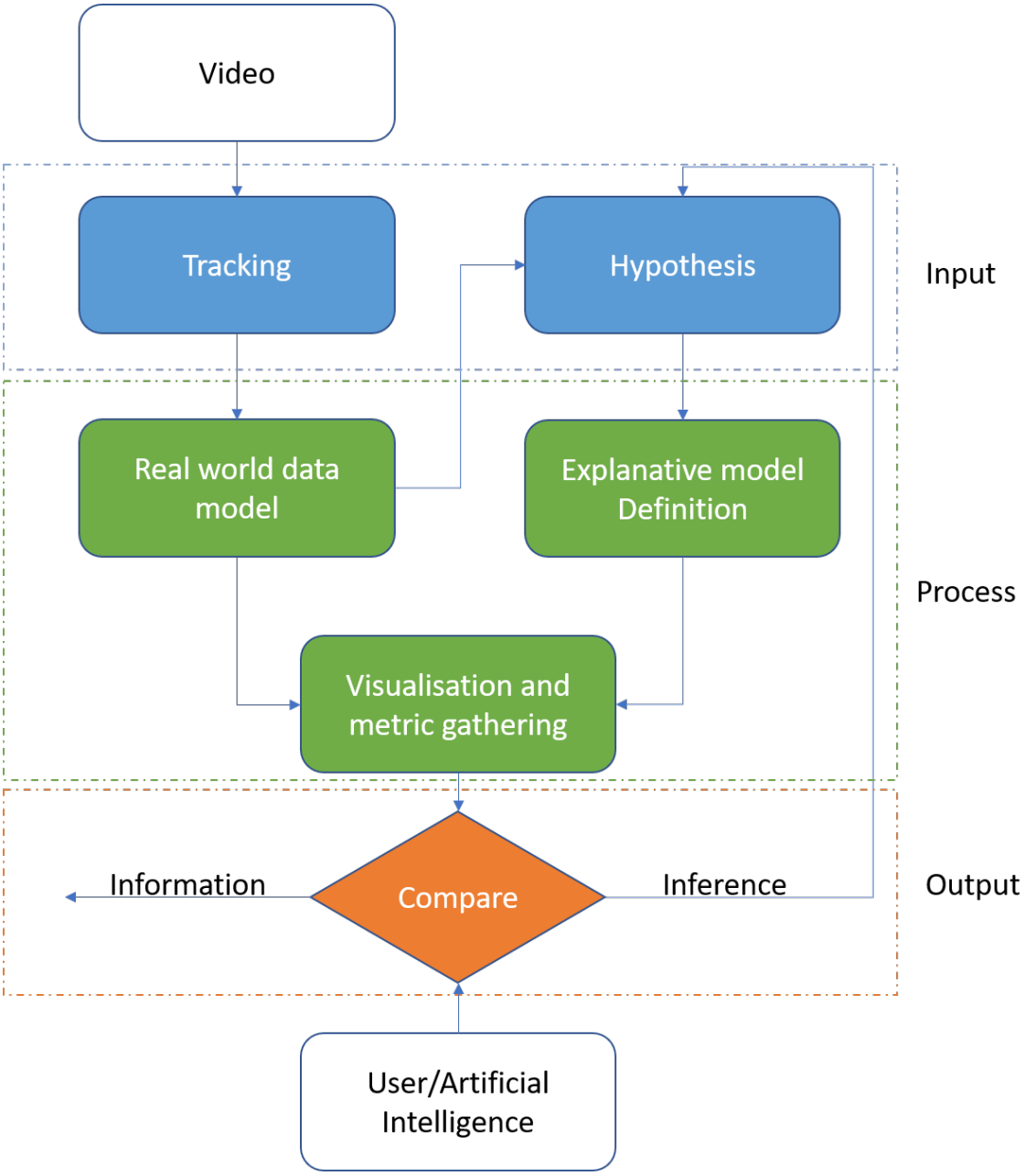



Figure 1.4: The framework level design intends to enable input of tracked real data to a digitised model representation informing hypothetical causal conceptual models for explanative computational model definition. Both explanative and real-world data representative models are processed through visualisation and metric gathering tools for comparison. Finally comparison of both model results leads to further hypothesis development and information generation.

# 2 Cancer cell movement - Design and first application

## 2.1 Introduction

Cancer is a cell level disease, however, recent work has tended to focus on a molecular level or body wide scales. We aim here to design and develop a modelling framework for cancer at a cell level with individual free movement.

We wish to understand cancerous cell behaviour via observation without direct intervention, generating new knowledge of the driving processes. To understand the interactions that comprise such a complex biological system, we decided to focus upon movement over time, developing representative computer models for comparison within an overarching framework. We aim to generate insight by understanding and identifying possible causal interactions for population behaviour. Our hypothesis is that given *in vitro* biological data in the form of videos, we can extract entity positions over time and analyse movement to identify salient patterns for comparison. We identified a framework level approach to Agent Based Model (ABM) development with holistic visualisation and comparison tools that would be effective.

### 2.1.1 Summary

Herein, after development and application of the framework, we identified strand like patterns indicating cohesive cancer cell movement over time. Initial analysis also identified a trend of directed movement across populations. Representative models were developed to assess the relationship between directed movement and the strand patterns, then the possible existence of least resistance lattice following or path forging behaviours. We found a hybrid directed forging

movement with overlaid attractive lattice created compelling resultant movement patterns over time.

### 2.1.2 The cancer system

When observing cell level cancerous micro-environments, we need to remain aware of behavioural cancer cell heterogeneity [34, 35, 36, 37, 38]. Such breadth of possible behaviours necessitates a generalizable approach.

Cancerous tumours emerge from malfunctions in the process of cell replacement and the resultant life cycle of an affected cell [39]. As replication continues these cells are influenced by, and change, their local micro-environment such as through localised structural degradation or nutrient competition. Over time, changes can become detrimental on a larger scale leading to a variety of negative symptoms and eventually becoming fatal. Heterogeneity of cancerous cells from clonal evolution occurs across different tissue environments and cancer types [116, 117, 118]. Metastasis and cell motility play a major role in progression towards lethality or improved long-term treatment [44, 71]. Therefore, understanding the drivers and mechanisms that develop a cancerous micro-environment is important for successful treatment [42, 80].

Other cell level cancer ABMs still tend to focus on internal tumour arrangement, heterogeneity and cell to cell adhesion mechanics [119]. The relative lack of recent cancer cell movement models means that we need to take an abstract approach to pattern identification and attribution.

#### Expected movement patterns

Identifying and developing a causal understanding the different mechanisms and theories behind metastasis has been an important concern [120]. For example, Geiger et al [121] suggests a formalised multi-step ‘ *metastatic cascade*’. One common factor of importance among them is often cell motility. Seen as the process of chemotaxis, cells reacting to external chemical gradients [45], such movement can even occur on larger scales as groups [122]. Palmer et al [123] explored existing theories and chemical relationships to suggest that understanding and subsequent modification of cancer cell motility patterns might have important therapeutic impact. Application

of a motion focused framework for modelling and investigation should therefore be relevant.

Directed or random cancer cell movement with localised structural degradation can progress to the invasion of surrounding and remote tissue, eventual metastasis [124, 92, 80]. Metastatic sites can differ in behaviour considerably, meaning, despite being able to trace the origin or relationships between types we still need to be able to study them separately [125, 126, 127]. This means that once a metastatic stage is reached, the heterogeneous differences dictated by physical location and behaviour require a similarly diverse range of model representations and eventual treatments.

Movement and cell to cell interaction within the micro-environment can lead to larger scale behaviours and patterns such as group alignment or distribution trends discussed by Deisboeck et al [92]. Chang et al [122] refers to interaction with stroma; a supporting framework or matrix around cells. They suggest that since stromal stabilisation is stronger in expanding tumours cells create larger more invasive stromalised clusters. They also found that such interactions made collective migration more effective within noisy environments. Collective motion and interaction are often observed in a similar manner in bacterial colonies [90] or other groups within nature [92] but not exclusively. Simple spatial interactions can give rise to large scale changes. A spatial component is therefore important in modelling cell to cell interaction. They are also fundamentally comparable behaviours that can be approached from within a generalised model development framework.

#### Heterogeneity and behaviour

Heterogeneity can be observed in both cancer cells and their surroundings. For example, in dealing with brain, prostate, breast or lung the environments and applied investigative approaches all differ greatly [128, 129, 130, 48]. Many studies incorporate and investigate the effects of heterogeneous cell variables and environments such as growth rate, movement speed and cellular integrity [92, 131, 80, 117, 30].

Emerging from a single cell, clonal evolution [132] means genetic traits spread amongst a cell

population through clonal genetic retention. However, over time and with the pressures of a hostile environment even this can lead to wide variety of different cell types, even within a single isolated micro-environment. Differentiation leads to a varied heterogeneous landscape with a wide variety of cell-cell interactions and behaviours. The interaction between cancerous and immune cells can drive this evolution, even producing an effective arms race whereby tumour cells secrete immune suppressive factors to effectively camouflage themselves [46]. It has been suggested [71] that cancerous populations grow and behave in a similar way to organisms found in the wider natural world, a comparison which goes from displaying swarm like behaviours [105] to co-operative behaviour [40, 133].

## 2.2 Methodology:

### 2.2.1 The Framework

The framework design centres around two main output types; *digitized* and *representative*. *Digitized* output type comprises of *in vitro* data input via positional overlay with entities and iteration, creating a direct real-world digitised version within the framework. Then, we define *representative* models as a formalisation of a hypothesised systematic interaction (Figure 2.1). Both strands converge at the visual representation and analysis stages for comparison allowing us to observe entity interaction; representation and analysis can be sequential or independent. Beginning with real-world biological data as video, we track members in a population and digitize them within the framework. Subsequent observation then allows us to then develop a hypothesis and digitised representation. Sharing the same visualization and analytic tools eases the comparison and enables uniform application of metrics.

To utilize a Pattern Oriented Modelling (POM) comparative approach as well as iteratively improve both comparison and validation, a bias reduction loop is added to the design (Figure 2.1 Yellow). Crowd sourcing and neural net learning algorithms can be used as well for comparison and improve automation of the bias reduction loop. Comparing the model and digitized *in vitro* meta-data allows correlations to be drawn between similarity measures and effective model iteration (see implementation of artificial neural networks for model comparison in chapter 5). With less complex representative models, initial definition might apply semi-supervised tools for

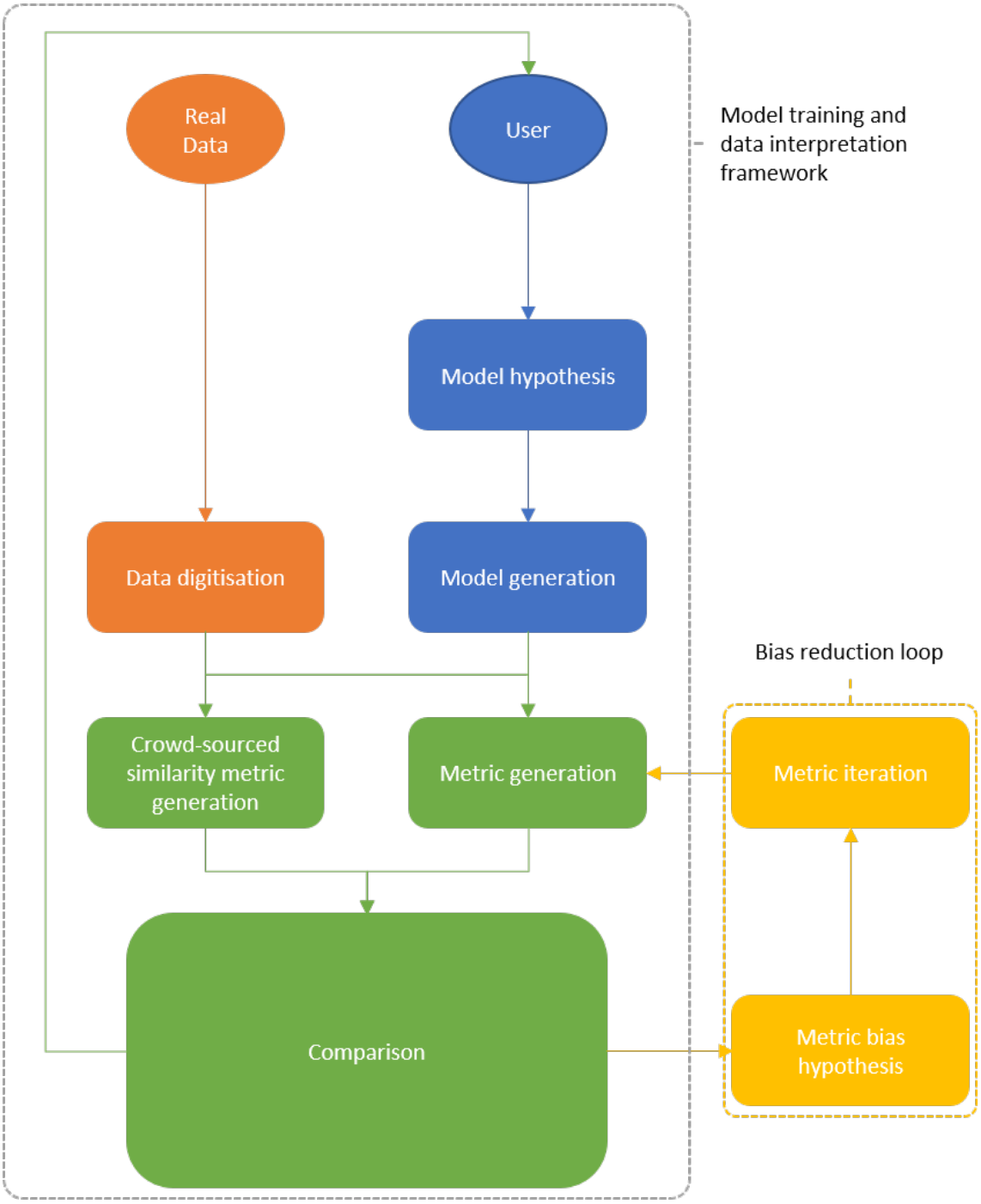



Figure 2.1: *in vitro* data is digitized (Orange), models generated (Blue), and both compared (Green) to inform further incremental model and hypothesis development, a bias reduction loop (Yellow) at the comparison stage can be used to identify appropriate metric significance.

automatic fitting but still require oversight of a user. Trend fitting or other comparison metrics would create a closed loop for trimming and automation. However, a fully unsupervised approach would risk over-fitting, hence it should only represent an initial step.

#### In vitro data input

Population movement within a video is usually processed via step by step entity identification where positions are assigned and then then linked into a set of tracks. A track then consists of known positions and the time index for an entity's movement. Tracks can be complete or partial with gaps in time where tracking lost the target entity. Once identified, tracks then become input for the *in vitro* data digitization.

After inputting track data, positional x,y co-ordinates are extrapolated to starting position, direction and speed at each known time point to smooth over missing positions. Digitised tracks within the framework are visualised and recorded the same way as model entities. The framework does not generate any behaviours or interactions for digitised tracks and only records; any population change or behaviours are entirely from the real-world example. For digitised *in vitro* runs, decisions are made based on the input trajectory data as time increments are iterated. Therefore, comparison of tracks and model population results within the framework allows almost identical data collection and visualisation, differences due to visual artefacts are minimized.

With the chosen approach, we can correct for some pre-framework tracking issues such as missing data. Other attributes such as population size can be difficult to approximate with fractured tracks each creating a new population member. Model with digitized *in vitro* runs is not the only possible comparison we can make. Multiple track files from *in vitro* experiments can also be compared with uniform analysis libraries in the framework for pattern comparison.

#### Model design requirements

Many populations, such as cancerous cells will move around with random directional bias, unless otherwise driven by their environment. Entities can also vary in individual size or shape and

replication rate along with overall population size. We require the ability to replicate many of these population traits when we represent a similar situation within a model.

A primarily Agent Based Model (ABM) approach was selected (1.4.6). However, the framework is also able to generate cellular automaton models for improved performance via reduced spatial complexity in the case of much larger population problems. In an ABM approach each cancerous cell can be represented within a model as a distinct moving entity. Individual movement leading to interactions that can produce similar comparable patterns to the real-world biological system. Effects can be broken down to population and environmentally based variables, an entities preferred behaviour and effects of topographically local phenomena upon them.

By first attempting to approximate neutral behaviour, we create a control population that interacts with environmental effects to generate the consequent emergent behaviours. Results can then be observed to inform further model design. Once established a framework can be used to address different modelling paradigms and data sources. Initial qualitative investigation utilising in a rapid iterative model development can drive explorative development.

#### Generating comparable in vitro and model patterns

Given movement data our framework creates a digitised representation, from which we observe and generate causative hypotheses. We use a hypothesis to define a representative model, comparison with the real-world digitisation then allows us to validate and iterate our hypothesis. When observing movement of a given cancer cell population, we identify where and when individual cells move, input that information to the framework and create meta data patterns that are compared with representative models. *Explorative* models constitute comparable behaviour and can act as an analysis step by identifying hyperparameters; we hypothesize strongly effecting interactions and develop our understanding. *Explanative* models also test our causal assumptions, designing a system with our assumed rules, and observing the predicted outcome to assess applicability.

Sharing the same visualization and analytic libraries eases the comparison between simulated

and real-world data: a pattern observed in one should be generated by movement in the same manner as the other. Quantitative metrics such as turn rate, position and movement distance can also be gathered in a uniform manner and form the basis for meta movement representations. We can approximate population size by observation of active entity numbers at several time points across a run. More complex metrics such as collision and population replication rate are known for model runs but necessarily presumptive for *in vitro* digitised sets. As such we have access to many metrics and information representations with differing levels of reliability for integration in POM validation.

#### Model generation and comparison

It can be difficult to move from a concept to computational model design when selecting appropriate representative hyperparameters; the underlying behaviour and interactions that comprise our model, movement types, population growth dynamics or environmental effects. Individual behaviour and movement over time lead to interactions and relationships within a population. A interactions rarely occur in isolation, each affecting others to create a network of other emergent behaviours. Therefore, patterns can emerge from single or groups of parameter combinations, sometimes the effects of small parameter changes need to be assumed and applied with others to represent emergent co-dependant patterns. We developed a *live-run* tool to incrementally introduce and visualise interactions within a model before progressing to more large-scale application and automatic data generation.

Our *live-run* tool allows us to visualise a population, adjusting individual and environmental parameters to observe their effects (Figure 2.2Yellow). Once an acceptable starting population and environment rule set is decided it can be used as a solidified model definition. We then implement this definition in a loop which will enable the computation of numerous runs without live visualization and its attendant computational overhead. It uses defined parameter ranges to generate multiple stochastic models from a similar state and attempt to avoid over-fitting while retaining hyperparameter effects. Summary visualizations of movement trends and other collected population metrics then enables greater POM validation scope.

Once hyperparameters have been provisionally identified a large-scale stochastic loop is de-

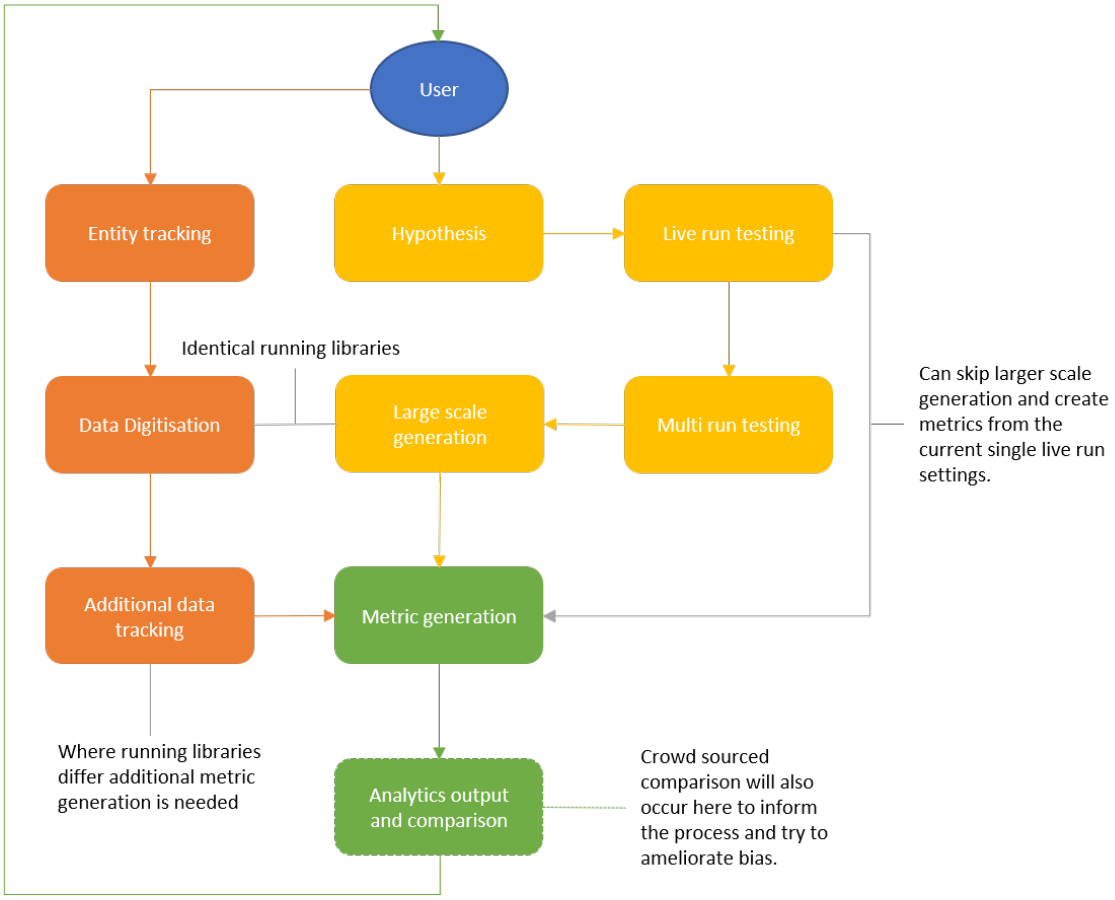



Figure 2.2: Diagram of the iterative process through the modelling libraries for digitized and hypothetical data. For digitized data (left, Orange), movement profiles are transposed over representative entities for digitisation and then metric generation for comparison. A hypothetical model (right, Yellow) is created in a live interactive run and then mass produced to account for an allow stochastic variance. Both can then be compared with the same analytic libraries and the model incremented.

signed to use the same algorithms as live model manipulation but over many runs to account for stochastic bias (Figure 2.3). To be able to repeat previous runs, we generate a 'seed' value that can be used to identically generate all stochastic elements if needed. With a seed a run producing data of interest can be moved from a large-scale set to the live visualisation suite for more direct observation. The framework can also be used for comparison of *in vitro* data sets, multiple input and digitization allowing side by side comparison.

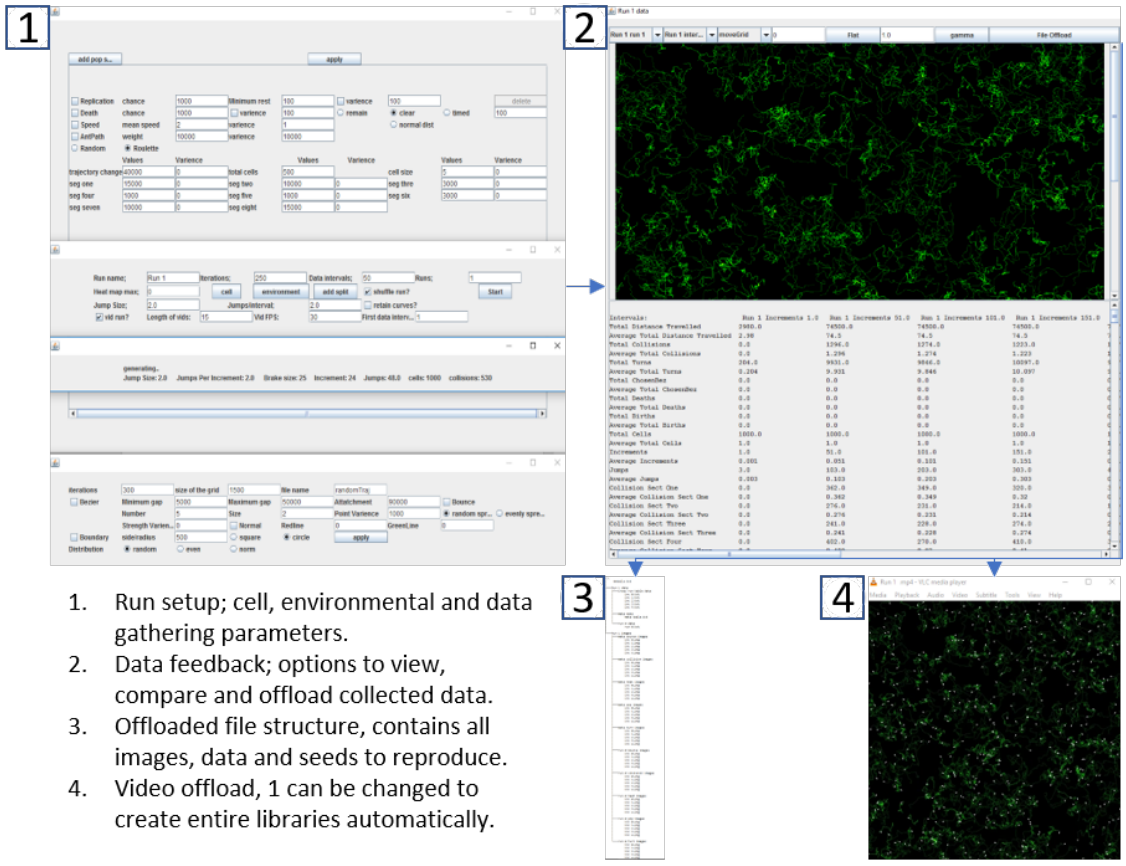



Figure 2.3: The interaction between user and large-scale run starts with parameter definition (1), metric generation and observation (2), metric offload and storage (3) and optional generation of video data (4).

### Comparative metric generation

To compare and analyse both *in vitro* data and explanative model results we need to be able to derive a metric representation. We can use the framework to extract positions over time and create cumulative meta data for general fields such as travel distance or more specific population movement (Table 2.1). We then automatically display this data. Our meta data visualization generally falls into the categories of two-dimensional movement heatmaps, directional diagrams

| | |
|---|---|
| Active cells | In each time phase, between two recording points, the number of active tracks or entities. It is more reliable than total entities over a run but not infallible, fragmented tracks will artificially inflate population metrics. |
| Total distance travelled | This refers to the total distance every entity in a population travels over a model or real data run. Less affected by tracking issues such as fragmentation but representative of population size and activity. |
| Total turns | Every time an entity changes direction the total turn count increments. The value should be tied to population size, variance in this relationship indicates artificial bias or significant entity interaction. |
| Directional preference | The framework also records entity directional choice at each time point. Relative to their previous heading, where do entities turn and how often across a population. |

Table 2.1: Metrics often reported across this work to identify possible representative patterns or for sanity checks to ensure expected relationships and interactions occur.

and auto generating graphs (Figure 2.4). Directional diagrams are segmented into eight or sixteen equal portions, in the case of a turn diagram the width is set for each segment but side length is relative to their occurrence rate. A segment that is longer than its neighbours is therefore representative of a trend in that direction. Directional diagrams can show turn preferences relative to previous direction, cardinal points or amount of distance travelled after a selection.

Cumulative movement heatmaps (2.2.2) are constructed by recording values with a spatial component, an environment is segmented into a grid overlay and every time an entity moves that segment is incremented. As such, a grid overlay can be extracted, the highest value found and a colour gradient constructed from the highest value to zero. Each grid segment is then assigned a colour based on the range: the more movement that occurs in an area, the brighter that area will be in relation to the rest of the grid. The top end value can also be manually defined to ensure a colour tone refers to the same value across runs and comparison. Heatmaps were implemented to represent movement density or population density over time, much like photographic exposure; the longer entities are present, the greater the localised effect it has on heatmap colour.

At the end of any run, wether *in vitro* or representative model, our automatically captured, represented, and generated meta data can be encapsulated and delivered as a framework level comparison tool set. In practical terms we create side by side number tables and visual metric displays from tracked real-world data and comparable models. Images can be zoomed in and out, traversed and modified for top end and gamma transforms. We output images and tables automatically or as a file structure to be read in for future comparative tool use without a new running cycle. Output of model data also includes the states of all variables at data gather time points and starting seeds for full reproduction of any model.

#### Automation

In any investigation we need to prioritise between development time and investigation of the resultant models. Automation requires a higher software development cost but becomes more efficient over time and allows investigative improvement. The more problems we apply our tool sets to the more valuable algorithms that save us time become.

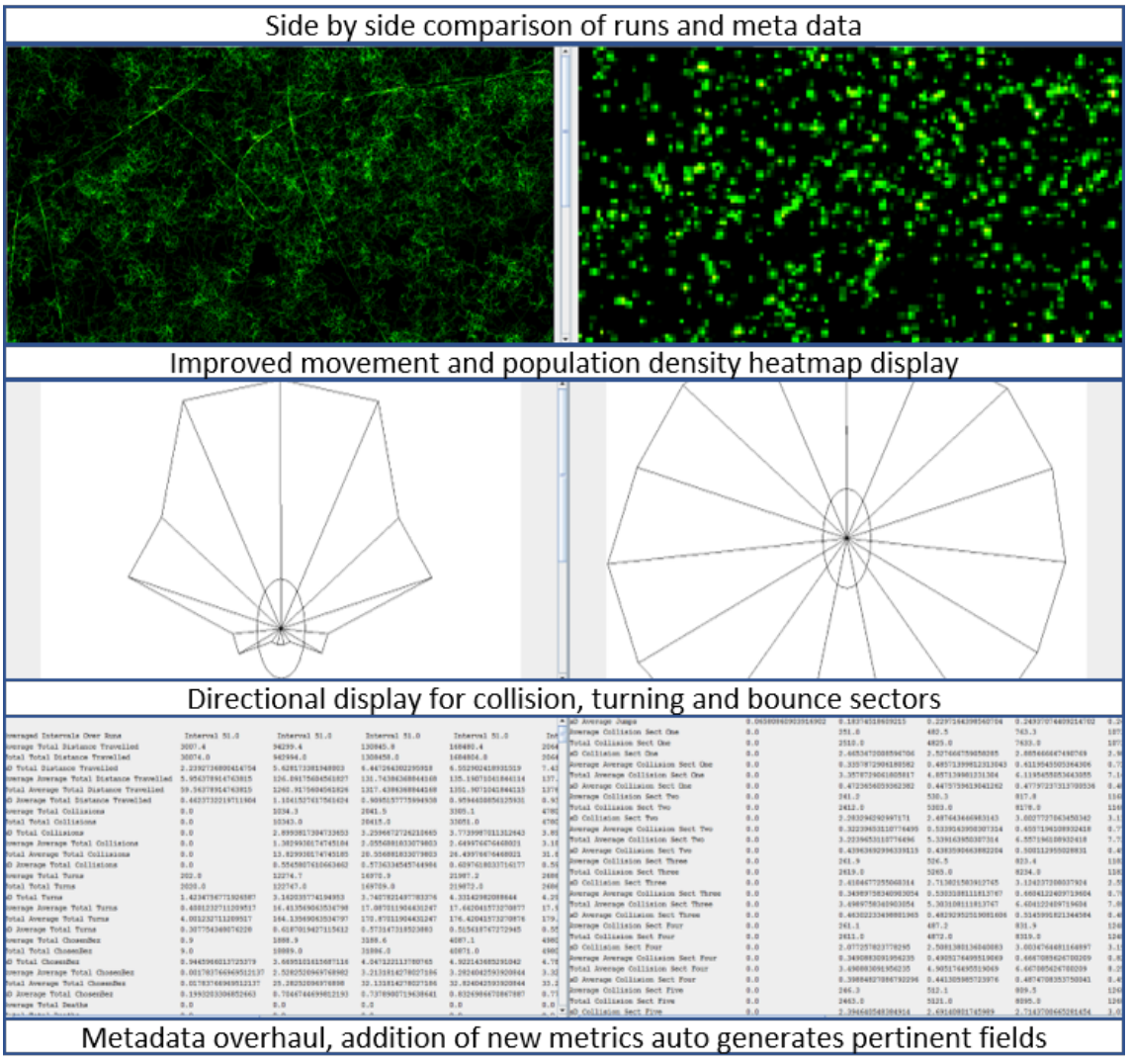



Figure 2.4: Users can select runs and individual snap shots in two side by side data views. These windows can also show turn circles and detailed metadata tables before being offloaded.

Many development pipelines require significant user input. To define a model, we have to define input parameters, ranges, stochastic effects, and apply interaction rules to recreate patterns from our chosen biological system. In simple cases such as trying to make a population grow at a similar rate to the observed cancer population or travel the same average distance, it can become a time expensive case of trial and error. Simple initial fitting is required but can also be acquired automatically.

Automated trend fitting allows us to first feed the framework *in vitro* data and cell positions over time, define modifiable variables and then run automatic comparison and improvement cycles. Our framework can then generate a group of models, with comparable results and similar trends to the real-world observed population.

Automated comparison approaches have been enabled by designing the framework for large numbers of runs and for reproducibility. For example, AI interpretation methods like neural net classifications systems (see Chapter 5) become more effective and accurate with varied and greater numbers of available pattern examples. Framework reproducibility enables scalable training data generation with the possibility of mixed training.

In creating a complex heterogeneous simulation environment capable of generalized model development, the number of possible variables quickly leads to combinatorial explosion [134, 135, 104]. A complex combinatorial landscape also causes difficulties when applying crowd-sourcing approaches: for comparison, we require a much greater number of comparisons to be made; too wide a field with too small a group makes statistical inference more difficult [106, 136]. An opportunity therefore exists to address both issues with a stochastic candidate thinning solution. In this case, an evolutionary optimization algorithm could be used [44, 13]. A very general definition of similarity and bias towards inclusion of outliers would thin the field of possible comparison but preserve diversity for similarity classification. A more rigorous and constrained definition of similarity is then applied after multiple iterations of the bias reduction loop (Figure 2.1). Thus, the process can be further automated and supported with minimal additional development time in the future.

#### Framework implementation

**Development** Within the running cycle of the framework, a modular approach was used to improve process clarity and reduce additive overhead. Whenever a new environmental or population parameter is added to allow approximation within models it's encapsulated within a executable module. Modules are then loaded into model cycles; this level of separation means there are no checks for unused variables and the encapsulation reduces unintended cross effects. Emergent behaviour therefore is caused by the interaction of effectors and dependant variables being forced to work through entities themselves.

The framework was developed with the Java [137] software development toolkit. As an object-oriented language there is strong conceptual correlation with an ABM approach. Its performance is comparable and often faster than other high-level languages such as Ruby, C++ or Python.

**Hardware** The framework was primarily developed and tested on local desktops with GIT cloud support and version control. More extensive computational resources were also available for intensive application cases and larger scale tests. CaStLeS [138] is a compute resource designed to support computationally intensive Life Science research. Two virtual machines (VM) with distinct use cases were deployed for the project. The first, a workspace, taking advantage of substantial computational power the models produced can be of a larger scope with greater moment to moment accuracy. The second VM was designed to host a crowd-sourcing web framework (https://ccf.bham.ac.uk) allowing greater security, staged roll-out and constant server up-time. Working in this manner also creates ongoing backup of work and gathered data.

### 2.2.2 Cancer implementation

Of the available *in vitro* non-small lung cancer cell videos 5 were of sufficient length and population density to produce interesting positional data sets post tracking 1. The first four were 12 second 84-time increment data sets with a fifth longer 54 second video processed into 181-time

increments with attendant tracks. We developed an effective interpretation and analysis pipeline via application-based development of the framework. Cells were tracked as a *in vitro* data source and analysed to identify salient movement patterns. Once target patterns and movement metrics were identified a background general behavioural model was defined. New population wide or environmental effects were iteratively developed and added to explore the emergent behavioural changes and cross compare with the *in vitro* data source to generate an explanatory hypothesis.

**Materials: available data set**

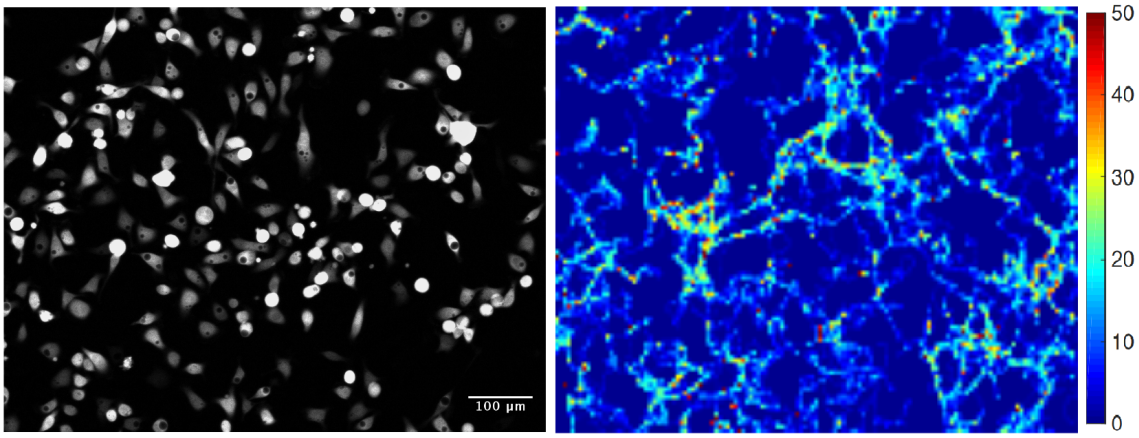



Figure 2.5: Left: a frame of a set of '*2D time-lapse microscopy image sequences of PC9 non-small lung cancer cells, incubated at 37 degrees Celsius in 5% Co2/Air in a humidified chamber.* ' right: a resultant frequency map for the low-density cell group developed by Credi et al [48]

In previous work by the group, sequences of microscopy images taken from non-small lung cancer cells after staining based upon mutational profile were processed by the image processing software *Fiji* [48, 62]. Extensive metrics such as transit frequency maps, repulsive force and distribution of cell turning angles both population wide and in relation to the heterogeneous mutational profiles were identified (Figure 2.5).

It was suggested that cell dynamics are heavily regulated by stigmergy, the indirect co-ordination of cells via incremental modification of the environment itself. This communication via modification of the environment was also suggested to be affected and disrupted by noise at higher numbers of cells; as the population size increases, contradictory or uninformative data also increases with a greater overall noise inducing effect. At lower densities, movement heatmaps

show clear network like movement and interconnections (Figure 2.5).

With such a wide range of possible development avenues it was reasonable to begin from a position that takes advantage of existing resources. Work focused upon modelling the given positional data from 2D microscopy image sequences of non-small lung cancer cells, collected in the form of videos. Several different densities of cancerous cells move and interact over various time frames. Analysis already developed by this group focused upon community clustering and motility within the micro-environment. *Fiji* was used to again track cells from the same data videos and confirm previous positional integrity. Repeating the tracking step also established a reproducible modelling and validation pipeline for future data sources.

#### Visualisation of results

Comparison across both models and *in vitro* observed data is made possible by digitising trajectories within a framework of common analysis libraries. Lower artificial variance improves comparison with common available generated meta data representations. We can take a POM approach to compare meta data patterns (Figure 2.6) with augmentation via statistics such as population size and travel distance (Figure 2.9). After digitising, cancer cell changes in turn direction and position can be recorded to identify population wide behavioural trends; generate patterns. Initial comparison is often with visual inspection, then a wide array of metrics is also generated to aid trend fitting or other more quantitative approaches.

After tracking entities across a run, a movement grid containing localised transit information over time can be used to generate a heatmap to visualize population wide movement patterns (Figure 2.6). By default, the highest movement score on a grid is found and a uniform colour gradient defined from 0 to this top score, black at 0 to the top-end value, white; the brighter an area of a heatmap, the greater the relative movement over time in that area relative to the top-end value. Gamma transformation can be added, or the top-end value manually set to a uniform number to compress the range and minimise outliers. By setting the top value for heatmaps we can also improve cross comparison, every shade of the gradient in one image then represents the same value as the same shade in another, in this chapter all heatmaps are normalized to the same top value.

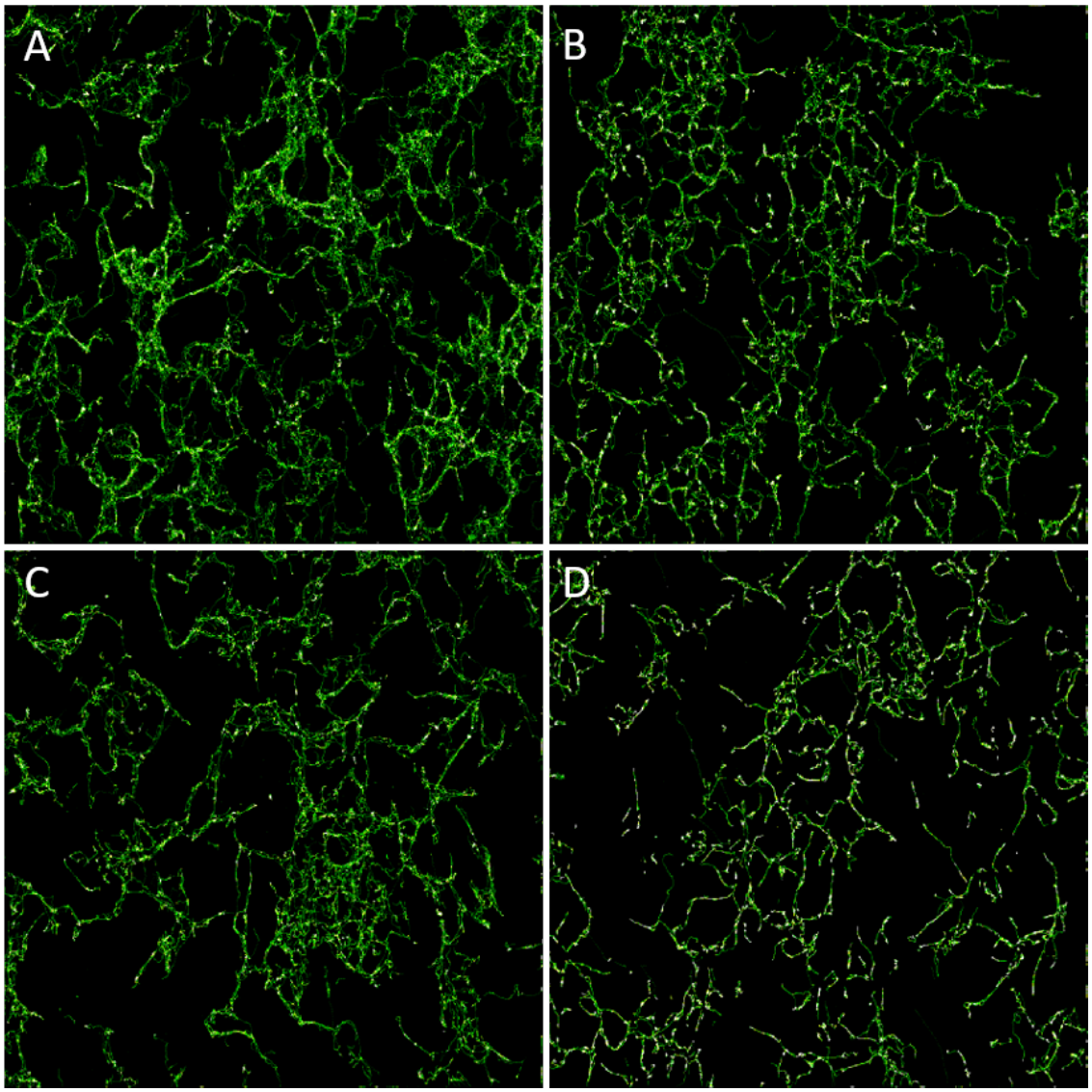



Figure 2.6: Movement over time heatmaps (2.2.2), representation of several 1-4 (A-D) short length (12 second) tracked videos of non-small lung cancer cells after framework representation.

### Directional visualisation

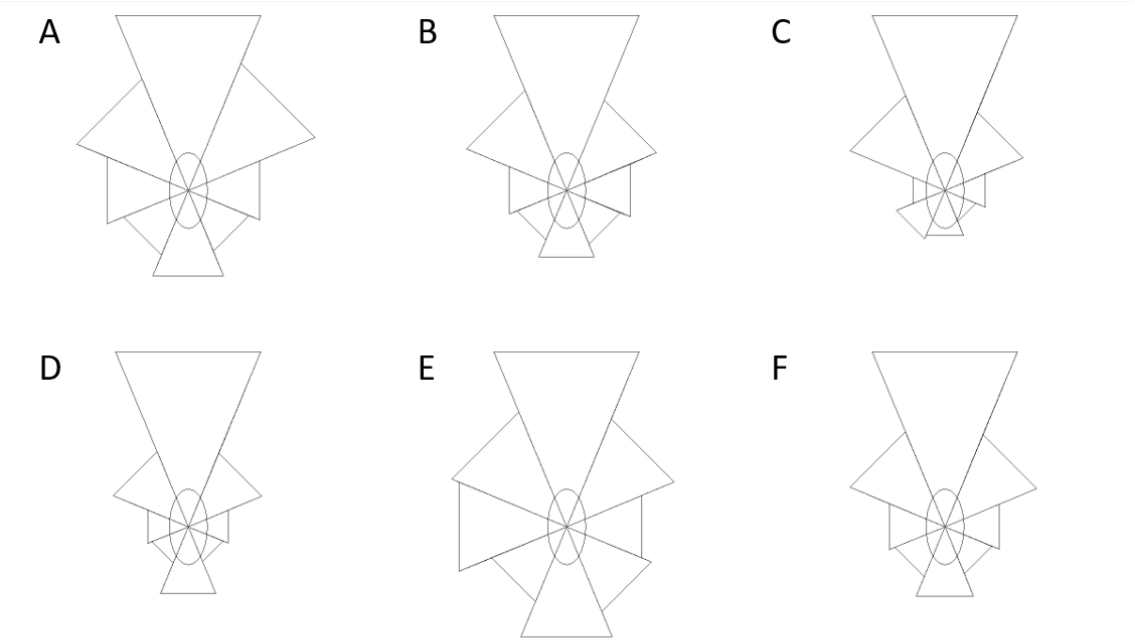



Figure 2.7: Every time a population member turns the angle relative to their current heading is recorded. By summing these relative turn choices across a population and time we can create directional preference diagrams relative to 0 at true north. Cancer sets A-D and the longer set E along with a meta diagram across all sets after 84 time increments (F) can be automatically generated.

Preferential turn directionality of group movement is not immediately clear by only quantifying population metrics and observing movement heatmaps. Turn diagrams represent population wide directional preference being generated by recording angular relativity (Table 2.4) at change points (Figure 2.7). When making a turn, cells are measured for the new angle relative to previous direction or simulated north. Turning circle is broken down to eight evenly sized segments, the relevant box being incremented when a new turn is registered. To show directionality relative to cardinal observable angles within a video, absolute diagrams are used, each segment being representative of turns in the shown direction: the larger a segment, the more turns occurred. Relative turn diagrams represent turns relative to the forward direction of movement, so each angle box represents the likelihood of directional selection relative to an entities previous direction, north is considered forward when building the diagrams (Figure 2.7). There is also a modified version of directional diagrams that adds the total distance travelled in each direction to allow analysis of speed and direction combination.

### Model hyperparameters

We took a two step approach to initial cancer motion modelling based on the input data, separating generalized population movement from environmental effects. Sub-population interaction can be modelled by population definition, lattice paths environmental and path forging with following behaviour as a combination of cells affecting environment to drive ongoing behaviour. We can attempt to replicate general movement by comparing directional and distance trends across input runs and modifying directionally weighted decision-making in models. Directional preference variance can be set at population level as with other variables such as replication, death, directional change rate and following for path forging implementations. Conceptually, we implement path forging behaviour as heatmap following: the higher a movement heat, the higher a chance that an entity will turn towards it.

Lattice path simulation was achieved by addition of a cubic curve in the space. A cubic curve is the result of four points in two-dimensional space, relative point position determines the bend, length, and overall shape. Our implementation determines the cubic curve and attempts to mirror it to a given path thickness; these mirrored paths then function like variable resistance strands. Each strand is segmented, and each segment given separate attraction or reflection strength, cells can identify a resistance path and join or leave based on local variance.

We simulate path forging behaviour by implementing an ant path like heatmap following component to the model. Ant pathing is known as an optimization algorithm mimicking real-world ant's pheromone tracking process. As ants travel, they lay a pheromone signal, the more ants the stronger the chemical signal and the more likely other ants will follow. In principal we can conceptualize the modelled process as a similar interaction via the movement heatmaps, when deciding to turn entities observe their surroundings and modify their roulette weighted decision making to favour higher movement concentrations.

Hyperparameters for our model definition primarily pertain to the inclusion of environmental and population effects. Then, modification of parameters dictating ranges or number of included

relevant entities give finer control and modification of a model set. In the case of cancer modelling, we seek to generate a basic population size and motility behaviour, lattice paths and path forging. Therefore, our hyperparameters should be replication, death, path inclusion and forging behaviour inclusion. We modify and explore the effects of variables for replication and death rate along with path number and following strength to generate representative model results.

| **Hyperparameters** | **Description** |
|---|---|
| **Movement type** | Describes the logic used for entity directional selection, roulette uses a weighted system selecting randomly from directional groups with a weighted preference. |
| **Iterations** | The number of steps in any given model run. |
| **Entity number** | The number of starting entities in a model run. |
| **Direction weight one** | Weight of 0-45 degrees. |
| **Direction weight two** | Weight of 45-90 degrees. |
| **Direction weight three** | Weight of 90-135 degrees. |
| **Direction weight four** | Weight of 135-180 degrees. |
| **Direction weight five** | Weight of 180-225 degrees. |
| **Direction weight six** | Weight of 225-270 degrees. |
| **Direction weight seven** | Weight of 270-315 degrees. |
| **Direction weight eight** | Weight of 315-360 degrees. |
| **Movement speed (px)** | Number of px moved by entities on average per iteration. |
| **Attractive pathing** | Whether an ant like attractive path implementation is applied, entities tend to prefer movement towards heavily travelled areas. |
| **Replicating** | Whether entities replicate during a model run. |
| **Replication chance** | At each increment, the replication chance. |
| **Cubic curves** | Whether attractive or deflective curves are applied within the model. |
| **Curve attraction/deflection** | Upon intersecting a curve, an entities average chance to attach, deflect or continue along at each increment out of 100000. |
| **Number of cubic curves** | The number of included curves total. |
| **Size of cubic curves** | Curve width, entities can move within broad attractive curves and interaction with other entities may not bounce them out. |

Table 2.2: Key hyperparameters for the chapter with a short description. A full list can be found with the code repository https://github.com/Benkwitz-Bedford/AB-FABS

| Hyperparameters | 2.12 | 2.13 A1 | 2.13 A2 | 2.13 B1 | 2.13 B2 | 2.14 A | 2.14 B | 2.14 C | 2.14 D | 2.15 A | 2.15 B | 2.15 C | 2.15 D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Movement type** | Roulette | | | | | | | | | | | | |
| **Iterations** | 80 | | | | | 480 | | | | | | | |
| **Entity number** | 400 | | | | | 500 | | | | | | | |
| **Direction weight one** | 15000 | | | | | | | | | | | | |
| **Direction weight two** | 4000 | | | | | | | | | | | | |
| **Direction weight three** | 1983 | | | | | | | | | | | | |
| **Direction weight four** | 2152 | | | | | | | | | | | | |
| **Direction weight five** | 8000 | | | | | | | | | | | | |
| **Direction weight six** | 1977 | | | | | | | | | | | | |
| **Direction weight seven** | 4000 | | | | | | | | | | | | |
| **Direction weight eight** | 15000 | | | | | | | | | | | | |
| **Movement speed (px)** | 0 | 1 | | | | 2 | | | | | | | |
| **Attractive pathing** | FALSE | | | | | | TRUE | | | | | | |
| **Replicating** | TRUE | | | | | FALSE | | | | | | | |
| **Replication chance** | 350 | | | | | 3000 | | | | | | | |
| **Cubic curves** | FALSE | TRUE | | | | FALSE | | | | TRUE | | | |
| **Attraction/deflection chance** | 90000 | 95000 | 100000 | 95000 | 100000 | 90000 | | | | 98000 | | | |
| **Number of cubic curves** | 20 | | | 10 | | 20 | | | | | | 10 | |
| **Size of cubic curves** | 5 | | | 10 | | 5 | | | | | | 10 | |

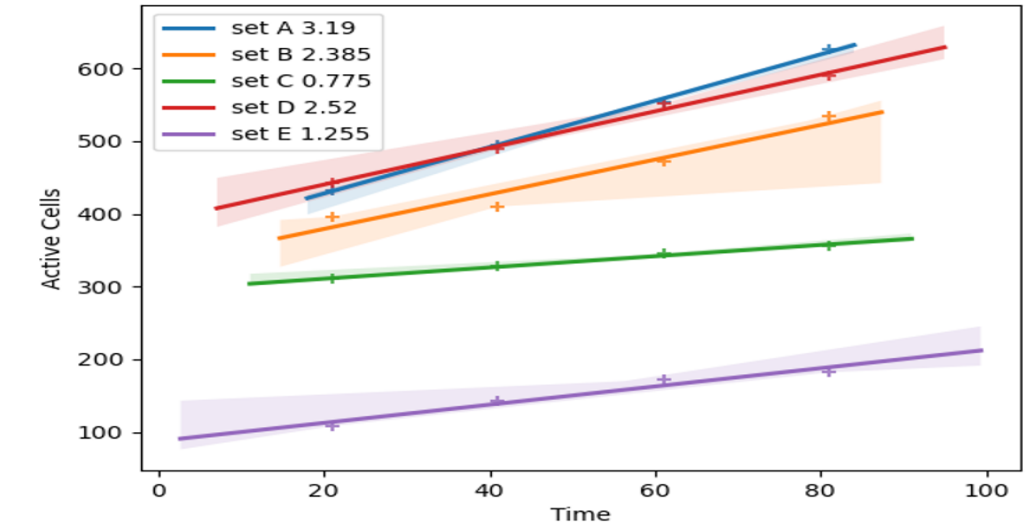
Table 2.3: Key hyperparameter values sorted by figure for the chapter. A full list can be found with the code repository https://github.com/Benkwitz-Bedford/AB-FABS

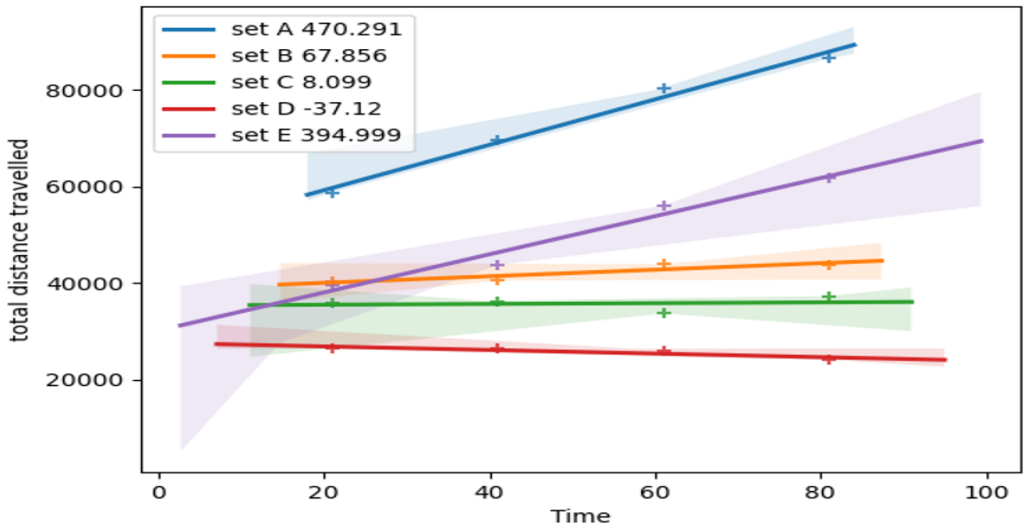
## 2.3 Results

### 2.3.1 Cancer population movement

#### Population trends: growth and movement characteristics

For the short videos (A-D) active population seems to grow slowly 0.78-3.19 population members per increment (Figure 2.8). Population growth rate is roughly proportional to existing size and continual, population controls such as nutrient availability are therefore unlikely to have been reached within the observed time frame. Over the shorter time frame of 84 increments, the longer population set E shows a much smaller population size, 1/3 of the smallest A-D set. The greater striated movement path pattern clarity of A-D may be due to larger populations accelerating exploitation or indicative of a more time advanced and developed micro-environment. However, while set D clearly shows less strand like behaviour, strands in C are present but not as distinct as in A and B, its likely that low population is tied to lack of strand development (C) but less impactful than interaction time (E) (Figure 2.6C2.10B). Travel distance does not appear to be directly associated with population size, greater cumulative distance does have a correlation with strand thickness however (Figure 2.8 E), lowest distance leading to the sparsest branches. Several of the quantitative graphs show steep drop offs for the final data snapshot, the final break only records 4 increments of activity for an 84 increment data set and should therefore be ignored (Figure 2.9). Turn frequency ties almost exactly to active population, at each time increment



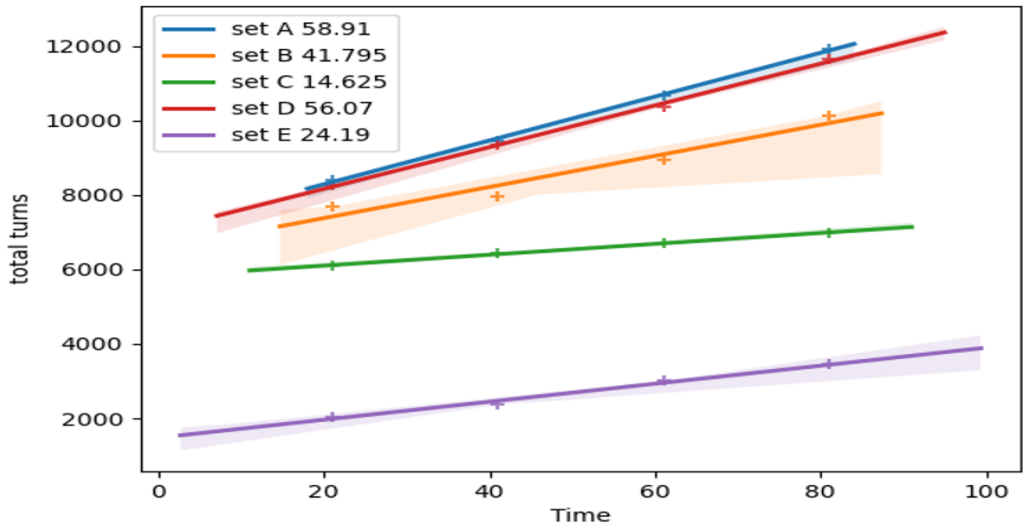




Figure 2.8: Here we generate summary statistics every 20 increments to generate quantitative graphs of active population members, distance travelled and turns made across populations can be made. Each line represents a different data input, the number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

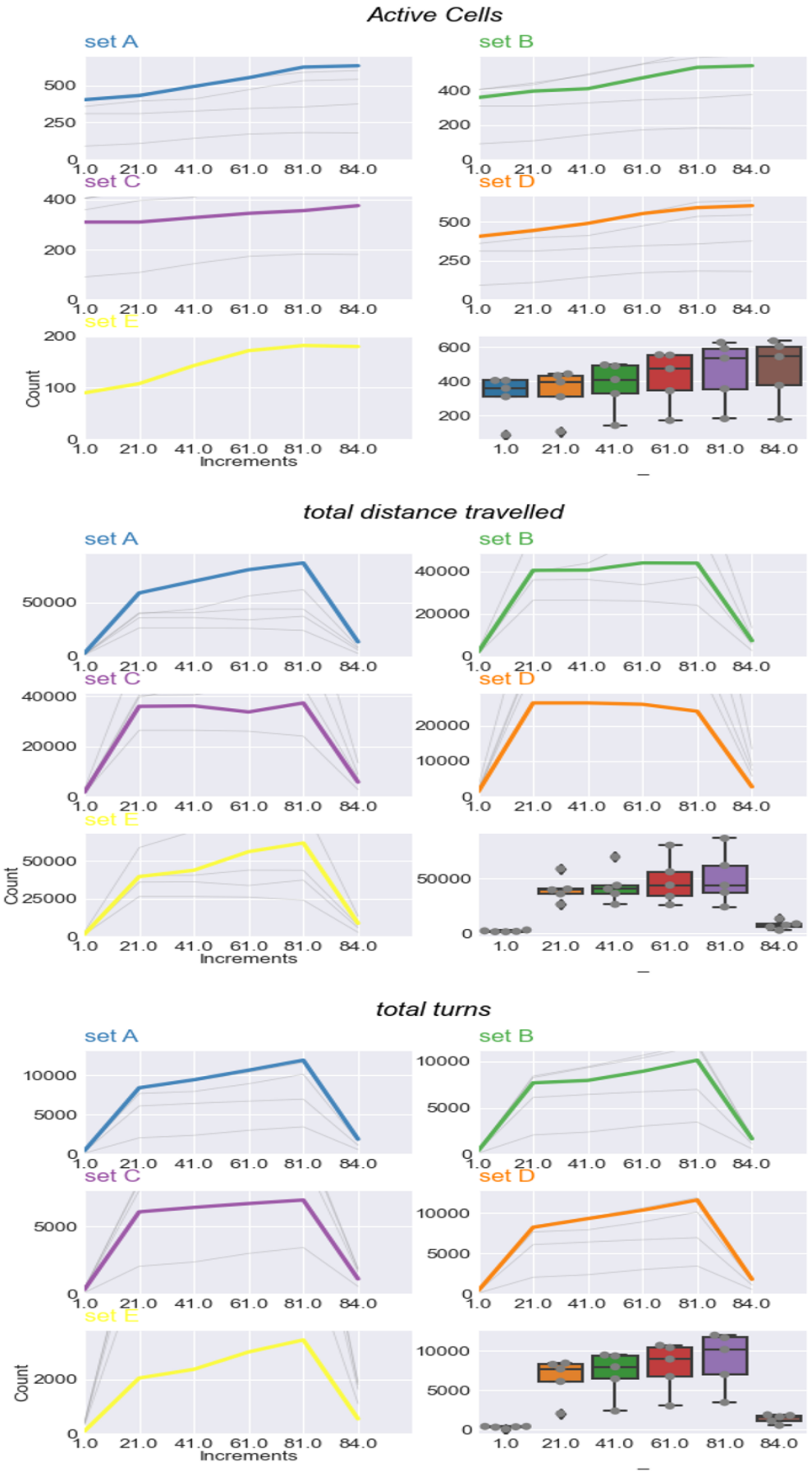



Figure 2.9: Active population members, distance travelled and turns made across populations for cancer sets A-E across 84 time increments. Per segment tracking graphs show a sharp downturn for the last snapshot because at 84 increments the last segment comprises only 4 active time steps.

every population member makes a turn in a direction that can be recorded, only the longer fifth population shows any departure from this trend but with a smaller population any variation would also be more noticeable (Figure 2.9).

**Longer time-frame** Observing a population over a longer time frame usually creates a greater amount of comparative points for identification of fundamental behavioural trends. Much like a computational model, few data sets are perfect but they can be useful. With longer sets, behaviours can be lost and noise generation more noticeable. In the longest cancer data set (E) bright lines of striated strand like movement are less clear but present (Figure 2.10). We can still identify defined strand movement of brighter attachment within increased general movement areas. Some areas become homogenised and no longer strand like indicating general random movement, existing strands also thicken but become more pronounced. If we take active cells as indicative of overall time stage this longer set is still prior to the other cancer sets(Figure 2.10 C). E is likely taken from an earlier formative point but runs for longer. We may observe an earlier less developed environment moving to a situation better represented by the shorter data sets. Again, cancer cells turn constantly, the nature of turns should be identified, constant micro turns being less obvious than pronounced directional change. Per member travel distance falls with population growth, perhaps suggesting post proliferative exhaustion or population based localised movement restriction via compression.

### Directional preferences

In line with path following turn data shows a strong foreword bias across runs A-D (Figure 2.7 and Table 2.4). Directed motion usually requires consistent small turns to maintain driven exploration. High likelihood of turns at every available increment might indicate high micro-turn prevalence in following curves or slightly indirect paths. While micro-turns are prevalent, they do not preclude all other directions. The longer fifth video shows more generalised heatmap movement and least forward biased turn preference although it is still pronounced (Figure 2.7 E).

The first of the four shorter data sets (A) shows a more pronounced general turn pattern.

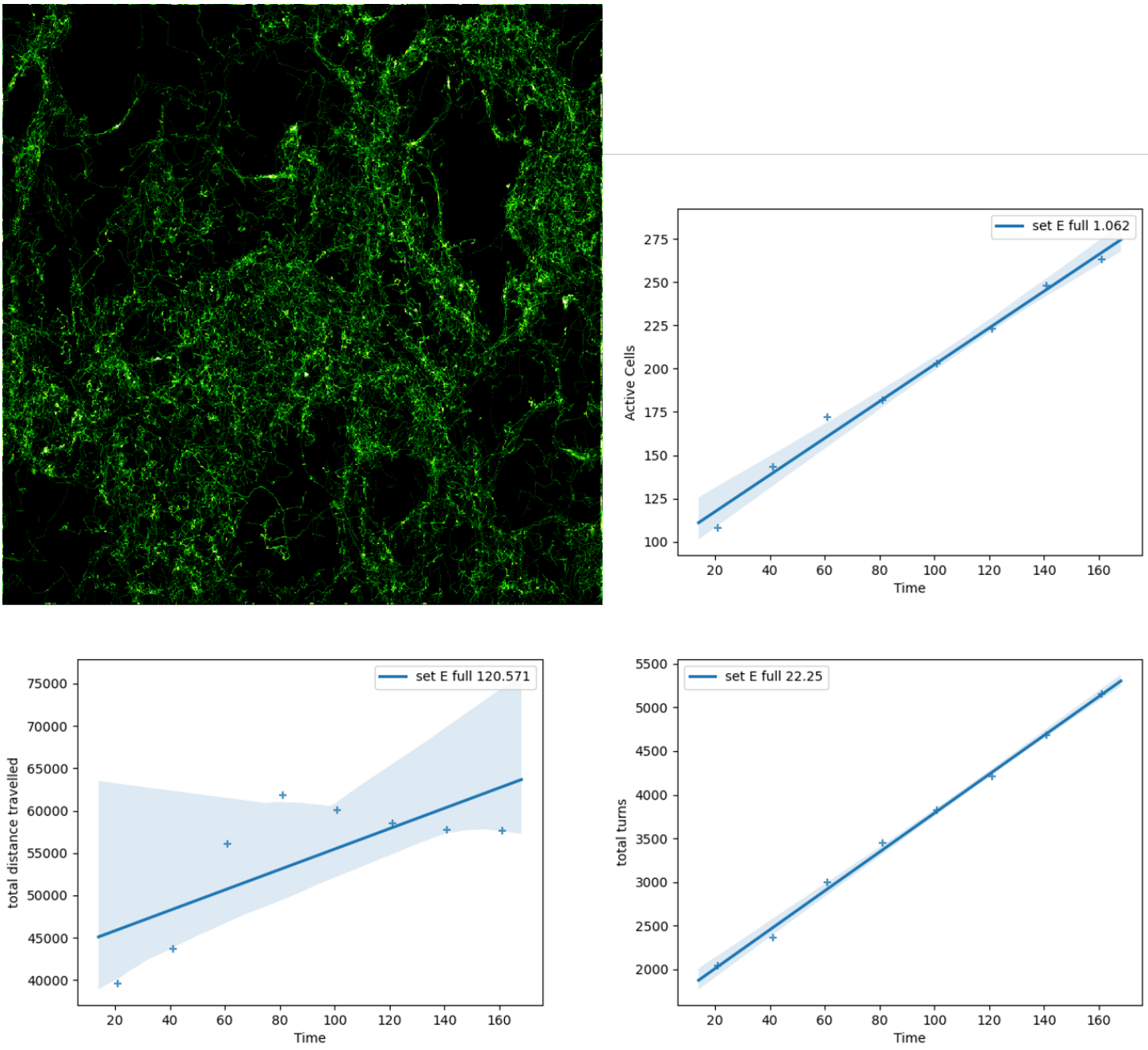



Figure 2.10: Movement heatmaps (2.2.2) are made after 181 increments. Keeping a snapshot interval of 20 increments active cells, distance travelled and total turns by a population can also be observed. The number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

| | Time | zero | one | two | three | four | five | six | seven | s.deviation | mean | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 21 | 27.78% | 14.61% | 8.81% | 7.90% | 9.87% | 7.81% | 8.96% | 14.27% | 509.71 | 946.25 | 7570 |
| | 41 | 24.32% | 13.77% | 9.06% | 8.65% | 11.70% | 8.68% | 9.09% | 14.73% | 454.55 | 1063.25 | 8506 |
| | 61 | 23.78% | 14.57% | 9.47% | 8.56% | 11.77% | 8.61% | 9.05% | 14.19% | 497.03 | 1203.38 | 9627 |
| | 81 | 24.14% | 14.77% | 9.33% | 8.29% | 11.48% | 8.55% | 9.11% | 14.33% | 577.94 | 1352.00 | 10816 |
| B | 21 | 29.59% | 15.61% | 7.88% | 6.78% | 9.14% | 7.22% | 8.12% | 15.67% | 544.66 | 874.38 | 6995 |
| | 41 | 28.30% | 15.38% | 8.42% | 7.36% | 10.06% | 7.57% | 8.14% | 14.77% | 512.85 | 899.75 | 7198 |
| | 61 | 26.11% | 14.17% | 9.51% | 8.12% | 10.50% | 8.55% | 8.65% | 14.39% | 488.67 | 1013.38 | 8107 |
| | 81 | 25.57% | 14.63% | 9.49% | 8.00% | 10.25% | 8.20% | 9.00% | 14.87% | 546.77 | 1152.75 | 9222 |
| C | 21 | 35.81% | 14.42% | 7.46% | 6.11% | 8.40% | 5.82% | 6.66% | 15.31% | 554.35 | 684.88 | 5479 |
| | 41 | 36.15% | 14.30% | 8.01% | 5.93% | 8.95% | 5.62% | 6.90% | 14.15% | 584.93 | 721.00 | 5768 |
| | 61 | 32.53% | 14.55% | 7.38% | 6.67% | 9.37% | 6.72% | 7.61% | 15.16% | 534.17 | 760.38 | 6083 |
| | 81 | 31.67% | 15.11% | 8.70% | 6.41% | 8.45% | 6.82% | 8.40% | 14.44% | 527.64 | 784.25 | 6274 |
| D | 21 | 36.24% | 13.14% | 6.59% | 6.69% | 10.41% | 6.97% | 6.23% | 13.73% | 758.94 | 943.00 | 7544 |
| | 41 | 33.30% | 13.53% | 7.01% | 7.08% | 10.89% | 7.03% | 7.31% | 13.85% | 759.28 | 1066.50 | 8532 |
| | 61 | 33.31% | 13.31% | 7.55% | 7.89% | 11.12% | 7.24% | 7.22% | 12.37% | 829.19 | 1183.75 | 9470 |
| | 81 | 32.74% | 13.39% | 7.30% | 7.60% | 11.39% | 7.91% | 7.39% | 12.28% | 910.55 | 1333.75 | 10670 |
| E | 21 | 19.60% | 12.95% | 10.77% | 9.28% | 13.98% | 9.86% | 9.34% | 14.21% | 61.17 | 218.13 | 1745 |
| | 41 | 21.57% | 14.70% | 10.06% | 8.87% | 12.76% | 8.62% | 9.52% | 13.90% | 87.25 | 250.88 | 2007 |
| | 61 | 21.96% | 13.80% | 10.16% | 8.82% | 12.27% | 9.33% | 9.41% | 14.24% | 111.09 | 318.75 | 2550 |
| | 81 | 23.61% | 12.75% | 9.53% | 8.62% | 11.80% | 8.48% | 10.65% | 14.55% | 145.98 | 368.50 | 2948 |
| F | 21 | 31.40% | 14.34% | 7.88% | 7.07% | 9.80% | 7.20% | 7.65% | 14.66% | 2416.18 | 3666.63 | 29333 |
| | 41 | 29.57% | 14.22% | 8.24% | 7.47% | 10.68% | 7.43% | 8.03% | 14.35% | 2389.47 | 4001.38 | 32011 |
| | 61 | 28.18% | 14.09% | 8.67% | 7.98% | 10.94% | 7.97% | 8.26% | 13.92% | 2447.62 | 4479.63 | 35837 |
| | 81 | 27.91% | 14.27% | 8.74% | 7.77% | 10.72% | 8.02% | 8.63% | 13.94% | 2689.07 | 4991.25 | 39930 |

Table 2.4: Turn choices made by cancer populations A-E and averaged set F across time relative to the angles shown in turn diagrams; angles 0-7 clockwise from north (Figure 2.7).

Busy areas might reasonably be avenues for path joining or areas of lower path attraction and population spread. Across all data sources a meta image indicates that a similar fundamental pattern is present (Figure 2.7 F), we can use a model to try and identify whether this is an environmental or cell driven effect.

### 2.3.2 Cancer models

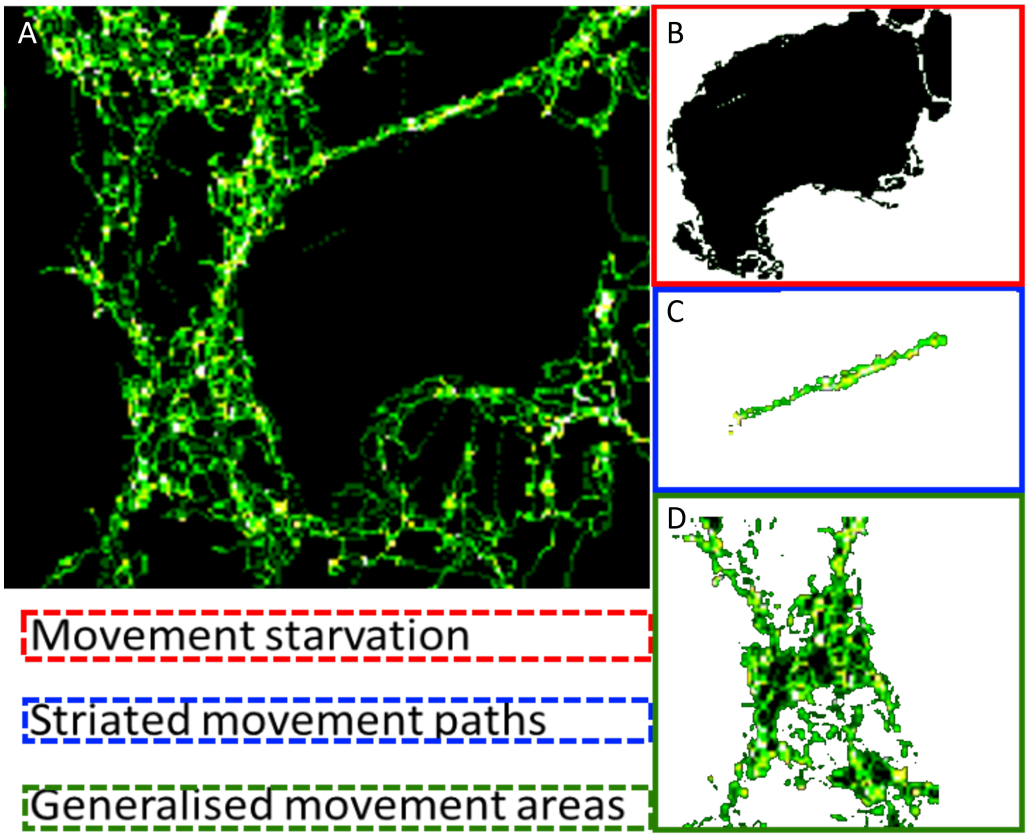



Figure 2.11: Key patterns identified via visualisation of cancer sets A-D, examples from A movement heatmaps (2.2.2) are made after 181 increments. Patterns shown are separated from a subsection of cancer set A (A) into movement starved (B), striated paths (C) and generalised areas of movement (D)

After observing tracks taken from the five-cancer cell movement videos we have several patterns to replicate and explain with ABM representations (Figure 2.11). We may seek to identify whether striated movement paths are caused by built in paths of least resistance or developed over time by path forging. Which is more likely, are sub patterns generated and could both be present. Similarly, we wish to identify whether the foreword bias observed in turns was environmentally driven or a preference for continued directional motion.

### Background movement

A simulated cell can be made by manually modifying model definition variables to generate similar patterns to real-world measured metrics such as population size (e.g. 400-500 cells), travel distance (e.g. 45000 steps) and for example strongly forward biased turn preference with constant selection (Figure 2.12). Since we can create a model that quantitatively resembles these population metrics in observed real-world *in vitro* data without many of the observed patterns we can hypothesize further. Primarily, the strand like movement patterns are environmental interactions, either cells are creating an environment of paths or exploiting existing paths of least resistance. Interconnecting paths are much clearer in real-world sets (A-D), even over short time periods, the attraction of the phenomena is likely substantial. Forward dominant motion is not in isolation sufficient to create the observed patterns. Furthermore, with dominant directional travel environmental exploration is less pronounced, perhaps exemplifying the advantage of least resistance travel.

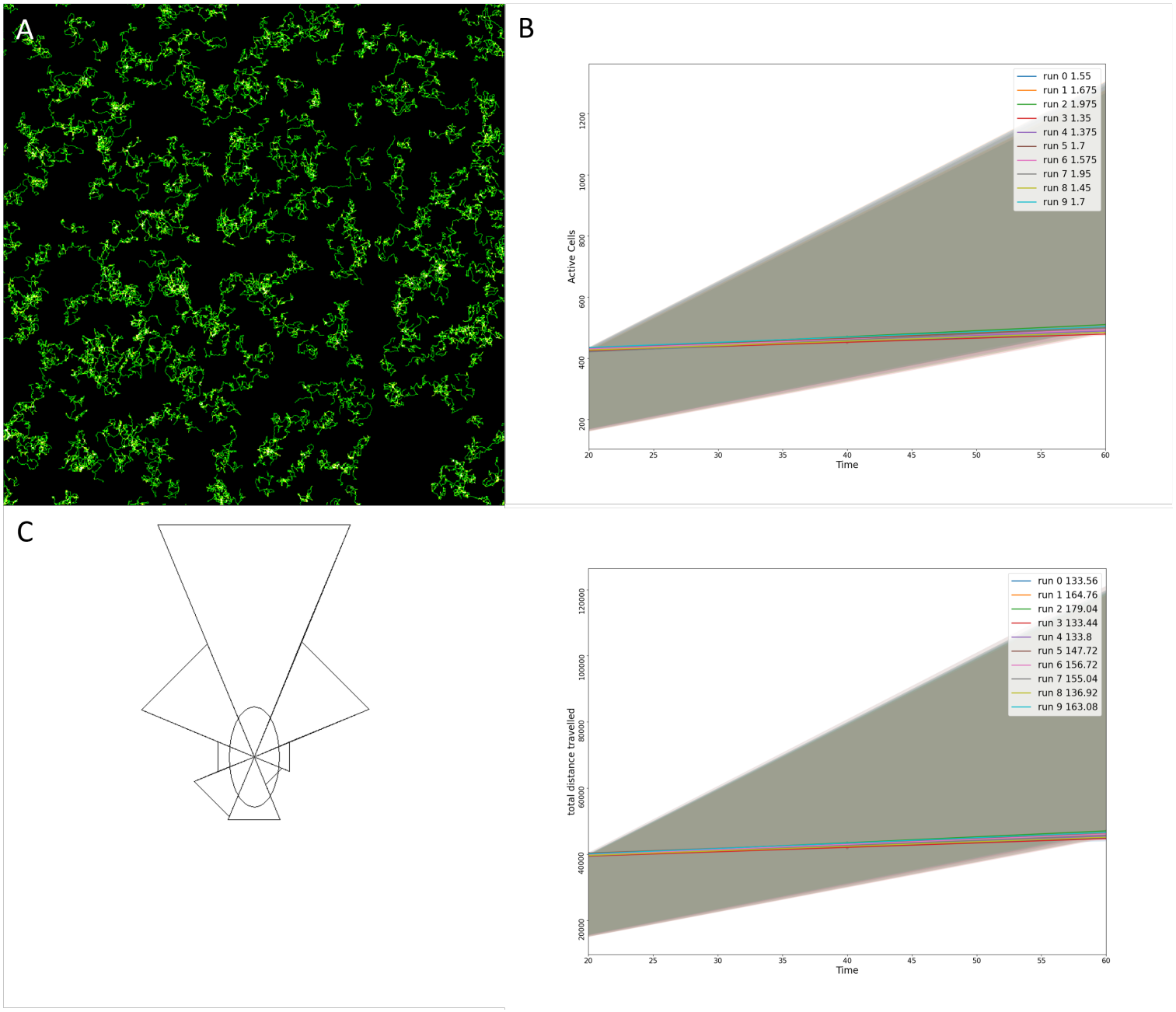



Figure 2.12: Movement over time displayed as a brightness gradient heatmap (2.2.2) at the end of a simulated run with observed numerical trends representing background movement for the cancer data set. Active entity participation, travel distance and turn preference can all be approximated and recreated across multiple separate model runs. Each line represents a different data input, the number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

**Lattice paths of least resistance**

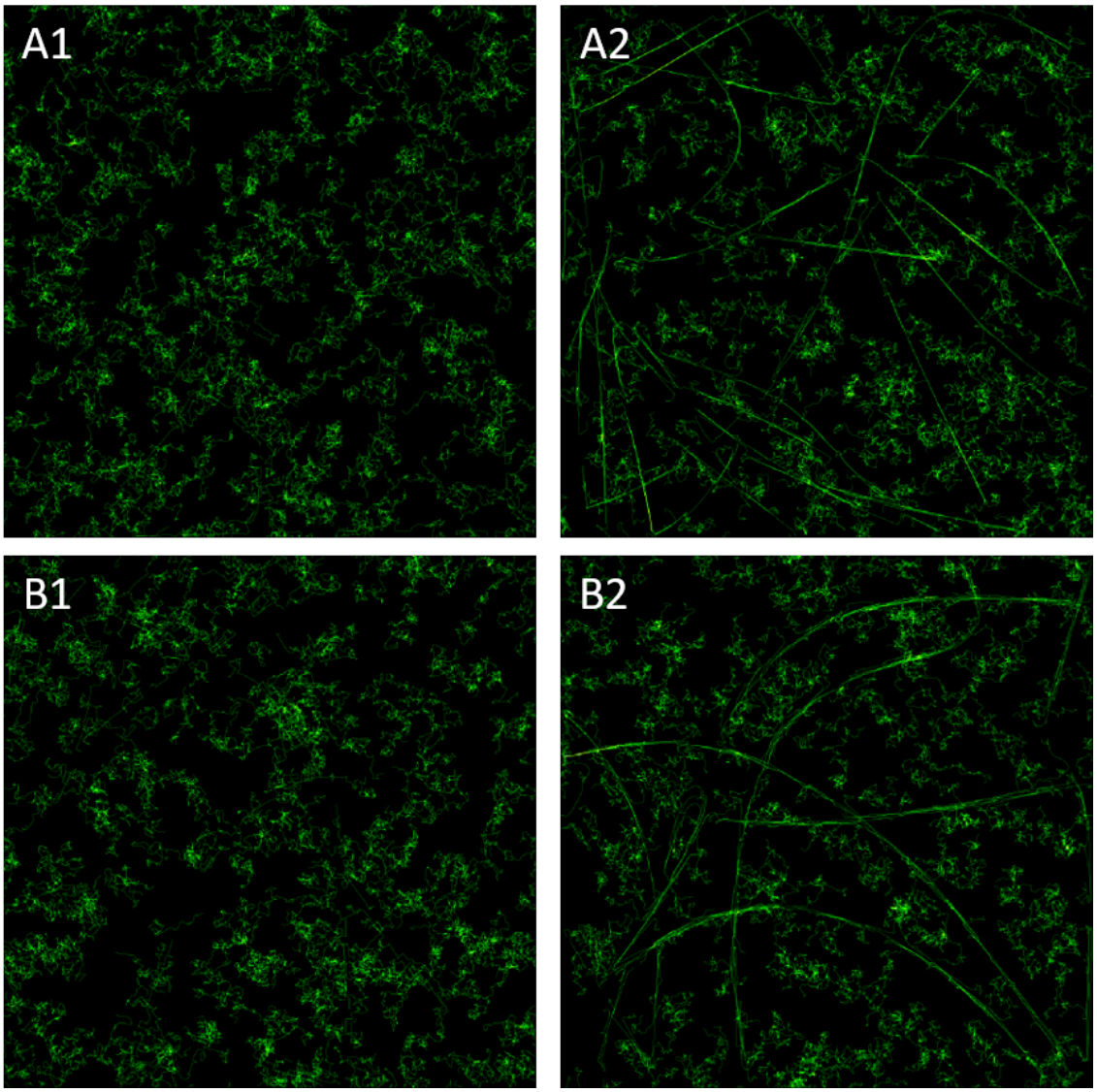



Figure 2.13: A,B) A range of movement density heatmaps (2.2.2) after operation of a cubic curve-based path of least resistance model for 84 time increments. Small width (A) and double width (B) attractive curves are applied with low (1) and high (2) attraction parameters.

We can add to the previous model of generalized background movement behaviour, a stochastic definition of cubic curves acts as attractive paths of least resistance (Figure 2.13). Paths cut through more homogenised random movement patterns with high strength attractive curves. However, such model does not remove the patterns of generalized movement before finding and attaching to a path or with lower attraction. The lack of generalised movement in the real-world sets indicates videos were taken at a time point where most entities were already using paths.

Embedded paths of least resistance create the stark path-following we see but not the population clustering. Some paths in the model show similar localised clustering but without the clarity of *in vitro* data, these lattice paths also cleanly separated through existing movement clusters. Model results for low attraction paths show much fewer clear patterns overall but also merge with general movement patterns as in real-world sets A-E.

### Path forging

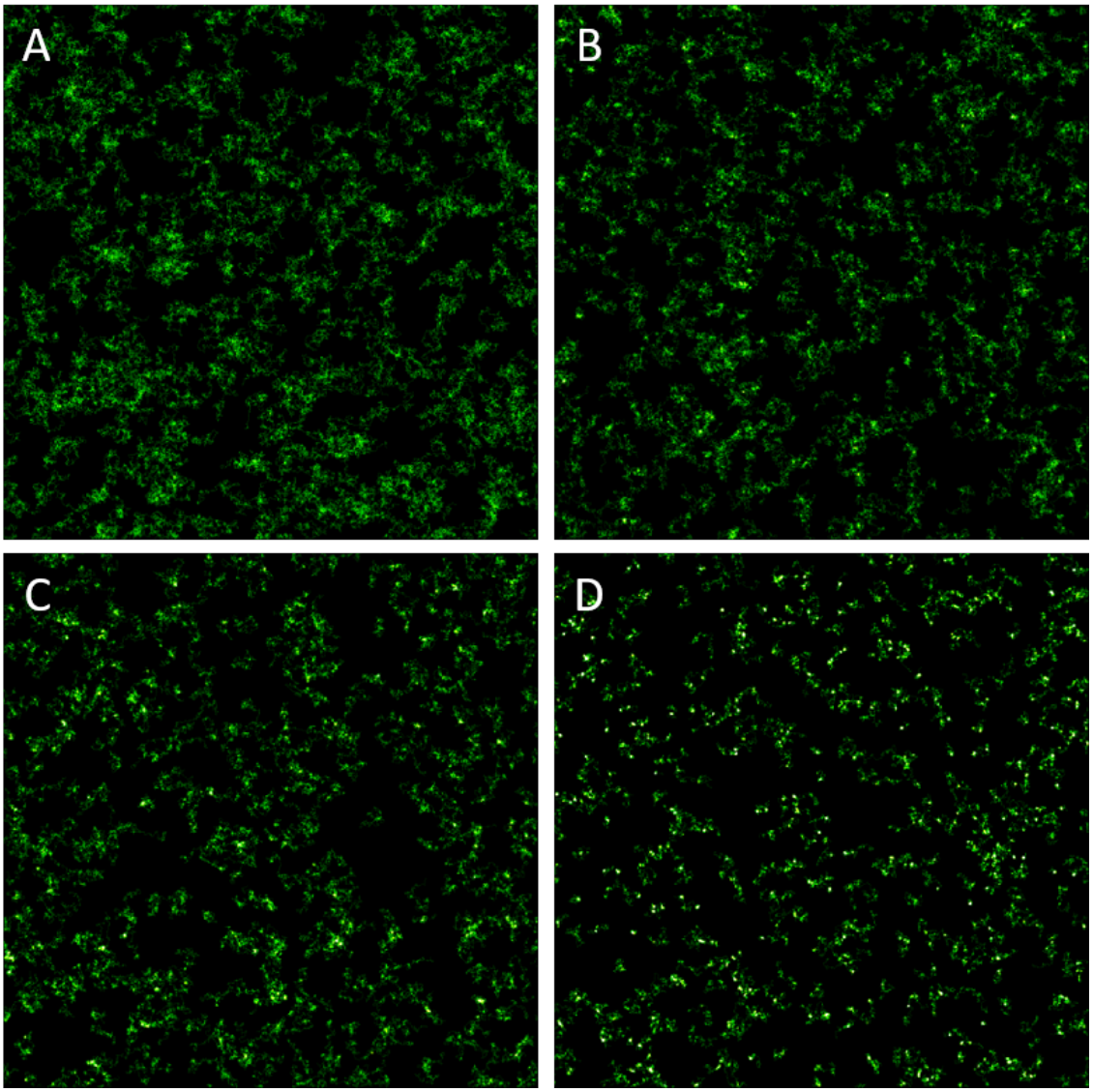



Figure 2.14: Both background movement approximation (Figure 2.12) and path forging behaviour, models can be stochastic operated and movement recorded over 84 time increments to show preference and population behaviour within gradient heatmaps (2.2.2). Heatmaps are from models with no following influence to low and then doubling following strength as they progress in the series (A-D)

Implementation of the path forging proxy via ant path implementation increases entity grouping as attraction strength increases (Figure 2.14). Very thin interconnections between more generalized movement hubs can be observed but with more emphasis on hubs than strands with reduced interconnection as strength increases. By increasing the attraction strength of forded areas, we can observe movement clumping and a reduction of generalised exploration, this is however without clear strand like following.

### Lattice and path forging combination

Due to the modular construction of the framework we can overlay multiple effects with minimal additional development time. In this case, the path forging behaviour representation and environmental lattice paths of least resistance can simultaneously be applied with the model of background movement. Several new patterns suggest possible similar interactions to conditions within the extrapolated cancer cell results (Figure 2.15). Lattice following is still clear but integrates into general grouping and movement. Strands between clusters of random background movement create overlap similar to patterns observed in lower population short videos. Despite bright areas suggesting evidence of strong clustering, general exploration of the available space is still present. With the introduction of lattice paths to an exaggerated path forging model, path following leads to greater overall spread. Addition of path forging and following results in wider paths as entities are attracted to general movement and create more permissive space over time. Interestingly, in the case of already broad attractive paths, path forging seems to narrow resultant stand like movement patterns more than weak attraction variants. A process of path growth might therefore be based upon an in-built lattice, they serve as a spline for path forging exploitation and development over time. Cancer cells are not however restricted only to paths of least resistance built into the environment; permissive zones are also created particularly at the intersections of strands.

### Longer time frame

We have implemented model representations of general movement patterns, lattice exploitation and path forging behaviour, time can be a major confounding variable. Prior to observed record and post recording larger time scales can allow creation of additional patterns for comparison. Within the cancer data set we have identified that run E is likely taken from an earlier formative

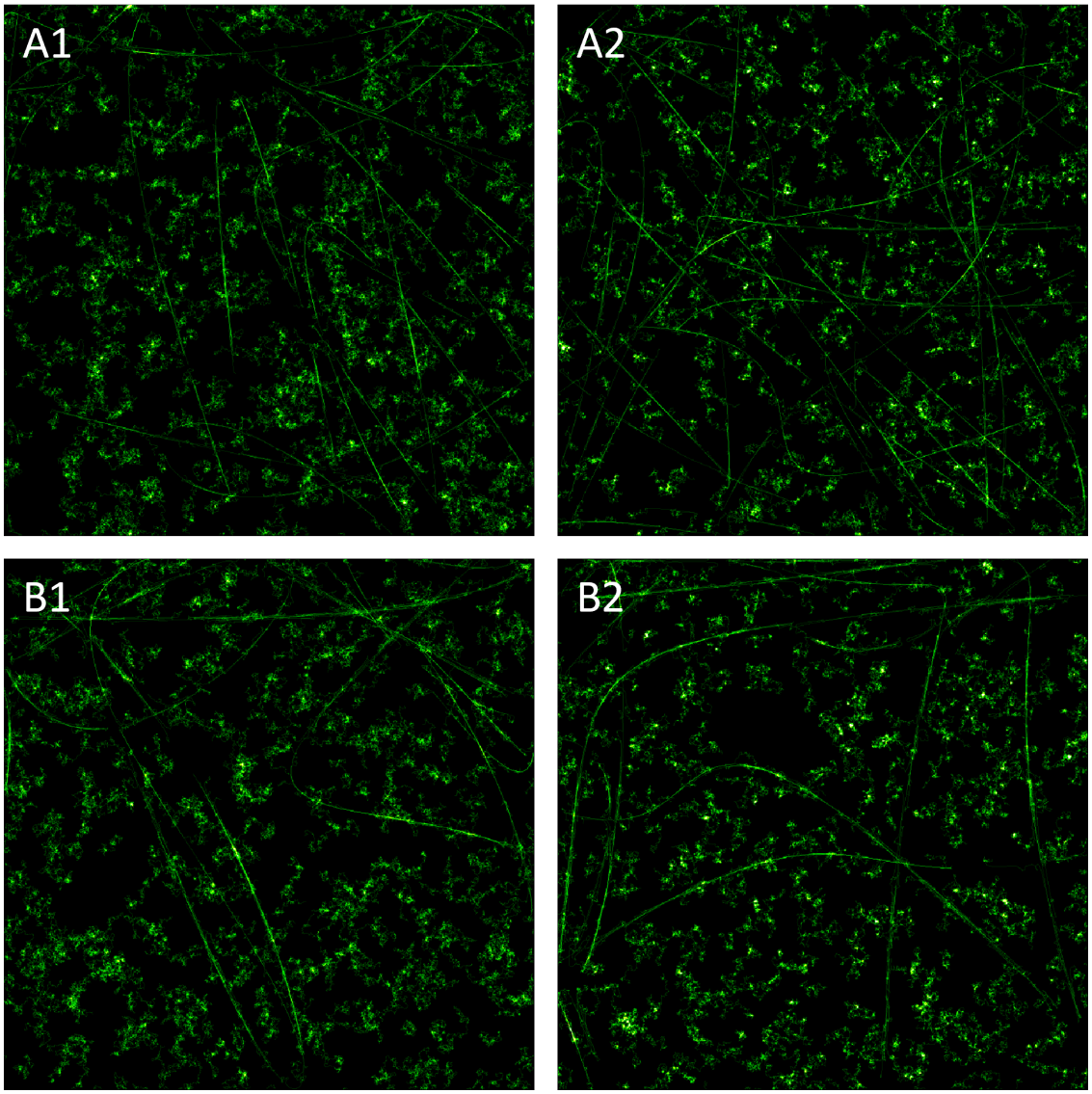



Figure 2.15: Models can be generated to include, background movement approximation, lattice of least resistance and path forging behavioural representations (Figures 2.12,2.13,2.14) for operation over several distinct run cycles with attendant movement heatmap (2.2.2) generation (A-D). Thin (A) and broad (B) attractive curves are combined with weak (1) and strong (2) path forging attraction.

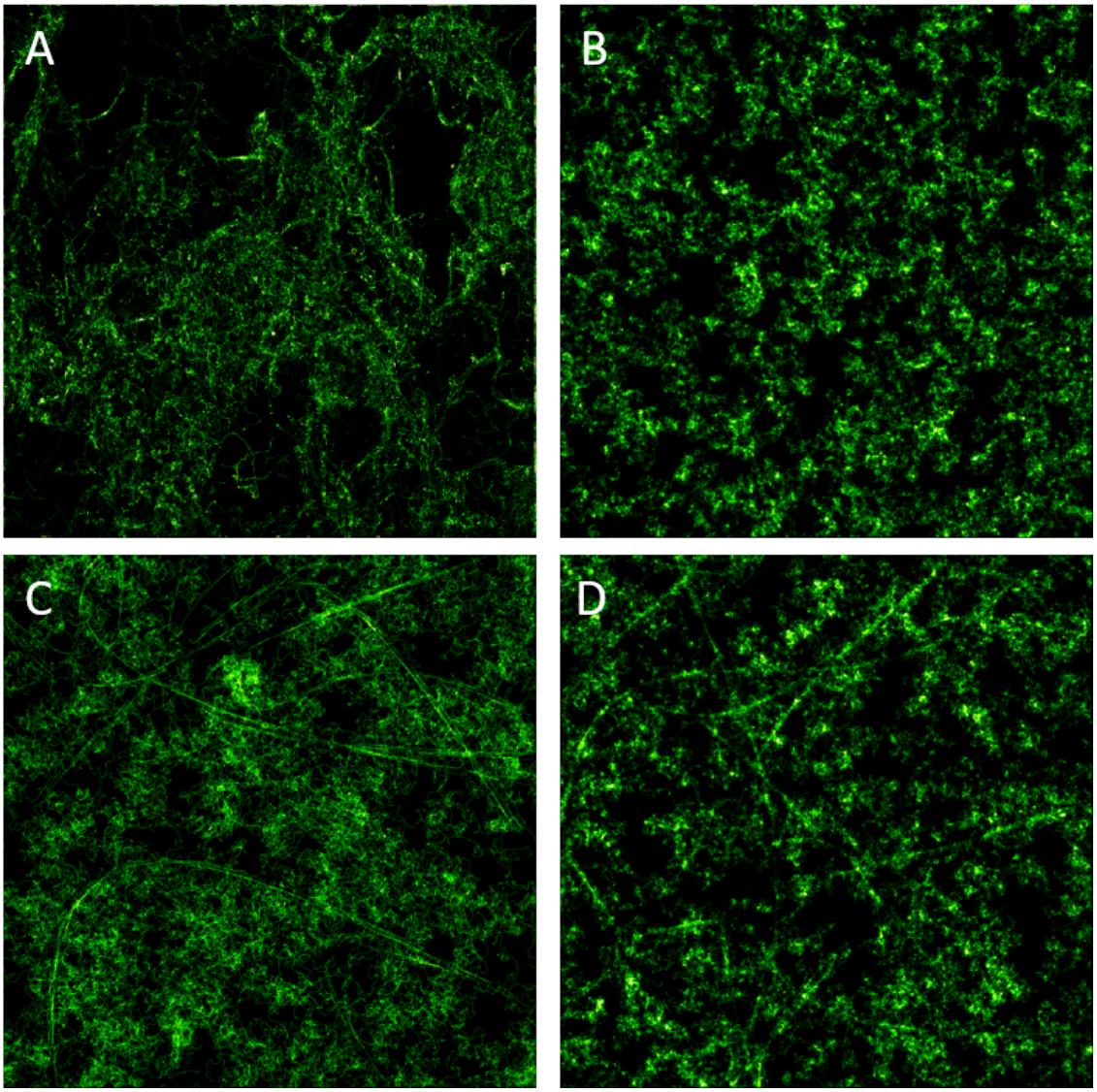



Figure 2.16: A) 181 increment cancer data set movement heatmap (2.2.2) can be compared to previous forging (B) lattice (C) and hybrid (D) models allowed to run for the same time frame, resultant movement is then visualized.

point but runs for longer. As such we can compare prospective formative models over a similar time frames to identify comparable pattern generation (Figure 2.16).

Comparing the longer 181 increment cancer data set with path forging and following models suggests a good analogy for general movement but without the presence of strong paths. Lack of directional preference in path forging behaviour leads to more generalised movement areas: despite a strong forward bias, for every step in an area turn towards pressure is only 1/20 weighted ratio to forward bias. As with shorter time spans, attractive curve implementation alone has similar issues of being too clear and cutting through general movement areas. Over a longer time span the attractive curves can funnel entities into general movement areas. Both least resistance paths and path forging sets seem to create small movement starved zones boxed by paths, entities are led out of the area by hitting lesser resistance and travelling away, a pattern also visible in the *in vitro* data set. Not all entities exist within the curves as shown by brightness similarity between curves and more random movement.

## 2.4 Discussion

With our framework, we were able to analyse real-world *in vitro* cancer data and iteratively develop explanatory models to support possible causal hypothesise for observed behaviour. We used a Pattern Oriented Modelling (POM) approach, quantitative population metrics, movement data over time and comparable models supported identification and association of several seemingly disparate patterns.

Some of our observed results, while easy to extract from a model, can be obscured when interpreting digitised *in vitro* representations. When tracking cells, loss of recognition can cause splintering into a separate cell representation. Fragmentation often results in new entities, once overlaid within the framework; a single *in vitro* moving cell becomes represented as two separate entities, one for each part of a track that was lost and subsequently split into two. Population measures and consequent metrics such as averages can become unreliable, taking readings at multiple time points of active population size can give a more accurate moving trend.

Regardless of splintering the total movement observed will remain the same. A cross section of population metrics and reasonable movement capabilities might be used to correctly approximate *in vitro* population growth. For example, selecting an arbitrary value, a simple filter may assume that entities existing for less than three time increments are likely fragments and should be excluded from total population metrics. We might then add other clauses such as concurrent overlapping entities as unlikely circumstances.

### 2.4.1 Cancer analysis

We identified incidental and population-based bias via comparison with a similar data set. Similarly, cross reference between initial work by the group [48] and new tracked extrapolations of the original data allowed confirmation of input and comparison during re-tracking of raw data videos.

We were able to identify trends such as following movement and least resistance exploitation. Qualitative visual validation at such an early stage created a necessary basis for further development. Clear striated common movement path patterns suggest cancer cells are following and travelling in similar areas to one another over time. Paths of least resistance and gradient following are well documented phenomena in cancerous behaviour and particularly metastasis [139, 124, 80, 92]. However, it is unclear whether least resistance patterns are due to a confinement effect forcing direction [140], nutrient starvation and following [139], localised environmental degradation [80] or other environmental interaction. Artificial effects such as an embedded medium based lattice paths or sub population behavioural differences are other possible explanations. Quantifiable metrics were also generated to provide further pattern comparison with the addition of modelled representative elements improving comparison points.

### 2.4.2 Model development

A first representative model was made by manually fitting measured metrics such as population size, cell turn rate and the addition of environmental paths of least resistance. Although

similar movement patterns were observed differences still exist, useful explanatory models not perfect ones. A loose definition of metric similarity or missing environmental factors along with misrepresented metric relevance, might be reasonable explanations. An implicit issue with data gathering is the temporal aspect, before our record environmental effects are difficult to approximate and make model development more difficult.

It is important to remember missing environmental factors can be consistent possible pitfalls, along with misrepresented metric relevance. Hence, we should focus on the useful aspect of model comparison, replication of specific patterns rather than overall replication.

It was difficult to identify whether the striated movement path patterns in *in vitro* results were completely random, driven by initial placement and path forging disposition or exploiting fundamental paths in their growth medium. Lattice based least resistance and path forging representative models can create pronounced strand movement. Simulated cells cover more of the environment indicating exploration and perhaps nutrient exploitation advantages.

Lattice paths seem more effective at replicating clear movement strands through areas of general movement. Path forging behaviour creates more complex curve patterns. A combination of the two may well be present in the *in vitro* environment. Both approaches showed some similarity and therefore possible representative phenomena.

Addition of the forging and following should have caused greater general movement reduction in favour of paths. However, the longer data set suggested that at an earlier time point more generalized movement is prevalent. Path cohesion and congregation seems to need to increase over time. Some similarity can be observed at different time points across path forging and lattice models, a combination may be present in the observed cancer micro-environment. One possible explanatory scenario is that of coalescence over time into stand like groups of cells in common areas but with lower generalized exploration. Much like other self-organizing systems, initial exploration of an environment is then exploited with convergence on common paths.

A pre-established network or interaction of large path forging entities and smaller following groups might explain the pattern differences; entity sizes do vary but aren't tied to path forging ability in the model. Addition of the path forging behaviour may not currently create attractive enough movement areas, movement patterns are more similar to the long data set than lattice alone, it is possible coalescence has not yet occurred.

### 2.4.3 Conclusion

We have applied the framework to *in vitro* digitization, direct analysis and generation of representative model results for comparison. As a first step design, targeting cancer movement created very generalizable libraries and several effective use cases for expansion, identifying future avenues for development.

When comparing with existing literature we can identify several interactions that we are currently unable to include or do not represent. For example, the framework includes the ability to define infinite sub populations, heterogeneity being fundamental to tumour arrangement and modeled before [141]. Similarly, observing the original cancer cell videos we can see clear cell morphology differences, morphodynamics can be important as both fundamental to cell navigation [142] and indicative of motility methods [140]. Further work could also include parameters for cell nutrition, environmental permeability and ECM stromal structure. Ultimately the scope of included models is limited but there are clear possibilities for hyperparameter or effect addition and possible model improvement.

Further, quantifying analysis can also be applied when possible, possibly utilizing neural nets or crowd sourcing to grade image similarity. A time phase tool would also allow more in depth discussion of pattern coalescence or transience. It is possible that widespread movement at earlier points is adding noise to the picture, such a change would highlight differences. We can also identify and apply sub population filtering by defining *in vitro* data set subdivisions based upon collected metrics. One simple subdivision is that of replicating low motility groups and high motility exploitative cells. We could split *in vitro* data groups on either movement amount or directed movement preference; both populations suggest trait separation. The framework can then use this information to create a model with both sub population groups present for comparison.

Initial next steps will focus upon application of the analysis framework to a larger and more experimental data set which required more interaction with the collaborative team in order to study of G protein coupled receptors and G protein interactions at cell surface and their movement data. Further development of the framework could then also be applied back to the cancer set.

# 3 Modelling G protein coupled receptor and G protein population movement and interaction

## 3.1 Introduction

GPCRs are a large and diverse group of cell surface receptors. They allow cells to sense and react to changing environmental conditions: many larger and important systems depend upon effective functioning of this fundamental cellular capability. To communicate extracellular behavioural changes, receptors interact with G proteins in the plasma membrane and react when,for example, co-localizing to send signals into a cell. Therefore, to further understand how a cell identifies, controls, and understands extracellular signalling, both receptor and G protein movement were captured as video sequences. A typical GPCR was selected, its observation can help towards a better understanding of the general behaviour of the whole class.

Detailed visual interpretation of G protein-coupled receptors (GPCRs) is an increasingly important, yet still developing, area. Therefore, we aim to ease the study and understanding of the relationship between GPCR and G proteins to understand how their environment affects patterns of movement and behaviour.

We demonstrated earlier that comparison of Agent Based Model (ABM) with video taken from cancerous real-world *in vitro* biological systems, could generate new insight into cancerous cell movement. Here, we apply this same framework and generalizable holistic approach to this GPCR and G protein based biological system. We were granted access to movement tracks from both GPCR and G protein by the Calebiro group [6]. Direct discussion with the research group

in question concerning analysis, capture and tracking which led to rapid improvement integration, and iteration of model definition, along with result interpretation. Thus, we could better understand the problem space and identify potential artificial bias from previous data gathering stages and our modelling implementation. Importantly, to our knowledge no similar GPCR and G protein ABM exists so, direct comparison with a prior study is difficult.

GPCR and G protein movement was tracked by the Calebiro group [6] from captured videos to identify patterns that could be associated with known and hypothetical mechanisms [54]. Subsequently, these tracks allowed analysis of movement for co-localization, clear patterns, and correlation with other environmental phenomena such as cytoskeletal boundaries. To improve visualization of movement over time, we included some movement heatmap observation; similar to our cancer movement analysis. Sets of paired movement tracks for separate channels (C) showing G protein receptor (C1 ) and G protein (C2) entities over the same time frame within the same cell were studied. Despite the difference in scale, with individual proteins (GPCRs or G proteins) being in the nanometer range and cancer cell entities around 0.1mm, we could digitize the given tracks. We were able to easily and efficiently process them with the analysis tools within our previously designed framework. The results are more dependant upon microscope resolution and tracking capability than framework representation.

### 3.1.1 Summary

Initial results supported the phenomena described within literature: clear hot-zones and preferred movement areas [54, 143, 144, 6]. We were also able to support the hypothesis that these hot-zones were designed to increase co-localisation of GPCR and G protein; comparison across paired GPCR and G protein tracks revealing co-localisation within many such areas [6]. Therefore, to include theorized environmental causes such as boundary-based catchment or attractive zones, framework development implemented new hyperparameters for inclusion in representative models 3.3.2. Models included GPCR and G protein background motion as Brownian movement, variable deflective paths for cytoskeletal representation and attractive pull zones.

Initially we were able to show that a purely cytoskeletal confinement representative model did

not clearly replicate digitised real-world patterns, comparing both showed some clearly divergent movement behaviour (Figure3.25). Unlike with confinement models, we observed some similarity between real and attractive area model results. Further discussion also focused on movement starvation by entity attraction; when a strong attractor centralised a large proportion of a population in one place, nearby environmental exploration decreased over time.

### 3.1.2 The GPCR and G protein system

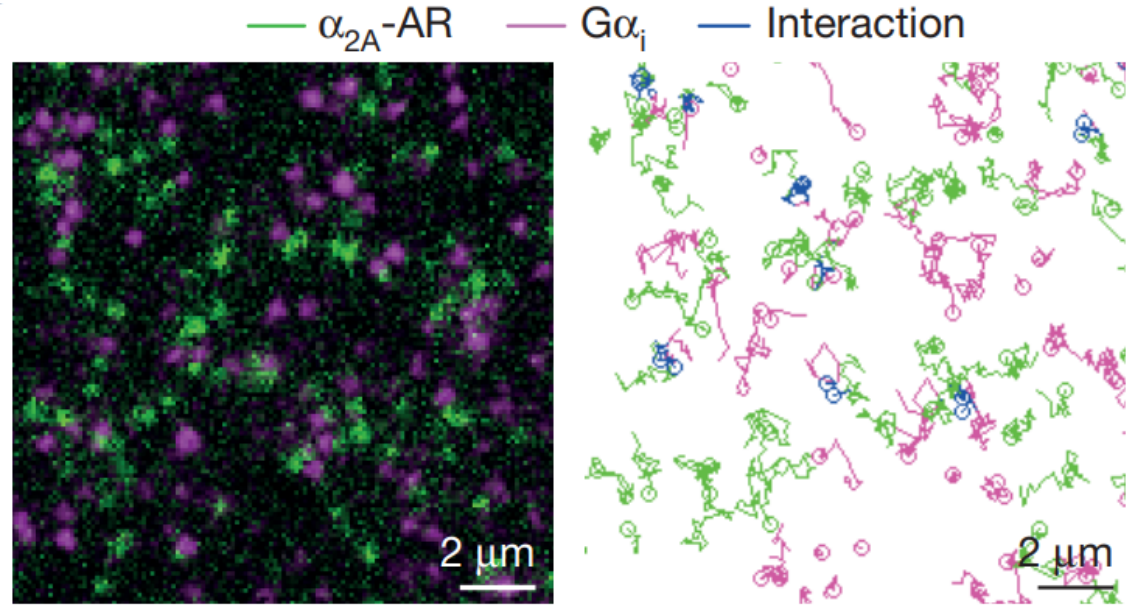



Figure 3.1: An image prior and post tracking of GPCR and G protein it depicts a: *'Selected frame from a fast single-molecule image sequence (left) and corresponding trajectories (right)'* [54]

To allow environmental sensing, a range of processes were evolved within a cell [6]. GPCRs are one of the most powerful examples of this environmental sensing and signalling whereby they can perceive their local conditions, comprising their own complex biological system. GPCRs enable cells to sense and then react to a wide range of environmental changes, also facilitating communication [6, 7, 8]. Receptors in the cell membrane can co-localize and then interact with G proteins, a signal is then transmitted to the interior of a cell, the exact scenario for this interaction remains uncertain and the subject of other research [54, 52, 53] (Figure 3.1). Therefore, to better explain the process the Calebiro groupo had chosen 'the 2A-adrenergic receptor (2A-AR), a prototypical family-A GPCR that couples strongly with the inhibitory G protein (Gi)' [52] for record and extrapolation via tracking [6]. In this work we therefore chose to observe and digitze these paired sets of GPCR (C1) and G protein (C2) tracks.

Primary limiting factors for framework input are usually population size, real-world entity tracking and reliable abstraction as sets of tracks. Despite the change in scale from cell to sub cellular level, the modelling framework still proved remarkably applicable. We were able to generate representative models with parameters for population variance and stochasticity. Much of the molecular behaviour associated with cellular function is described as stochastic and probabilistically driven [57]. Formation of complexes and co-localisations are often described as driven by stochastic regulation; mechanisms cause local environmental change compressing entities and enabling rapid transient interaction across areas filled with ordinarily random movement [58].

We enabled various locality based spatial representations [57, 58, 54, 6] with the implementation of our ABM approach. Our novel framework provides an opportunity to simultaneously generate qualitative visualization and quantitative analysis for these important biological systems.

#### GPCR and G protein as a fundamental cell process

We can improve our understanding of fundamental cell processes, pharmacological targets, and drug performance by developing and testing conceptual models of GPCR and G protein interaction and behaviour. Thus far, in other work done by the Calebiro group, analysing tracks with new observation techniques has revealed deep GPCR and G protein domain complexity [6]. We expect that important further insights should be gained by introducing new combined approaches to system visualisation, modelling, and analysis.

Regarding pharmacological improvement, GPCRs often influence and translate the biological effects of hormones and neurotransmitters, in doing so they present reasonable targets. While a large portion of currently marketed drugs target GPCR [49] only a fraction of targetable GPCRs are exploited [6]. With our approach, we could further define interaction, co-localisation and behaviour of GPCRs still not entirely understood by observing changes in local behaviour and diffusion [54, 7].

In the case of GPCRs, similarities can be observed across different cell types; improved understanding, and manipulation in one case possibly allowing similar work across several [54]. Early theories of cell membrane structure were also widely applicable across types [145]. Several representative models are likely needed: it has also been noted that the structures of the cell membrane are extremely dynamic [57]. Therefore, we could improve GPCR and G protein modelling by extending the framework to identify the heterogeneity of dynamic behaviour and micro domains.

### Observable patterns

Prior to our involvement, initial work described quantitative analysis via Hidden Markov Models and mean square displacement. Receptors had several different diffusion patterns, 11% virtually immobile, 38% confined, 45% simple Brownian and 6% directed [54]. Further, behavioural states were classified for both receptors and G proteins into four groups, a gradient from virtually immobile to fast diffusing. State change to more immobile patterns was particularly prevalent, it was suggested that the slower states were caused by compartmental trapping; semi permissive boundaries retaining entities much like a net. GPCR and G protein co-localization within compartmental areas lasted around a second.

Under ordinary conditions, proteins and lipids within the cell membrane have been observed to undergo *hop-diffusion*, i.e. jumping between membrane compartments within which they are trapped [6, 55, 56, 57]. A *fence and picket* model of the plasma membrane suggests that it is subdivided by actin-based skeleton *fences* and transmembrane protein *pickets*. The model was used to explain compartmentalization of nanodomains and barriers affecting free diffusion [56, 58, 9]. Such a model might enhance co-localisation and in turn biochemical reactions such as G protein receptor binding by confinement of GPCRs within these spaces [59, 6].

Areas of confinement appear on movement heatmaps as bright zones, termed 'hot-zones' (Figure 3.2). A confinement phenomenon was observed and further supported by overlaying cytoskeletal images with movement maps [54]. Actin fibres, microtubules and CCPs were also suggested as possible contributing factors for construction of the GPCR signalling nanodomains

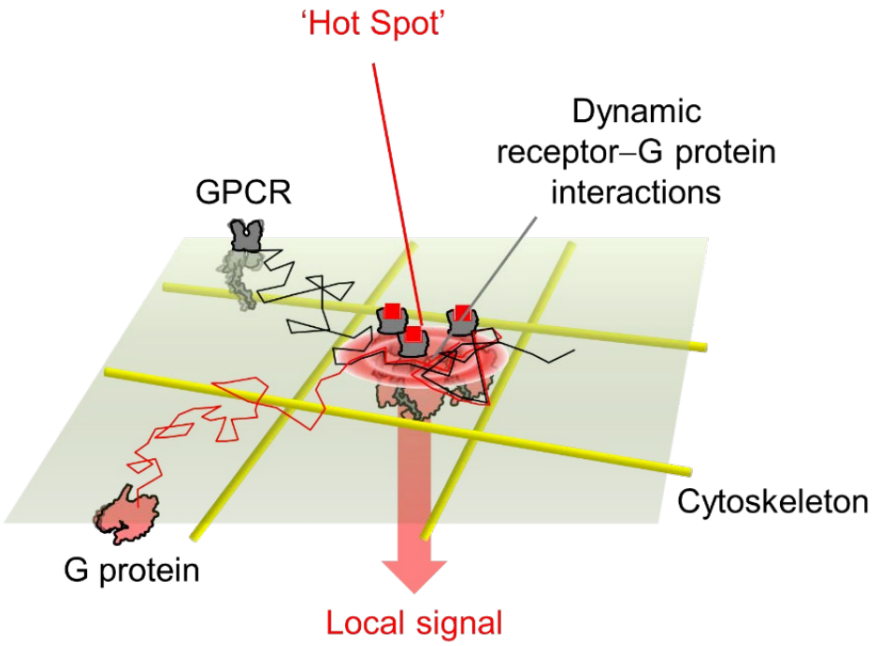



Figure 3.2: GPCR and G protein diffusion is often influenced by the cytoskeleton; dynamic interaction between the two molecules can happen in distinct domains created by the cytoskeletal mesh referred to as hot-zones. These hot-zones can increase G protein activation and in turn inter-cellular signaling [6].

previously postulated [54, 143, 144]. By promoting GPCR interaction through forced colocalization fast local signalling occurs at the expense of amplification, allowing mass control of signalling across the surface of a cell via dynamic nanodomains [6].

## 3.2 Methodology

### 3.2.1 GPCR and G proteins

We apply an improved framework from the initial cancer investigation to develop specific representative models for the GPCR and G protein system. Comparison is made between real and model results via quantitative measures such as population size over time and meta data representations such as movement heatmaps described previously 2. Insight is derived by observing, quantifying and interpreting the visualisation of real-world *in vitro* data, model design and mechanistic hypothesis exploration 1.

### GPCR and G protein video data

The GPCR video data comes from Calebiro et al [54] whose team *'visualized individual receptors and G proteins at the surface of living cells with high spatial (around 20nm) and temporal (around 30ms) resolution [146]'*, *in vitro* observation. At this size, it is possible to observe the movement of individual GPCR and G protein, therefore allowing interactions such as co-localization can be identified over time. It is also possible to examine environmental effects by assessing repeated changes in entity movement at specific locations across the system.

Due to the generalization of the framework, scale matters less than tracking accuracy; a population of physically large entities will be represented by scaling relative to one another and their environment. Therefore, the accuracy, length of time and number of tracks dictates our ability to apply our model from a cancer to GPCR and G protein system. The GPCR data set contains 4000-7000 tracks represented as separate entities over 400 time points.

Within the framework, tracks are converted to vectors and speed at each time point by taking two positions and calculating the angle of the line between them and distance travelled per increment. We can handle missing original positions by converting to this vector representation; an entity with missing positions will continue to the next known position rather than being eliminated. Covered area is also expanded or reduced by multiplying speed by a uniform ratio; a transformation of input space to desired visualisation space. In this case, the chosen area was reasonably arbitrary at 1500px but can be expanded for better image production at the expense of computational overhead.

For the paired GPCR (C1) and G protein (C2) tracks two channels were used to clarify instances of co-localization and population identification. There may be some small misalignment across sets prior to framework input caused by recording equipment misalignment. We used the data sets labelled 641-646 and 679-682 with complete C1 and C2 track sets, all without significant changes in environmental conditions or treatment, referential continuity [54].

### Novel imaging techniques

When attempting to observe motion at such a small scale, accuracy, consistency and differentiation can become significant problems. There is also an extremely limited time frame to consider interactions. It is possible to use fluorescence resonance energy transfer; an innovative microscopy approach that can make low scale observation reasonable [6, 147, 148, 149]. However, total internal reflection fluorescence (TIRF) microscopy was applied by the Calebiro's group, using total internal reflection to illuminate cells with the ability to differentiate if they are at least 20nm apart and recording an image once every 30ms [150, 151, 6].

Fluorescent single-molecule video images can be used to track GPCR and G protein molecules as they move within a cell membrane [146]. By using two colour imaging, multiple populations can be differentiated and observed within a single environment. In the set we are using, interactions between GPCR and G protein were observed with reasonable levels of labelling accuracy for each set, 'labelling efficiencies were approximately 90% (extracellular) and 80% (intracellular); non-specific labelling was below 1%' [54]. Population tracking was performed once particle and receptor colour differentiated videos were obtained. A MATLAB based tracking library called 'u-track' was used by the Calebiro group to define sets of tracks for subsequent analysis, we were provided with these and able to use them as input for our framework [54].

### Prepossessing GPCR and G protein data to identify and remove artificial bias

Identifying artificial and misleading patterns or 'bias' is an important part of model analysis and real-world digitisation across data sets. As such, we initially developed a preprocessing pipeline for the GPCR and G protein system. We incrementally applied techniques to isolate and identify possible artificial bias, subsequently filtering the effects or removing the source. A robust and consistent bias identification process is essential for the definition of any model and enable meaningful comparisons.

**Identifying trends and possible bias** To calibrate and ensure the we can effectively apply the framework to a new biological system, we use test data sets from a well understood situation (Figure 3.3). Initial observations are compared to known and expected patterns, differences might

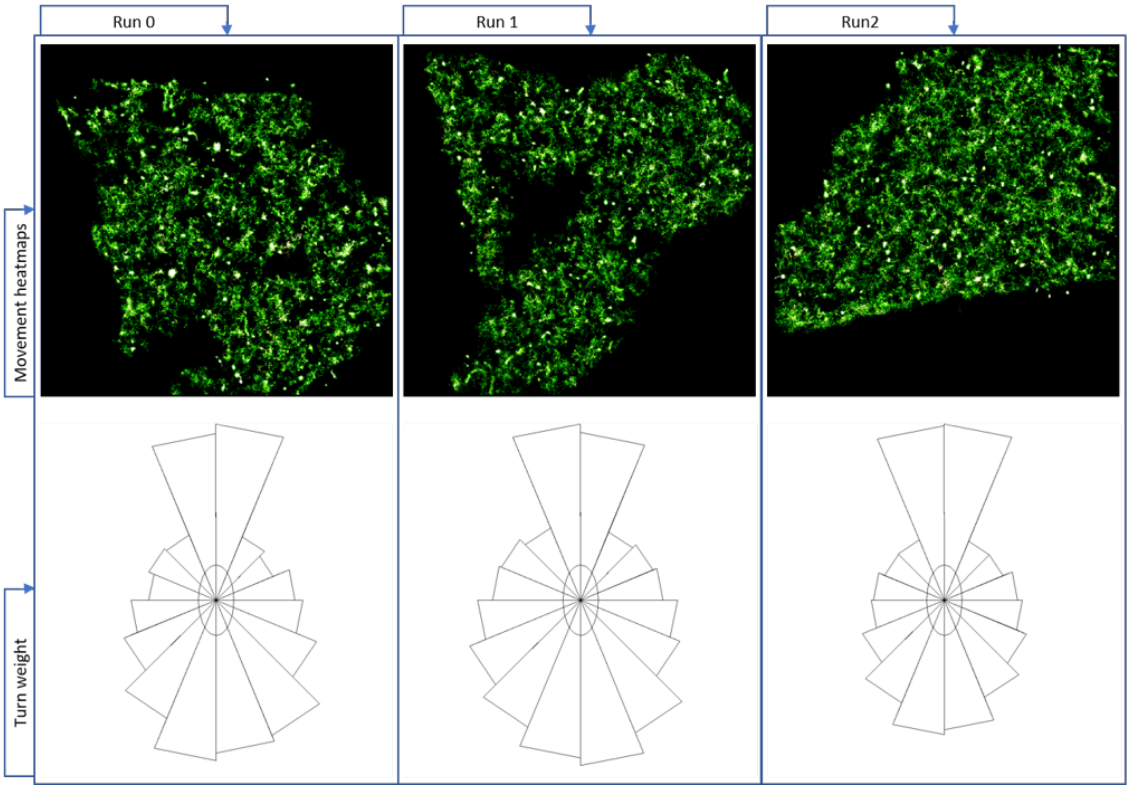



Figure 3.3: Movement heatmaps (2.2.2) of example data sets with their correlated turn trend analysis after initial framework real data digitisation analysis and still present strong artificial forward bias.

be caused by either model limitation or artificial bias. For example in the case of GPCR and G protein movement, we expect heatmaps with hot-zones, cell boundaries and general Brownian motion [56, 58]. In the case of GPCR and G protein turn distribution showed strong forward and reverse trends, neither was expected in general Brownian motion. Therefore, we next investigated framework interpretation of input data and entity tracking, internal and external processes.

**Framework Interpretation bias** To asses possible framework interpretation bias we model a range of known behaviours that we can simulate and compare the resultant patterns with an input format identical to the real-world data. For example, Brownian motion is expected for general GPCR and G protein movement and to produce random directional preference over time. Therefore we generate simulated Brownian motion within the framework and compare it with test data (Figure 3.4 C).

When designing a filter for possible artificial bias a strong causal narrative explanation is im-

portant to address the source rather than the result. In the GPCR and G protein set, initial forward preference may have been the result of actual micro turns; when a tracked entity travels in a straight line, its position can be recorded as multiple micro turns. Therefore, any angle turns below 0.05 degrees can be filtered out to remove most of the bias. However, this is a filter on results based upon an assumption and therefore we further refined our hypothesis: single point tracks and completely stationary entities were the real cause of strong forward preference. Upon initialisation entities were checked for a difference between their previous and new direction, and starting vectors were set to the same direction as their first turn so the returned first check was always 0. When combined with a highly fragmented set of trajectories, each loss of tracking results in a new track and increase in the artificial forward bias registered. Zero and single frame vector stats were removed, resulting in a representative turn diagram with only rear bias for real-world system(Figure 3.4 B).

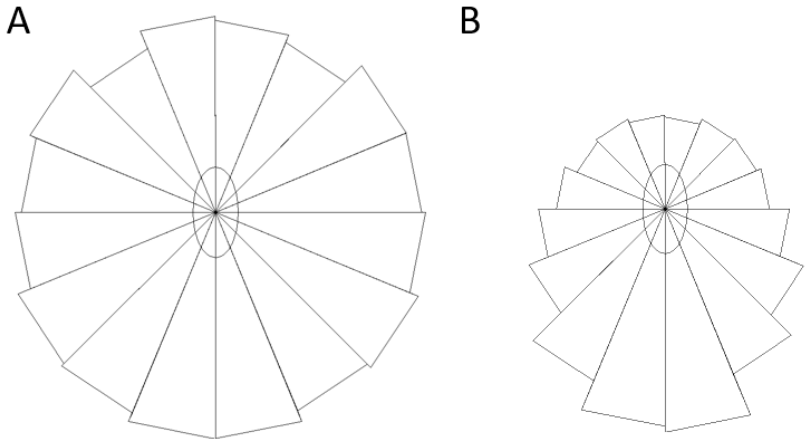



Figure 3.4: Simulated Brownian motion turn preferences as a relative turn diagram (A) for comparison with *in vitro* receptor and G protein turn preference data (B).

**Tracking bias** To identify possible bias in the tracking step, a well understood system is required; if we know the underlying movement patterns and perform tracking before the final analysis, then the loss or change of information should represent the effects of tracking. To assess tracking effects, manually distorted movement patterns were generated, i.e. test simulations with known bias to compare pre- and post- tracking turn trends. The magnitude of the biases, size of population and movement distance were comparable with the real-world system. Known test simulation videos are then tracked to check for artificial change.

**Including possible confounding conditions for tracking checks** Population size and travel distance may also affect tracking results. Other possibly strong confounding variables for any tracking algorithm are overlapping and stationary entities. The framework was used to generate videos of moving populations with known turn preferences and biases with random interactions followed by localization, random stopping and then random stopping when overlapping.

In the case of GPCR and G protein tracking with confounding variables, no rear bias was created by the tracking step. However, there was a minor trend modification where patterns were obscured rather than removed or added. Also, simulated stationary entities were interpreted correctly, any movement back and forth in a small space is likely a real behaviour. Tracking followed a more conservative approach to pattern identification and allows us to observe the existence of many clear trends and patterns.

### Smoothing

One of the main focuses of analysis for the GPCR and G protein system is population arrangement and interaction with micro domains. Visualization of movement over time via various heatmaps is a central part of our pattern-oriented approach to population definition (Figure 3.5). Within a heatmap values are assigned to a tracking grid as entities move within an area and then visualized on a colour gradient, the brighter an area the higher the movement value: the brightest areas indicate highest concentrations of movement and entity presence.

When extrapolating vectors and speed from sets of positional data each location in space sits on a continuum of time points. When we expand the covered area from the input set to our larger grid, the distance travelled between each step, i.e. speed, of entities can increase to the point of jumping. This can be fixed by introducing intermediate time steps to smooth the observations (Figure 3.5). The intermediate time points added enable entities to make smaller jumps that can still be visualized and tracked. Added time steps don't increase the total distance travelled but they do assume straight line movement between the known input positional points. This may become an issue if tracking is very fragmented and starting time points disparate. In the

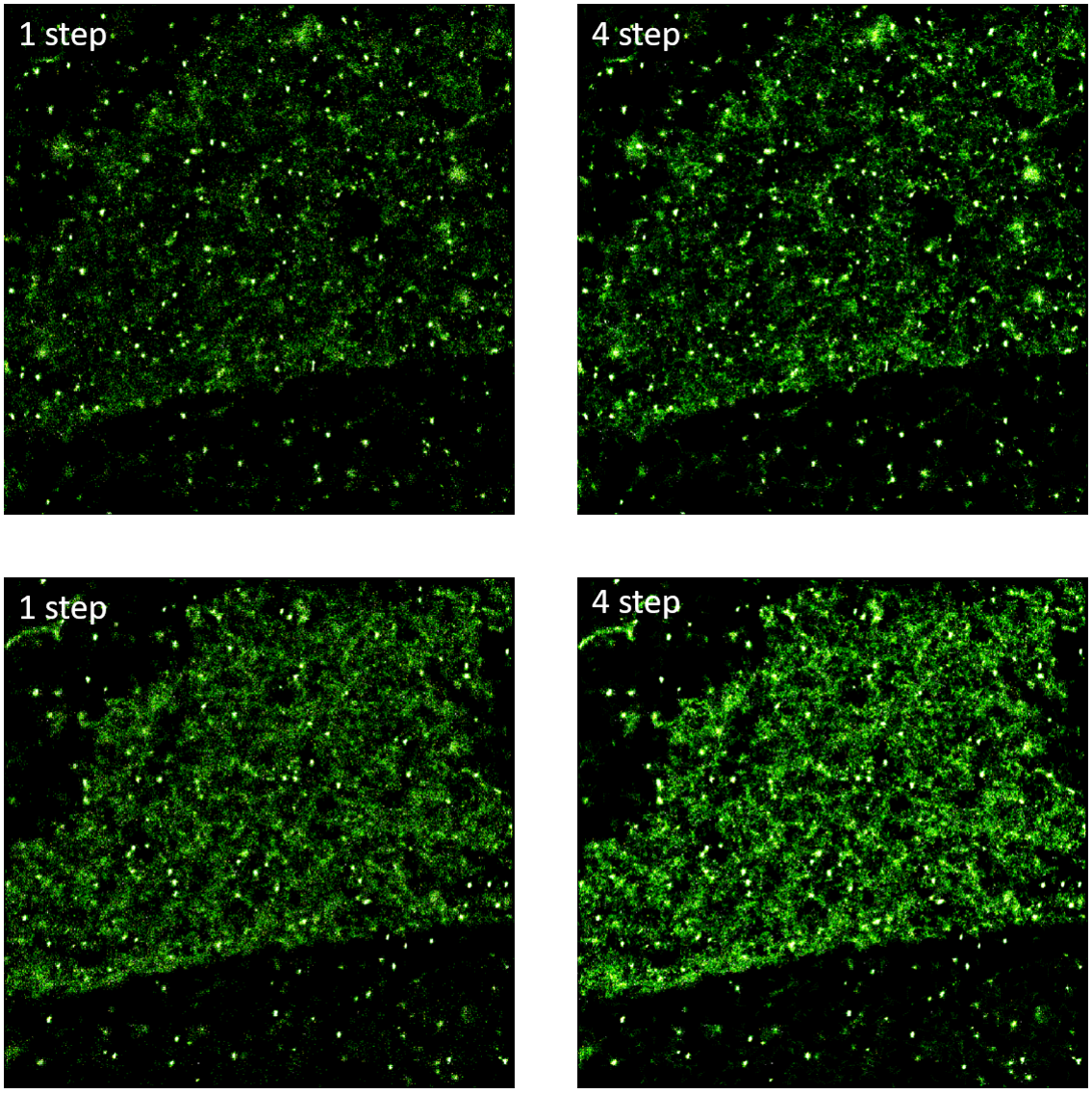



Figure 3.5: Cumulative movement heatmaps (2.2.2) of the accumulation of motion over the entirety of a positional data set, channel 1 (C1) above and channel 2 (C2) below. For each time point given in data set 643 positions are artificially separated by one or four smaller jump steps to smooth and decompress movement patterns in the resultant images.

case of the GPCR and G protein data, time points are very close and the measured area very small: when we expand this area in the framework to an observable size we add intermediary jump steps for a clear image. Specific hot-zone shape is also important enough that tracks with gaps make proper identification more difficult so conversion to our vector based representation is necessary ( 2.2.1). We record movement on 4 step sub increments for the extrapolated positions.

#### Post framework

We can use the framework to generate heatmaps and other visual meta data representations. Further comparison can also be made by using framework output as an input for external tools that focus on highlighting or separating sub patterns, (Figure 3.13). For example, here we overlay images utilizing the luminosity to RGB overlay process in *Fiji* [48, 62] to show several disparate images overlaid for contrast. C1 and C2 population movement maps are overlaid and differentiated with colour, with intensity still indicating movement density over time.

Population placement heatmaps can be created for the data snapshot time points, in this case every 50 increments (Figure 3.14,3.15). By using a courser grid of large cells, we can gain a visual representation of general entity placement. Different time points can then be overlaid with a final movement heatmaps in *Fiji* [48, 62] to give a view of population placement across time and relative preponderance of patterns and environmental effects in an area. Population placement heatmap intensity is relative to the number of population members in an area; across time points we should expect consistent bright areas to indicate a large sustained group presence in the area. Therefore, differentiating between single stationary and groups of trapped population members should be easier.

#### Model hyperparameters

By modelling possible explanatory phenomena, we will attempt to support the creation of hot-zones and investigate the localized movement starvation phenomena. Across the models hyperparameters pertain to inclusion of Brownian movement, population growth and decline, restrictive boundaries and attractive zones. From model to model hyperparameters may be included and modified by sub parameters such as growth rate or attractive zone size.

Brownian diffusion differs greatly from the directed movement of path following cancerous cells. However, our generalised approach allowed reasonable modification and application of existing representations. In this case Brownian diffusion was represented as turning at every time point with completely random direction and normally distributed jump length. Therefore, to create diffusion representation, we define an entity within the model with 100% chance to turn (it can also select an angle of 0) and even directional preference at each time point with the addition of a normally distributed jump distance modifier.

Population trends for entity removal or addition are accomplished by random chance for the effect per entity per time point. Restrictive boundaries are simulated by inversion of the attractive curves used previously. Chance to attach to a strand becomes the chance to deflect and can vary per portion of applied strands. Attractive areas are accomplished with a localised entity direction modification, they make a Brownian like hop but with a modification towards the attractive centre when within its catchment zone.

| Hyperparameters | Description |
|---|---|
| **Movement type** | Describes the logic used for entity directional selection, Brownian at each time point selects a random direction and jumps a normally distributed distance. |
| **Iterations** | The number of steps in any given model run. |
| **Entity number** | The number of starting entities in a model run. |
| **Cubic curves** | Whether attractive or deflective curves are applied within the model. |
| **Deflective curves** | Whether cubic curves within the environment deflect movement of entities that intersect them. |
| **Curve attraction/deflection** | Upon intersecting a curve, an entities average chance to attach, deflect or continue along at each increment out of 100000. |
| **Number of cubic curves** | The number of included curves total. |
| **Size of cubic curves** | Curve width, less important for deflective curves but can impact ease of meta pattern observation. |
| **Attractive zones** | An area of the environment where entity trajectories are modified towards the centre of the area. |
| **Attractive zone size** | The catchment zone of an attractive area, how far out from the centre entity tracks are modified. |
| **Attractive zone strength** | Determines the strength of track modification towards the centre. |
| **Attractive zone falloff** | Defines a gradient to track modification, weakening as entities gain distance from an attractive zone centre. |
| **Attractive zone count** | The number of applied and present attractive zones. |
| **Attractive zone eye size** | A movement permissive area (where trajectories are unmodified) within an attractive zone, this determines the size. |

Table 3.1: Key hyperparameters for the chapter with a short description. A full list can be found with the code repository https://github.com/Benkwitz-Bedford/AB-FABS

| Hyperparameters | 3.18 A | 3.18 B | 3.19 A | 3.19 B | 3.19 C | 3.20 | 3.21 | 3.22 A | 3.22 B | 3.22 C | 3.22 D | 3.23 A | 3.23 B | 3.23 C | 3.23 D | 3.24 | 3.25 A | 3.25 B | 3.25 C | 3.25 D | 3.26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Movement type** | Brownian | | | | | | | | | | | | | | | | | | | | |
| **Iterations** | 400 | | | | | | | | | | | | | | | | | | | | |
| **Entity number** | 400 | 450 | 400 | | | | | | | | | | | | | | | | | | |
| **Cubic curves** | FALSE | | | | | | | | | | | | | | | | TRUE | | | | |
| **Deflective curves** | FALSE | | | | | | | | | | | | | | | | TRUE | | | | |
| **Curve attraction/deflection** | 90000 | | | | | | | | | | | | | | | | | 80000 | | 40000 | |
| **Number of cubic curves** | 20 | | | | | | | | | | | | | | | | | | | | |
| **Size of cubic curves** | 5 | | | | | | | | | | | | | | | | | | | | |
| **Attractive zones** | FALSE | | | | | | | TRUE | | | | | | | | | FALSE | | | | TRUE |
| **Attractive zone size** | 20 | | | | | | | | | | | | | | | | | | | | |
| **Attractive zone strength** | 100 | | | | | | | 80 | 60 | 40 | 10 | 60 | 40 | | 5 | | 100 | | | | 10 |
| **Attractive zone falloff** | 0 | | | | | | | | | | | | 20 | 0 | 1 | | 0 | | | | |
| **Attractive zone count** | 20 | | | | | | | | | | | | | | | | | | | | |
| **Attractive zone eye size** | 10 | | | | | | | 0 | | | | 10 | 0 | 20 | 10 | 0 | 10 | | | | 0 |

Table 3.2: Key hyperparameter values sorted by figure for the chapter. A full list can be found with the code repository https://github.com/Benkwitz-Bedford/AB-FABS

## 3.3 Results

### 3.3.1 GPCR and G proteins

#### Population size, turns and travel distance

**General** We can initially start to understand the GPCR and G protein population groups, landscape, and movement patterns with very general metrics. Interesting observations can be made by comparing the behaviour and interactions of GPCRs (C1) and G proteins (C2) on the cell surface. Set number (641-682) is a unique key for result identification purposes, not indicative. Initially we recorded metric and metadata at 50 increment intervals across all C1 and C2 runs to compare positional grouping within the sets and compare between. We can observe population activity, the emergence of hot-zones and their relative effect upon entities over time by looking at population size, distance and turn rates (Figures 3.6,3.7).

In the C1 sets we observe consistent decline in the population size. C2 also shows some decline but with more variance across sets, even sometimes including growth (Figures 3.6,3.7). The active population sizes within each C1/C2 are reasonably consistent but with a greater spread in C2 (Figures 3.6,3.7). Across the sets both C1 and C2 remain between 200 and 600 active entities with the majority around 400.

For both C1 and C2 sets, total turn numbers are directly dictated by population size; indicating in both cases entities turn at every available opportunity. The general distance travelled for both C1 and C2 sets rises and falls with population size but shows some variance. Some clear differences can be observed in the travel distance of outlying sets, for set 644 C1, and to a lesser extent 646, C1 travel distance is a great deal larger 180000-210000 than the C2 counterparts 160000 and the rest of the C1 set (Figures 3.6,3.7). When we also look at the turn rate, set 644 is still a clear outlier in particular in the C2 channel, most other sets are within a similar range of 6000 turns per population per set. Channel C2 of set 680 had shown the lowest distance travelled and is more pronounced in a lower turn rate as well. The confinement model could provide an explanation in tying initial movement to greater chance of ending in a restricted location, in the case of 680 C2 lower motility leads to lower bouncing within a restricted space and effect upon turn preference.

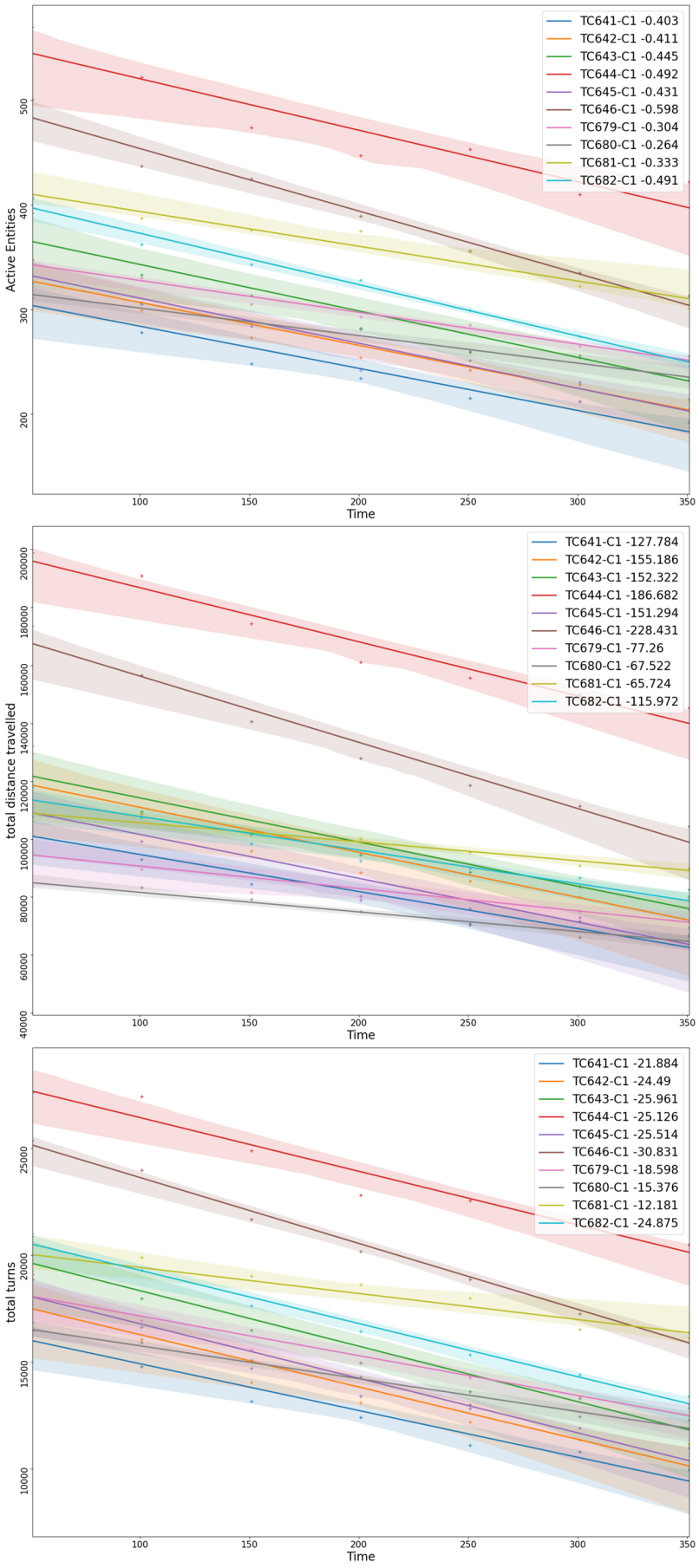



Figure 3.6: General quantifiable measures taken every 50 time increments from the 400 total. Active population size, distance travelled total and turns taken by the populations of C1. Each line represents a different data input, the number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

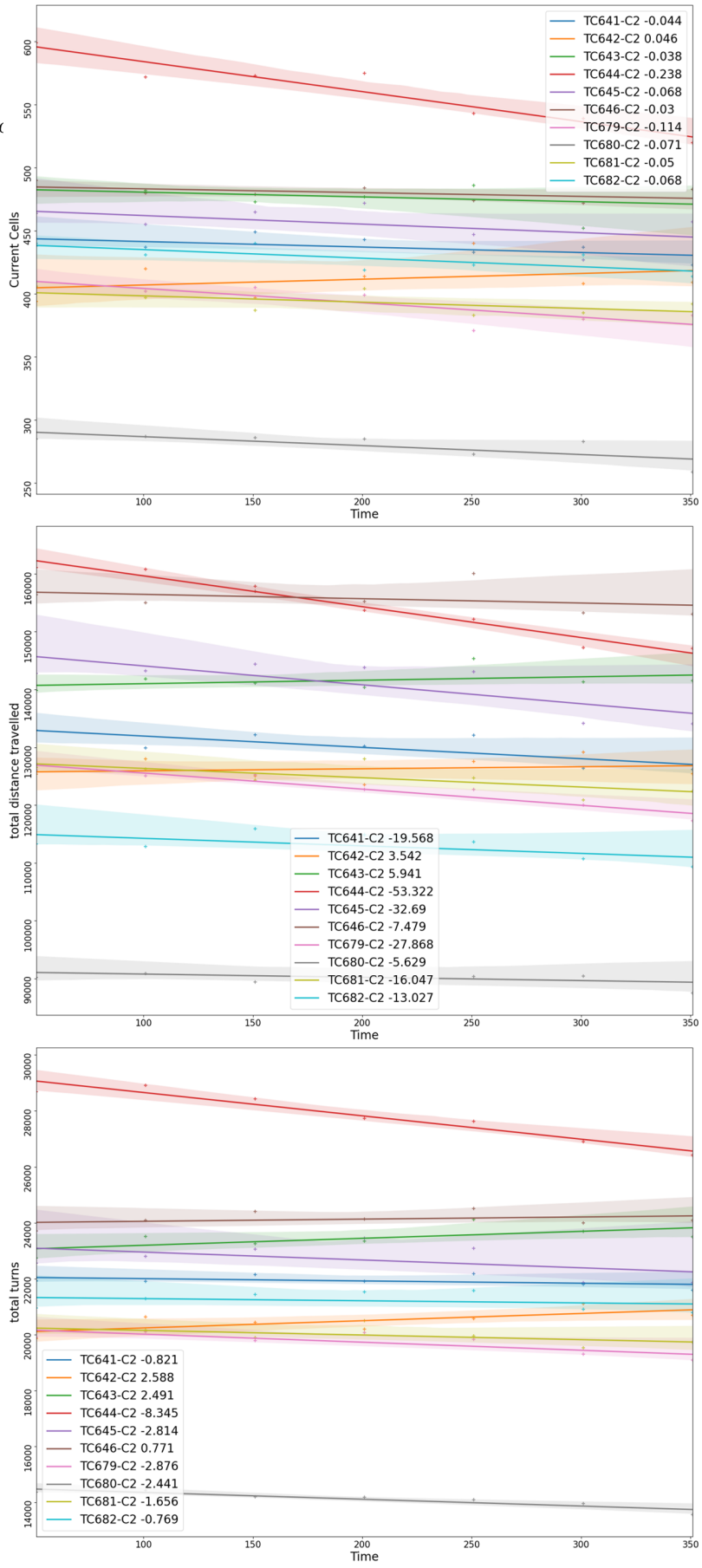



Figure 3.7: General quantifiable measures taken every 50 time increments from the 400 total. Active population size, distance travelled total and turns taken by the populations of C2. Each line represents a different data input, the number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

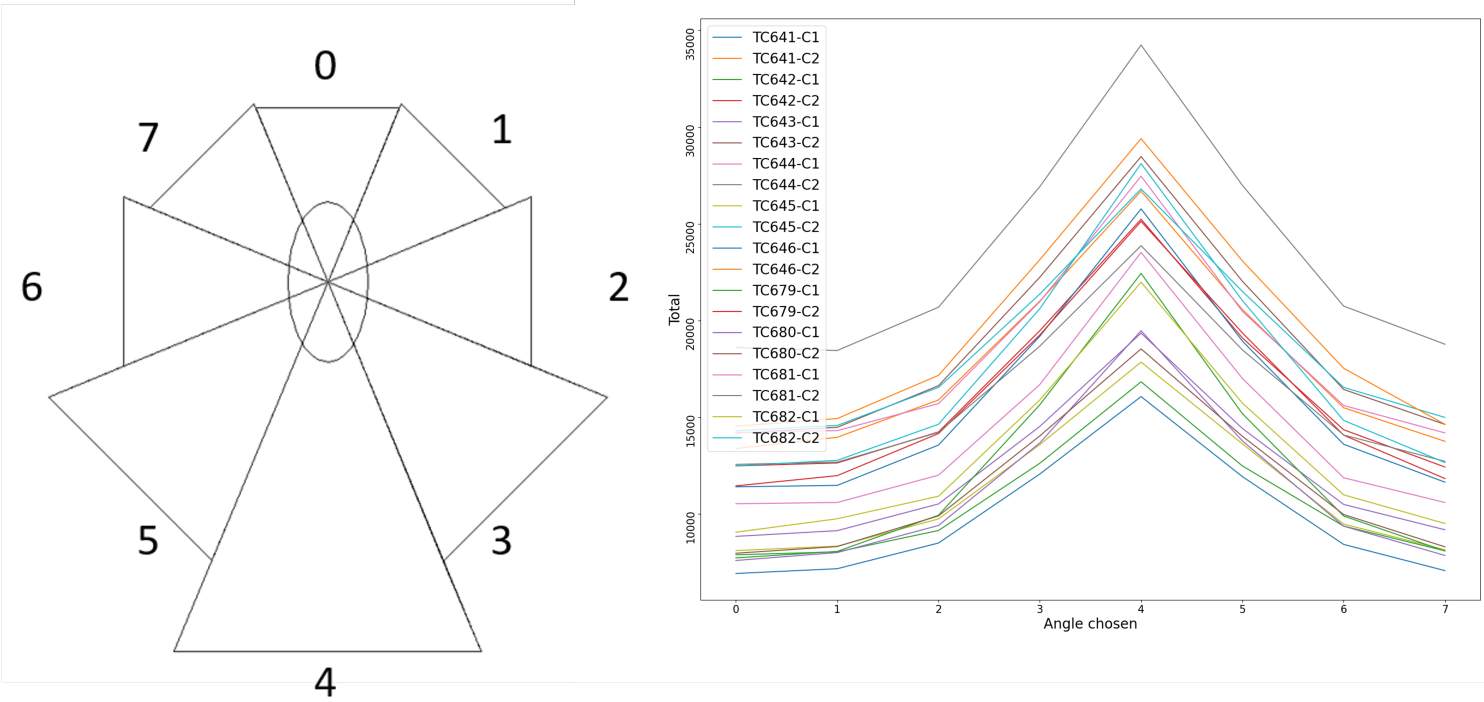



Figure 3.8: The general chosen turn trends of a population or model population over time can be represented as a weighted turn circle. A larger segment indicates a higher relative proportion of turns chosen relative to north or a 0 angle in this diagram (Left). We can also flatten the diagram out into a graph representation at any time point (Right). In this case over the entirety of a data set and population the x-axis indicates the angle box from a turn diagram.

**Turn trends** At each time point an entity has a direction, speed and position, and we can identify a change in direction as a turn. We measure turn preference by taking the angle between an old and new directions. In the case of GPCR and G protein movement, Brownian directional motion is expected to be the norm. Any clear bias in turn direction could therefore be indicative of environmental effects.

All the given GPCR and G protein data sets C1 and C2 show a rough Gaussian distribution based around a rear turn bias (Figure3.8). A rear bias indicates that at any turn point there is a higher chance that population members will reverse direction.

While the overall pattern is the same across all real-world sets we can observe greater similar-

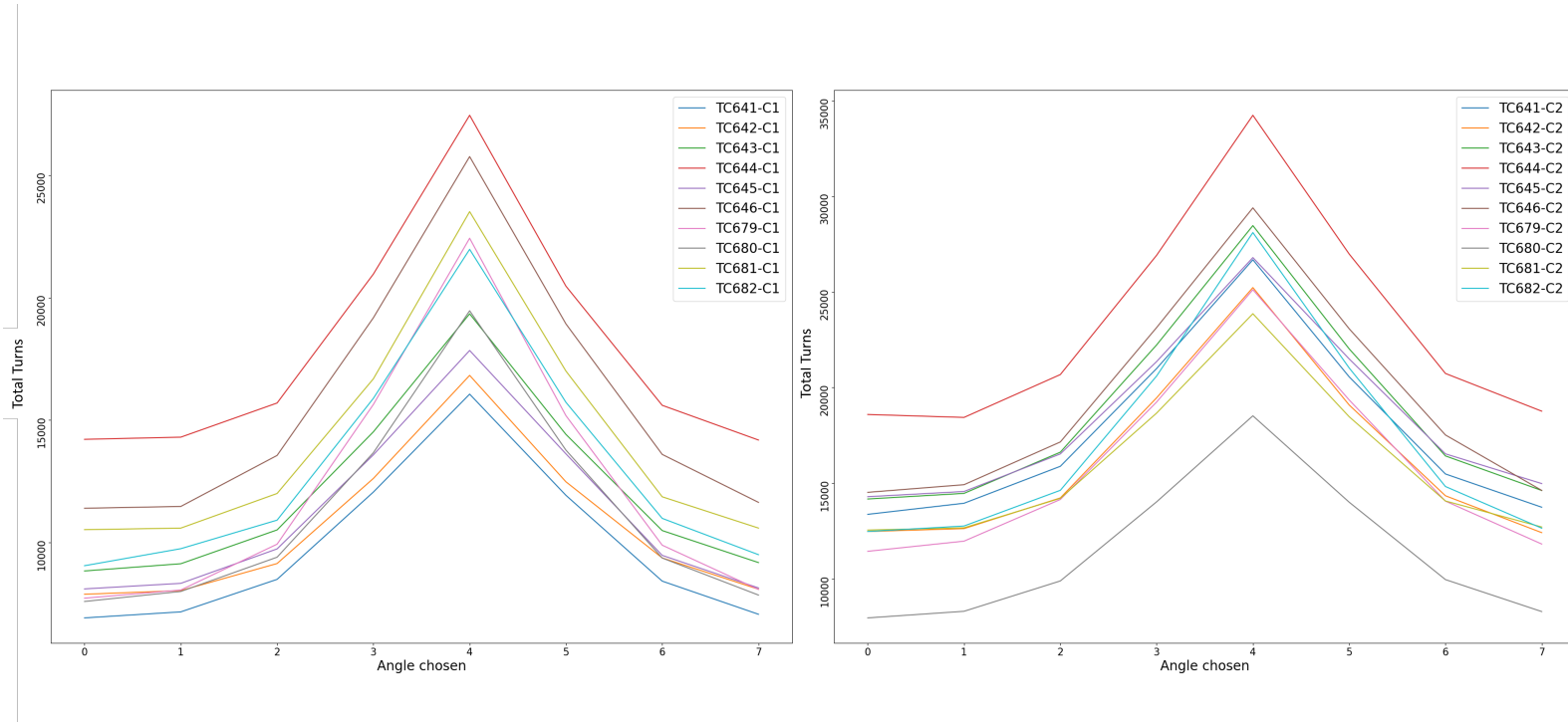



Figure 3.9: By tracking the number and angle of turns made across a data set we can represent preference across a population relative to their previous direction 3.8. 1) all the data for C1 and C2 sets with A-totals and B-percentage of total for the run. Turns 2) for the C1 data sets and 3) the C2 data sets.

ity between C1 trends than C2 in the averaged turn preference graphs (Figure 3.9B). While the amounts differ widely across real-world sets, the rear biased pattern is consistent. It would seem from these results that movement is likely reactive not directed or following such as in cancer movement; a leading phenomenon would cause greater 0 bias. As with raw number of turns, the range of C2 directed turns is greater because of the outlying sets 644 and 680 but the majority fall within a similar distribution to C1.

The number of directed turns (Figure 3.9) in real-world set 644 and 680 stand out in both C1 and C2 channels with a lower number of turns in all directions than other sets. 644 and 680 show the rear biased meta pattern as other sets; 680 shows a slightly greater proportional rear bias while 644 is below the average in C1 and C2 sets (Figure 3.9). Therefore, it seems unlikely that higher motility led simply to a higher confined rate and rear biased movement, it may suggest a threshold value for confined movement.

| | Time | zero | one | two | three | four | five | six | seven | s.deviation | mean | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 51 | 9.79% | 10.05% | 11.01% | 15.17% | 18.57% | 14.95% | 10.97% | 9.49% | 458.74 | 1729.88 | 13839 |
| | 101 | 8.98% | 9.21% | 11.00% | 14.93% | 20.23% | 14.98% | 11.25% | 9.42% | 466.35 | 1481.75 | 11854 |
| | 151 | 8.94% | 9.07% | 11.09% | 15.86% | 20.45% | 15.04% | 10.51% | 9.04% | 437.50 | 1303.13 | 10425 |
| | 201 | 8.50% | 9.33% | 11.16% | 15.16% | 20.85% | 15.82% | 10.41% | 8.76% | 429.25 | 1226.63 | 9813 |
| | 251 | 8.53% | 8.99% | 11.01% | 15.06% | 21.44% | 15.21% | 10.87% | 8.89% | 390.23 | 1091.38 | 8731 |
| | 301 | 8.15% | 8.54% | 10.26% | 16.18% | 21.49% | 15.55% | 10.87% | 8.96% | 401.00 | 1053.63 | 8429 |
| | 351 | 8.63% | 8.81% | 11.10% | 16.05% | 21.24% | 15.27% | 10.63% | 8.26% | 360.85 | 978.75 | 7830 |
| | 400 | 8.65% | 8.65% | 9.99% | 15.52% | 21.93% | 15.70% | 10.48% | 9.07% | 341.72 | 893.13 | 7145 |
| C2 | 51 | 9.29% | 9.84% | 11.49% | 15.02% | 18.88% | 14.12% | 11.23% | 10.14% | 579.97 | 2211.13 | 17689 |
| | 101 | 9.43% | 10.24% | 11.46% | 14.38% | 19.02% | 14.67% | 10.96% | 9.86% | 569.59 | 2165.00 | 17320 |
| | 151 | 9.81% | 10.10% | 11.15% | 14.71% | 18.98% | 14.73% | 10.94% | 9.58% | 596.35 | 2238.50 | 17908 |
| | 201 | 9.17% | 10.11% | 11.16% | 15.10% | 19.20% | 14.64% | 10.96% | 9.64% | 621.38 | 2223.63 | 17789 |
| | 251 | 9.60% | 10.11% | 11.21% | 14.57% | 19.01% | 14.58% | 11.06% | 9.87% | 588.46 | 2235.38 | 17883 |
| | 301 | 9.72% | 9.75% | 11.35% | 14.69% | 18.63% | 14.96% | 11.09% | 9.81% | 566.79 | 2173.63 | 17389 |
| | 351 | 9.72% | 9.66% | 11.00% | 15.37% | 19.14% | 14.40% | 10.97% | 9.73% | 617.56 | 2227.75 | 17822 |
| | 400 | 9.29% | 9.52% | 11.53% | 15.56% | 18.87% | 14.86% | 10.84% | 9.53% | 596.43 | 2112.25 | 16898 |

Table 3.3: Total turn choices made by GPCR and G protein set 641 C1 and C2 across time relative to the angles shown in turn diagrams; angles 0-7 are clockwise from north 3.8.

**Turn trend bias**

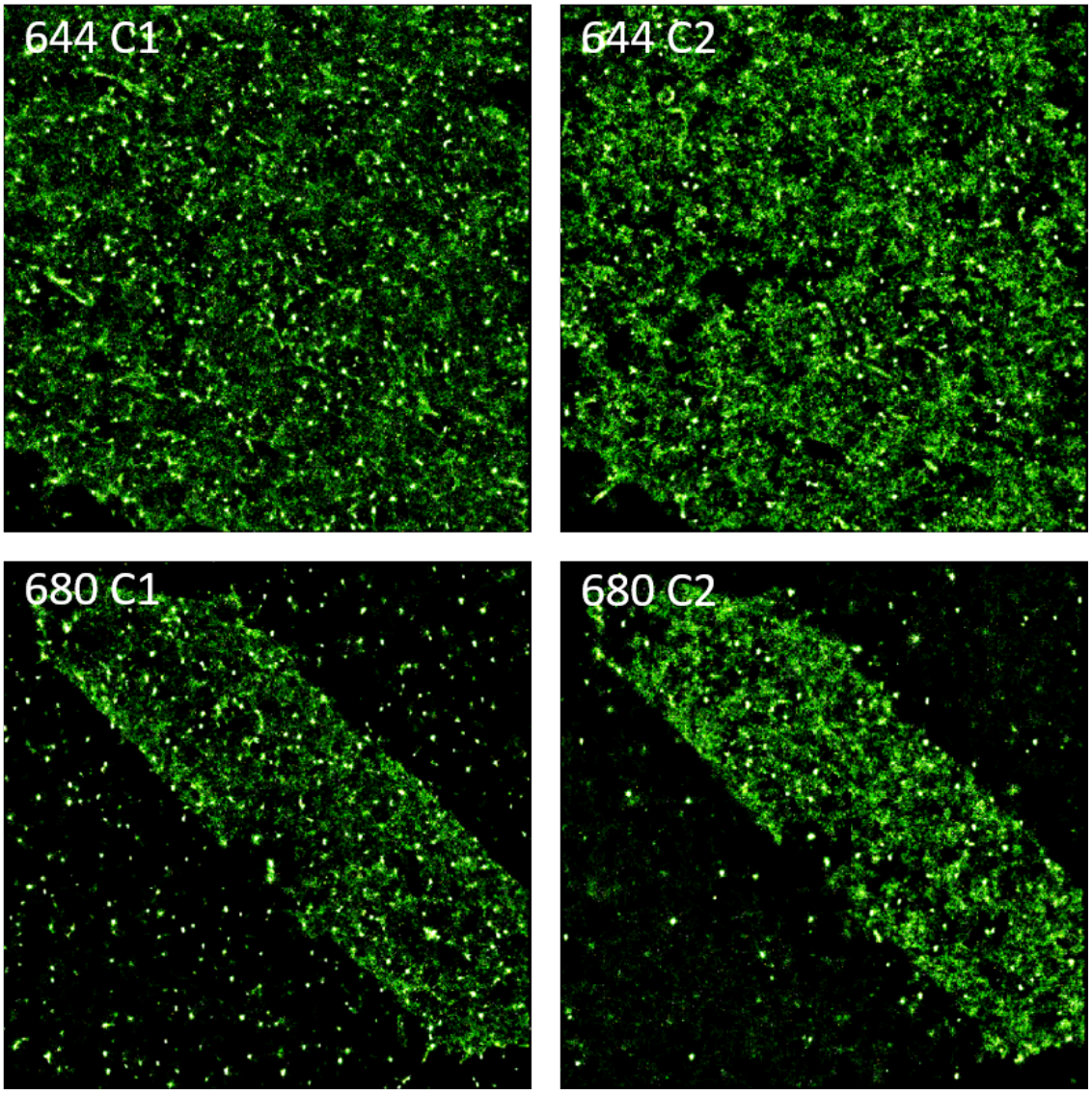



Figure 3.10: Identified by movement distance and turn number, outliers 644 and 680 C1 and C2 can be visualized as movement heatmaps (2.2.2) to see if the difference can be easily explained

We identified sets 644 and 680 as requiring further exploration based upon the patterns observed. Cell shape and accessible area for movement within sets 644 and 680 differs from that observed in other runs (Figure 3.10). Set 644 contains activity within a larger cell surface than the average and set 680 a much smaller area leading to higher and lower amounts of turns respectively. The surface area difference may also explain the low trend in 680 turn numbers and bias when comparing C1 and C2. Within this more confined space C1 has formed far more hot-zones than the C2 set.

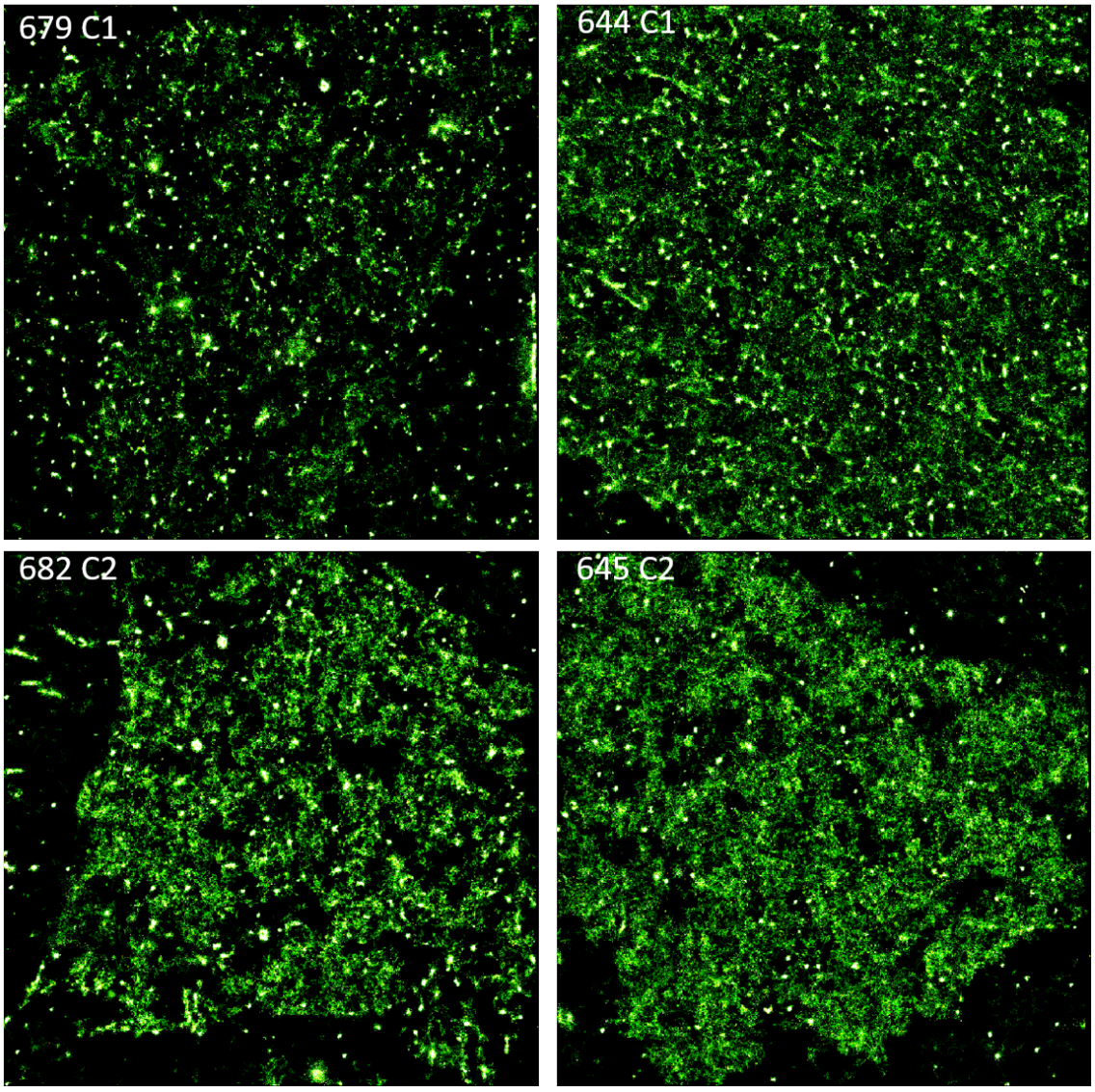



Figure 3.11: Population movement heatmaps (2.2.2) for outliers identified by differentiation of self-referential turn trends, i.e. the proportion dominance of rear turn bias (Figure 3.9B. 679 had the most pronounced C1 bias and 644 least, 682 the most pronounced C2 bias and 645 least.

We can further identify other potentially interesting outliers by observing noticeable differences across the quantifiable turn trend averages (Figure 3.11). We can observe that movement heatmaps suggest set 645 C2 has fewer and smaller hot-zones than set 682 C2, therefore, there may be a relation between hot-zones and rear biased movement. Comparison of sets 679 C1 and 644 C1, displays similar numbers of disparate small hot-zones with a greater number of large distinct hot-zones in 679, larger hot-zones may be more impactful in creating the bias. Therefore, in C2 rear turn selection bias is indicative of hot-zone interaction or immobility, in C1 it indicates hot-zone size. We may suggest that rear bias is closely associated to population and movement amount since there is clearly a lower general movement density, 679 and 682 having less members and movement than 644 and 645 respectively. 680 has the lowest population and number of turns with less exaggerated turn bias. What seems clear across both sets of outliers (Figure 3.10,3.11) is that general non hot-zone movement density correlates to the reduction of rear bias.

#### Tracks identified and visual trends

While quantitative values are useful, identifying patterns of movement over time can become extremely complex or lead to oversimplification of real phenomena. We use more qualitative visual comparison to identify or highlight patterns within the observed environments. In the case of GPCR and G protein we can identify and compare localized differences with overlap between hot-zones.

In the literature associated with specific GPCR and G protein, it was marked that populations seemed to be made up of 11% virtually immobile, 38% confined, 45% simple Brownian and 6% directed [54] behavioural types. Within heatmaps reproduced with the framework after input of the same data sets we identify similar patterns (Figure 3.12. Hot-zones can be seen across population sets, both C1 and C2 but with varying size. Larger hot-zones are more likely to be examples of restriction, smaller hot-zones may also be cases of restriction but more likely also represent immobile entities. Low back-and-forth motility leads to higher movement density in a more confined area with a lower associated population size requirement, a single entity can create a hot-zone area by remaining there rather than a group of entities traversing through one.

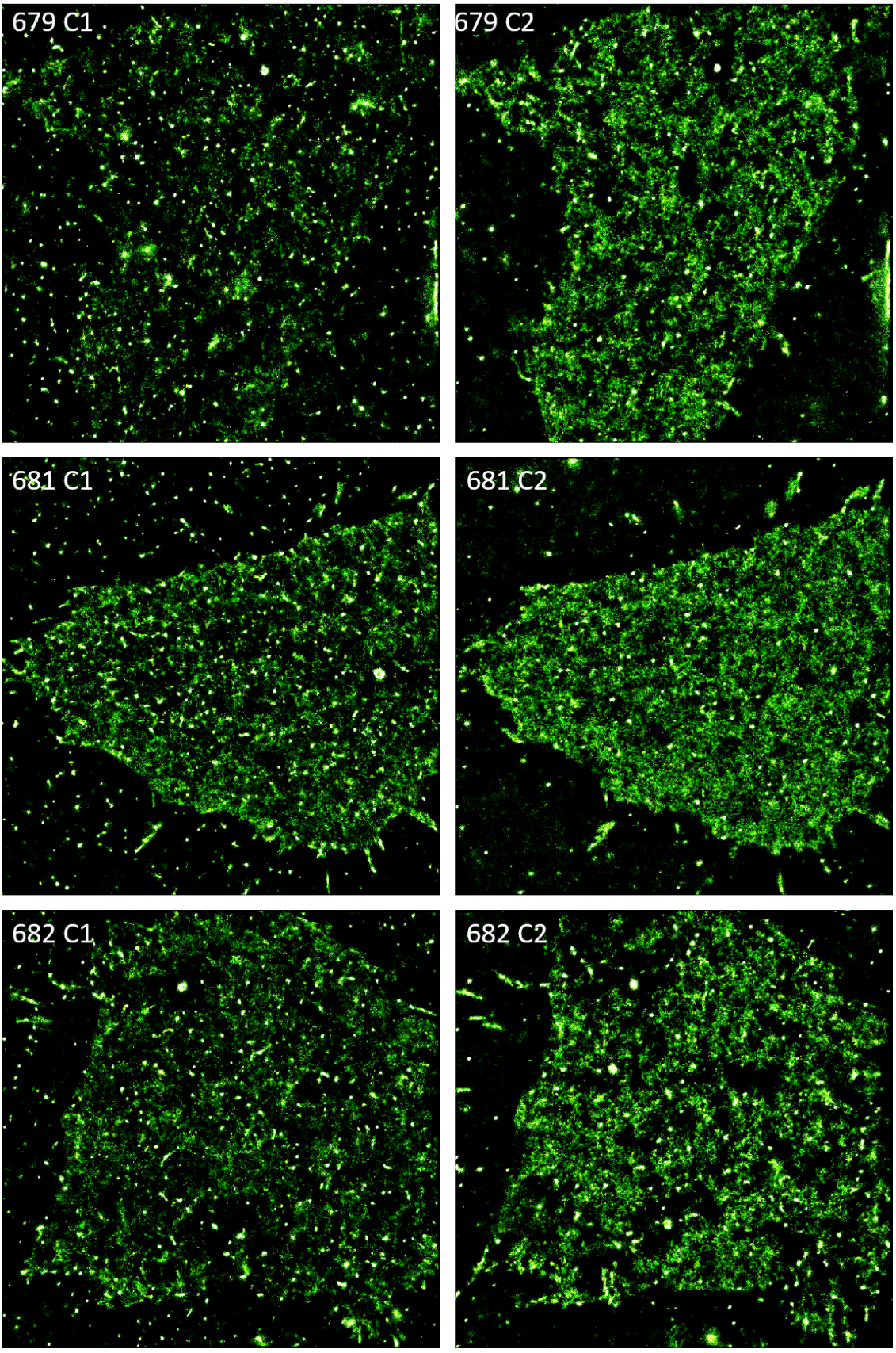



Figure 3.12: By using sets of C1 and C2 movement data over time a range of heatmaps (2.2.2) can be generated to visualize the behaviour of a population over time and their interaction with micro-environments. The brightness to activity relative gradient has a set top end to ensure that each image within the figures is easily comparable, each shade indicates the same value.

General Brownian random movement can be observed across C1 and C2. C2 shows predominantly uniformly higher movement while C1 creates large less cohesive movement hot-zones suggesting stronger compartmentalization. Near particularly large hot-zones, both C1 and C2 sets appear to display a reduction of localized general movement, a starvation effect. The same movement to heat ratio is applied across the entire population and each set, therefore, this starvation of local movement is not a simple bias of visual contrast. 682 C1 and C2 show several examples of this area starvation with C2 displaying clearer contrast with otherwise random general movement (Figure 3.12).

### Post framework input

**C1 and C2 overlay** We can attempt to visualize GPCR and G protein co-localization by overlaying the generated movement heatmaps for paired C1 and C2 sets of population. The larger hot-zones are often shared across C1 and C2 sets while few of the smaller zones were overlapping (Figure 3.13). Interestingly, some distinct non-overlapping larger hot-zones can also be observed, particularly in the case of set 643. There also appears to be some evidence of geographical disjunction between C1 and C2 movement zones; areas distinctly populated by one set but not the the other, this may be a visual bias due to greater clarity with reduced overlap. Set 646 shares some mutually disassociated adjacent movement channel like structures distinct from general overlapping movement that can also be observed with less emphasis in other sets.

Qualitative observation suggests that larger hot-zones are more reliably areas of co-localization with an associated likelihood of low motility as areas get smaller. Amount of co-localization also differs across real-world sets, as expected from circumstantial environmental change and stochasticity. However, hot-zone associated movement starvation remains. Some distinct movement starvation patterns appear to be channel specific restriction; areas where members of a specific set are restricted but not the other. In set 641 centre-left the red (C1) and green (C2) overlaid movement map highlights a large hot-zone displaying co-localization with no green movement around but some smaller red hot-zones (Figure 3.13). If small areas are low motility zones it might indicate a phenomenon similar to dense but viscous micro domains, precluding green inclusion and slowing red movement in the area by forcing slower or no local diffusion.

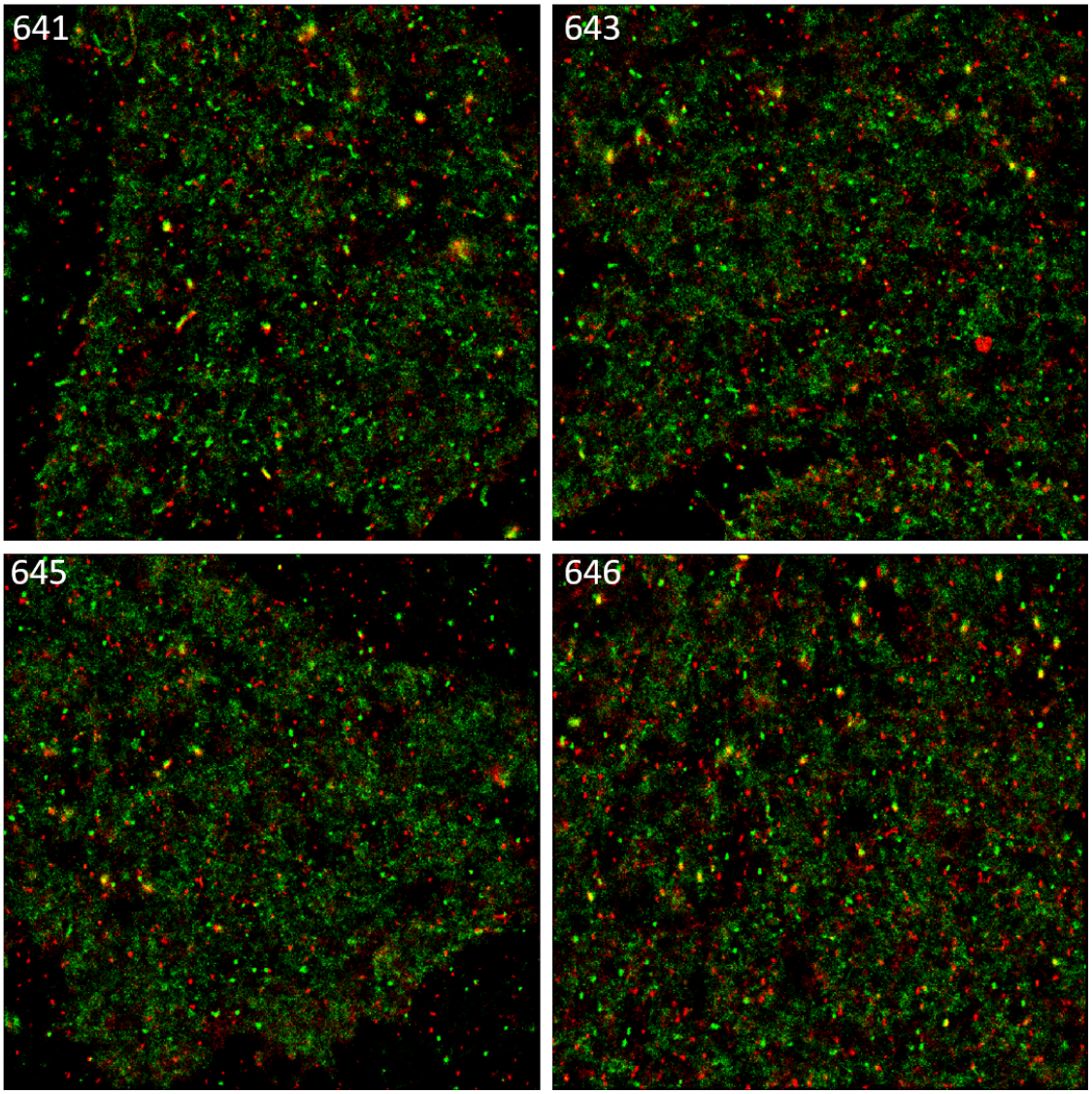



Figure 3.13: Population movement heatmaps (2.2.2) for 641, 643, 645 and 646 population data sets, C1 and C2 were overlaid for each. Red shows C1 preponderance in an area and green C2. By layering images in this manner, we can directly visualize co-localization of GPCR and G protein

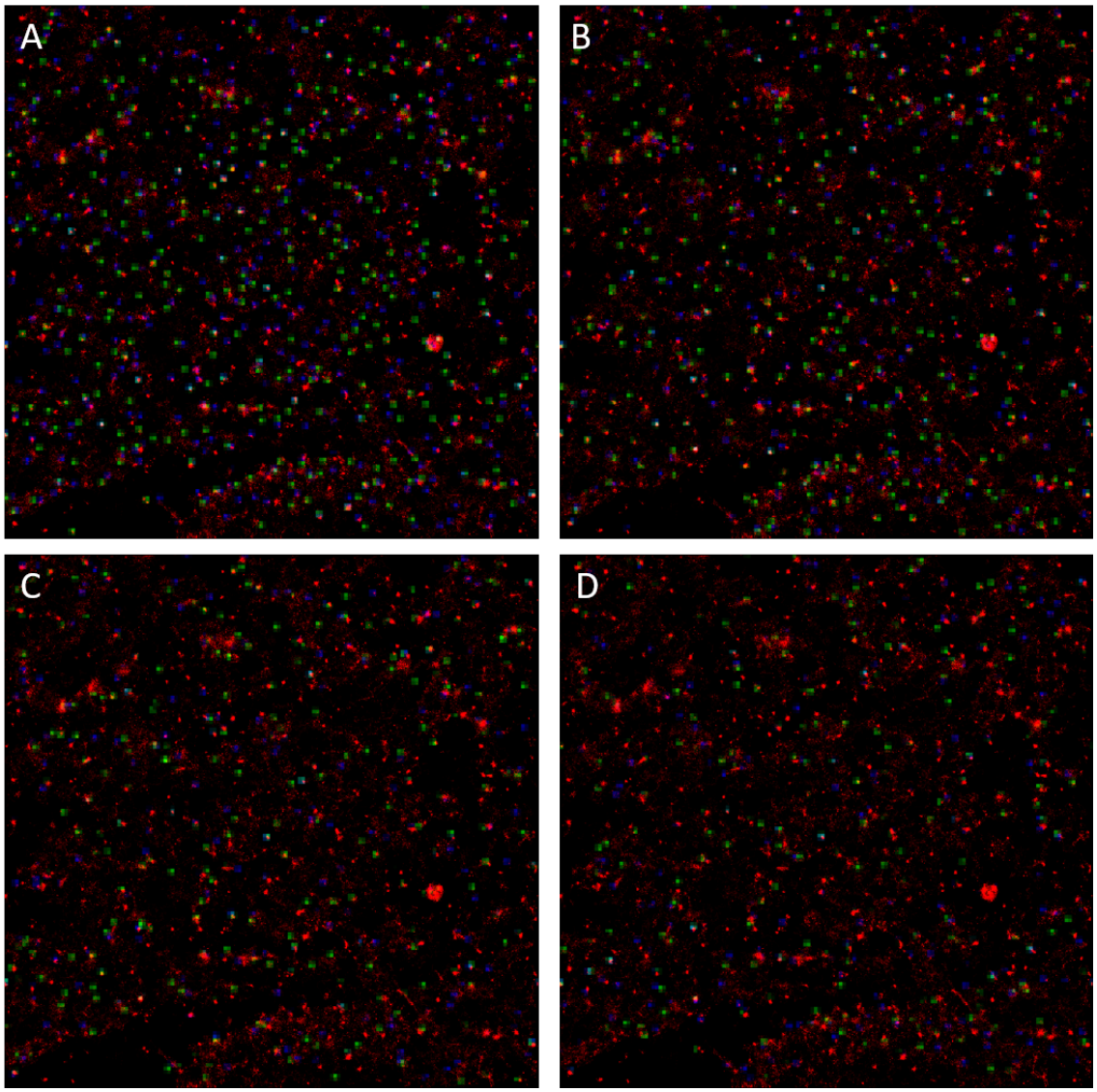



Figure 3.14: Population presence heatmaps (2.2.2) across time for set 643 C1, Red grid cells are the final movement heatmap of set 643 C1 after 400 time increments. Green and blue are two contiguous data snap time points selected from within set 643 C1, teal corresponds to the overlap of green and blue. A)0 (green) and 50 (blue) increments B)100 (green) and 150 (blue) C)200 (green) and 250 (blue) D) 300 (green) and 350 (blue).

**Population overlay** Within the larger hot-zones we can identify multiple small clusters of entities over several captured time frames (Figure 3.14,3.15). Therefore, suggesting that large hot-zones capture entities over time and can create co-localized groups as expected. We can also observe many scattered, inconsistent, and motile population placements differing across time, background diffusion. Small hot-zones may be consistent single population placement caused by restriction or other immobility; they show a low scale static population across time. However, some smaller hot-zones also have bright areas indicating larger populations; indication of co-localization similar to larger hot-zones. One of the distinct large hot-zones seen in C1 that

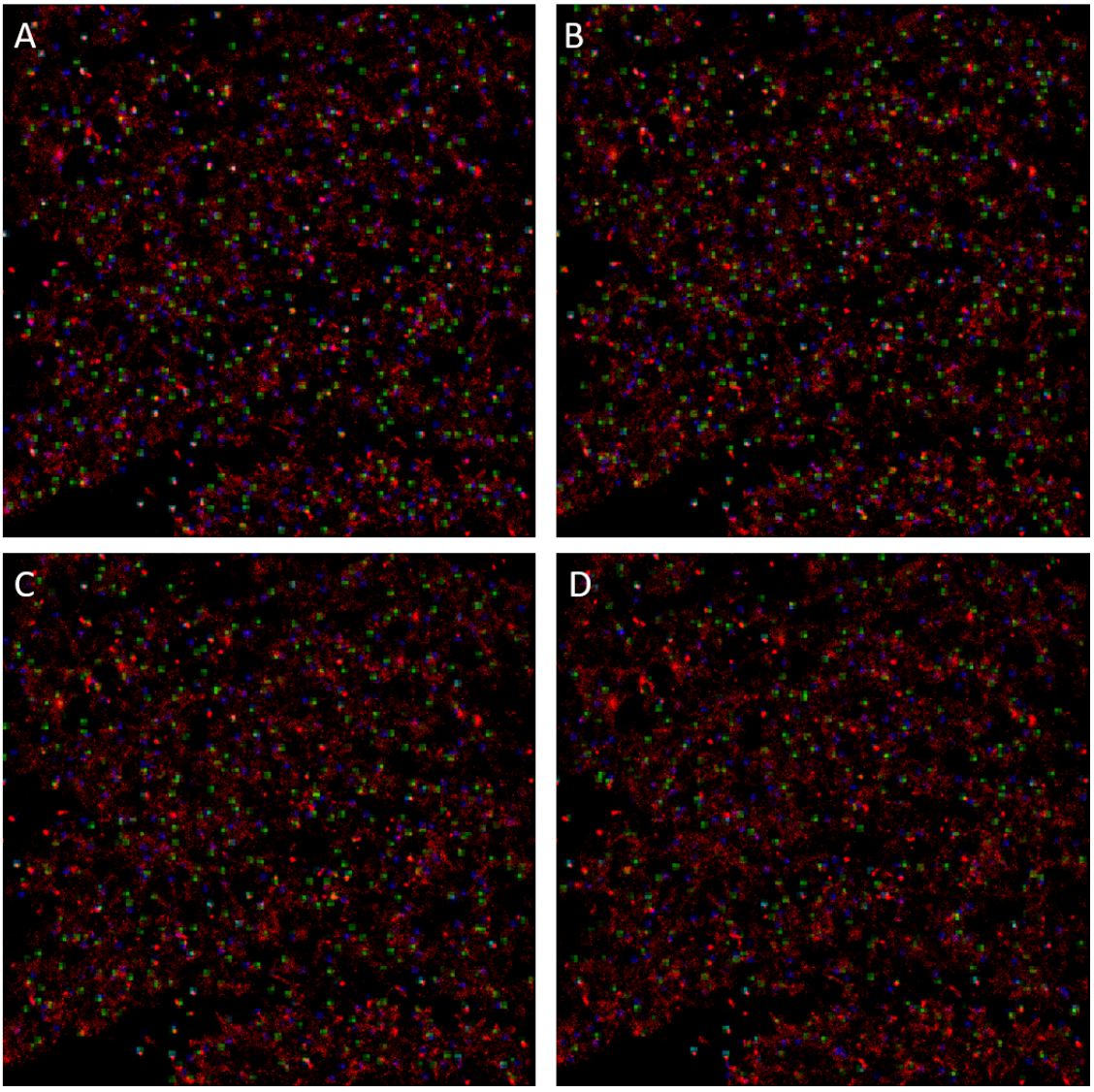



Figure 3.15: Population presence heatmaps (2.2.2) across time for set 643 C2; red grid cells are the final movement heatmap of set 643 C2 after 400 time increments. Green and blue are two contiguous data snap time points selected from within set 643 C2, teal corresponds to the overlap of green and blue. A)0 (green) and 50 (blue) increments B)100 (green) and 150 (blue) C)200 (green) and 250 (blue) D) 300 (green) and 350 (blue).

stands out as not overlapping with C2 is composed of a cluster of population positions. With this consistent disparate population placement within a slightly larger space we might suggest a cluster of several small adjacent micro domains or a large receptor complex is being observed. At later time points this disparate hot-zone shows marked decline in population members forming it but retains a clear pattern: possibly a large and very short-lived interaction (Figure 3.14).

### 3.3.2 GPCR and G protein models

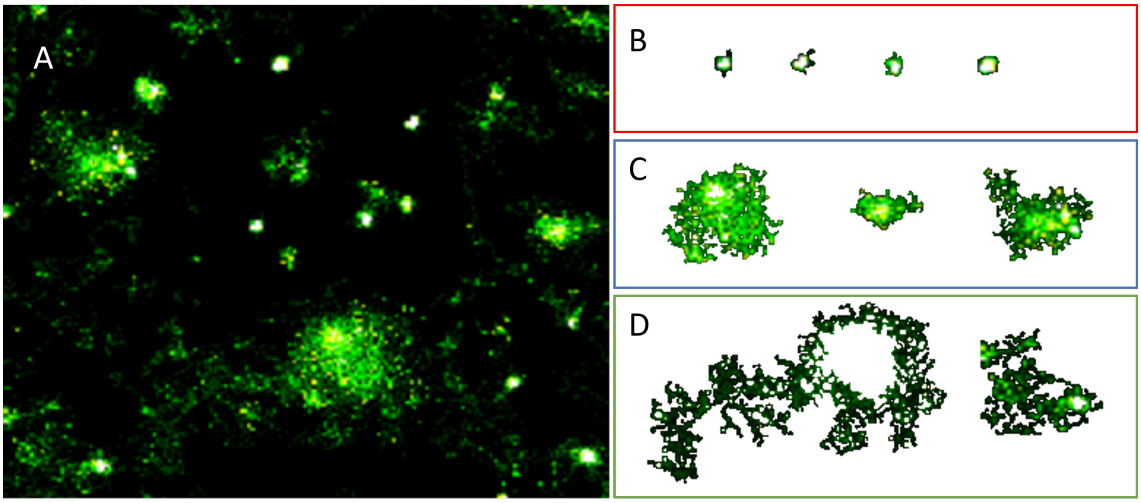



Figure 3.16: Key patterns identified via visualisation of GPCR and G protein set C1 641 movement heatmaps (2.2.2). Patterns shown are separated from a subsection of set 641 C1 (A) into small concentrated hot-zones (B), larger less concentrated hot-zones (C) and generalised areas of Brownian movement (D)

Observation of visual movement heatmap metadata has allowed the identification of distinct patterns within the given GPCR and G protein positional representation (Figure 3.16). Our modelling approach begins with the definition of general representative models and then the addition of hyperparameters to replicate identified patterns.

To compare hot-zones across different models and real-world, we can apply a uniform movement derived representative heatmap with the same heat intensity to movement gradient (Figure 3.17). By comparing, we can hypothesise about isolated observed patterns and the different causal conditions. Between immobile sub populations, attractive areas and deflective boundaries we can identify possible explanatory models for each of the major observed real-world *in vitro* patterns.

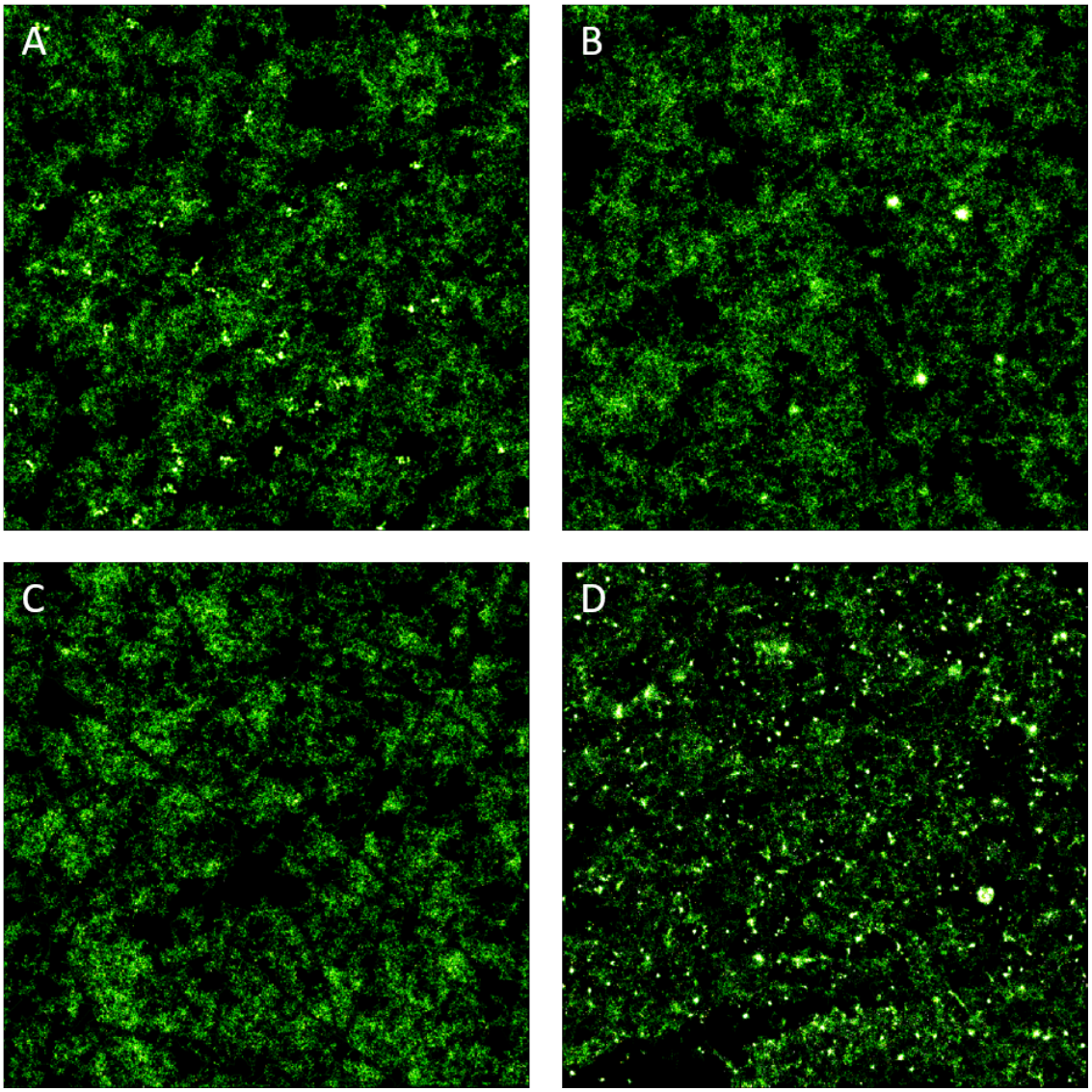



Figure 3.17: Uniform visualization with the same top-end for movement heatmap (2.2.2) gradient across several different model definitions and real data. A-C all include the C1 derived basic Brownian motion parameters, A) including 11% immobile population members, B) low strength attractive areas, C) deflective boundaries and D) real-world observed movement data 643 C1

### Model definition: basic movement patterns

GPCR and G protein movement is generally reported as Brownian diffusive motion; it is expected that particles and receptors will diffuse in random directions with Gaussian distributed distances. To produce a model of basic GPCR and G protein movement we therefore need to identify population size, average travel distance and turn frequency. Once a basic movement profile is implemented, we can add other population effects. In this case we will reproduce the C1 population reduction trend to visualize the effect on general movement density. To compare with later hot-zone explanatory implementations, we will also include an immobile sub population with a 11% incidence rate as reported in literature [54].

**Brownian movement C1 C2** Tracking can make accurately measuring population size difficult, however, the number of active participants at any time step was around 400 for C1 and 450 for C2 sets. Speed is slightly more difficult to determine since rather than a direct divisible distance by population total, we also need to account for the Gaussian distances. Brownian diffusion is simulated as contiguous undirected Gaussian distributed movements over time not a set distance, averages of 6.52 for C1 and 6.8 for C2 per increment showed similar distance results to real-world totals. We implement a general Brownian motion model within the framework representing C1 and C2 sets.

Our model generated Brownian C1 and C2 representations exist upon a constant flattened spherical surface area; upon hitting the edge of the displayed area they continue without vector change and appear at the opposite edge (Figure 3.18). We focus upon the general movement patterns, assumed extraneous parameters such as boundary conditions or collision have been disabled. The resultant heatmap shows no adherence within a clear boundary and therefore greater spreading than observed GPCR and G protein results. General movement patterns are similar to general motion areas observed but lacking hot-zones or non-random co-localization. Interestingly, movement starved areas are still common, this may be due to greater available area but more expansive data sets such as 644 still don't display similar starvation preponderance. Spread difference could be due to environmental parameters, indeed with the lack of hot-zones we would expect more Brownian motion to create clearer general coverage. Similarly, a model en-

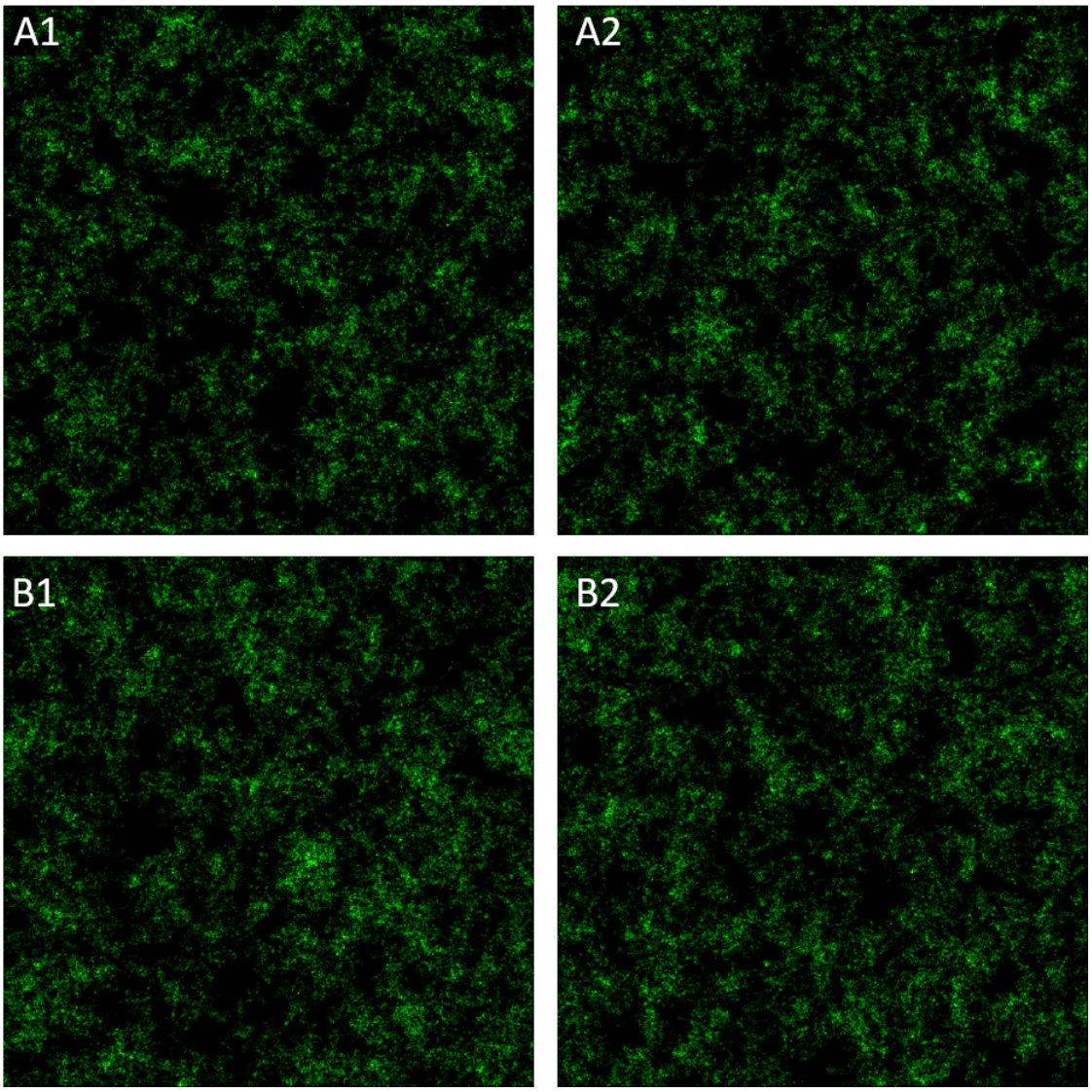



Figure 3.18: Representative model results for general C1 and C2 movement represented as population wide movement heatmaps (2.2.2). A) C1 run 1 and 2, B) C2 run 3 and 4.

vironment exists from the beginning to the end of record. However, a real observed environment would include prior interaction, we could suggest that greater starvation is due to the starting distribution.

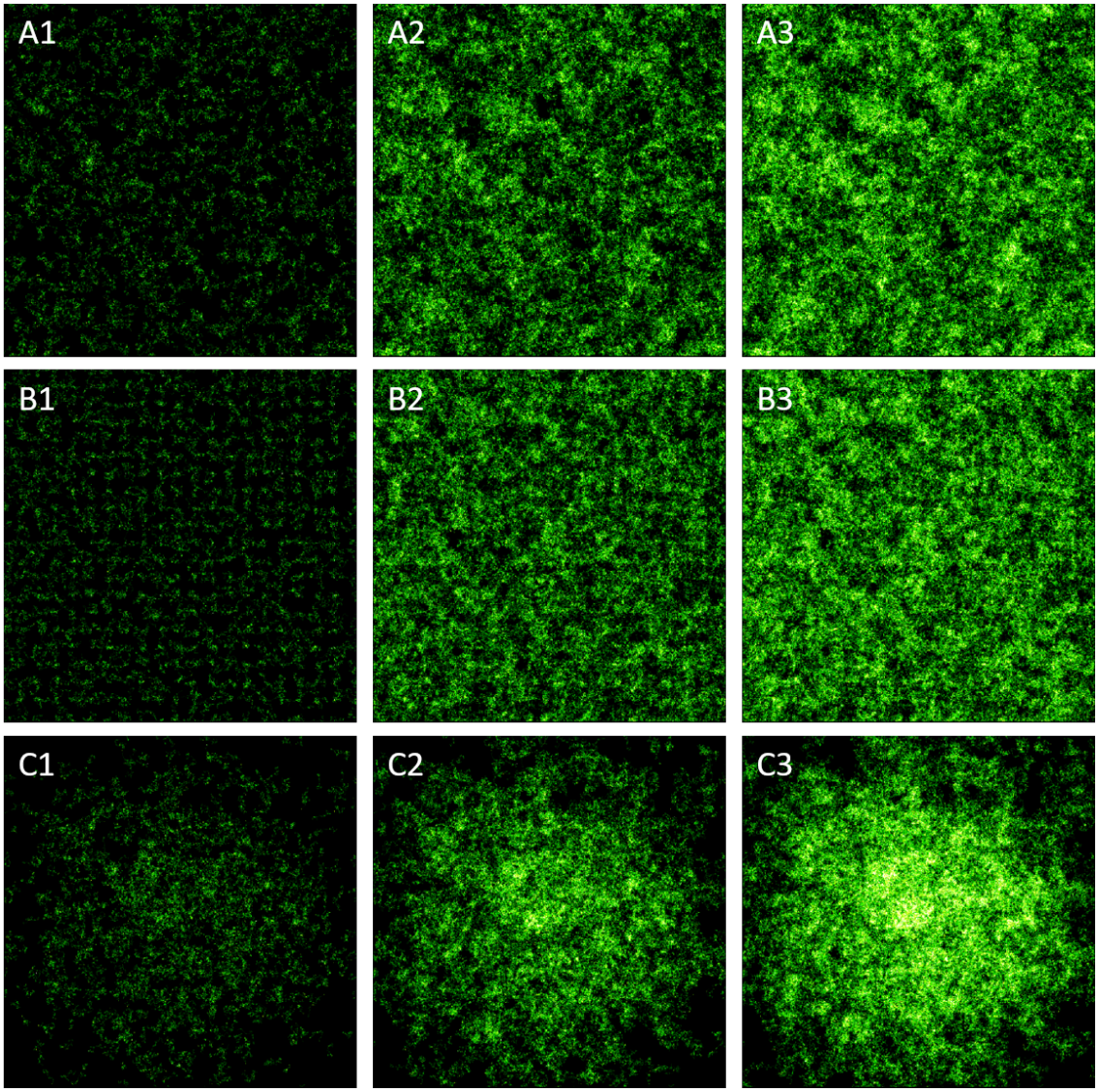



Figure 3.19: Heatmaps (2.2.2) for population movement across three simulations of random Brownian motion, overlaying all three creates an image for identifying repeated patterns, they were generated at 50, 200 and 400 time increments to observe initial entity placement. A) Random distribution, B) Even distribution and C) Normal distribution.

**distribution, populations size and coverage issues** Once the model is defined, the framework can generate different initial population distributions. Random initialized special distribution mimics the stochastic effects of unknown environmental starting parameters, even distribution

reduces positional effects on eventual patterns and normal distribution allows comparison with more clustered starting conditions (Figure 3.19). By taking the movement heatmaps of three runs with identical settings we can highlight repeating patterns. Random distribution aids in the creation of some movement starved areas but even distribution still creates more fragmented zones. Importantly, across representative models there seems to be no clear artificial bias, reproducible stochastic placement creates reasonably even coverage of movement over multiple runs as expected introducing minimal negative distribution bias.

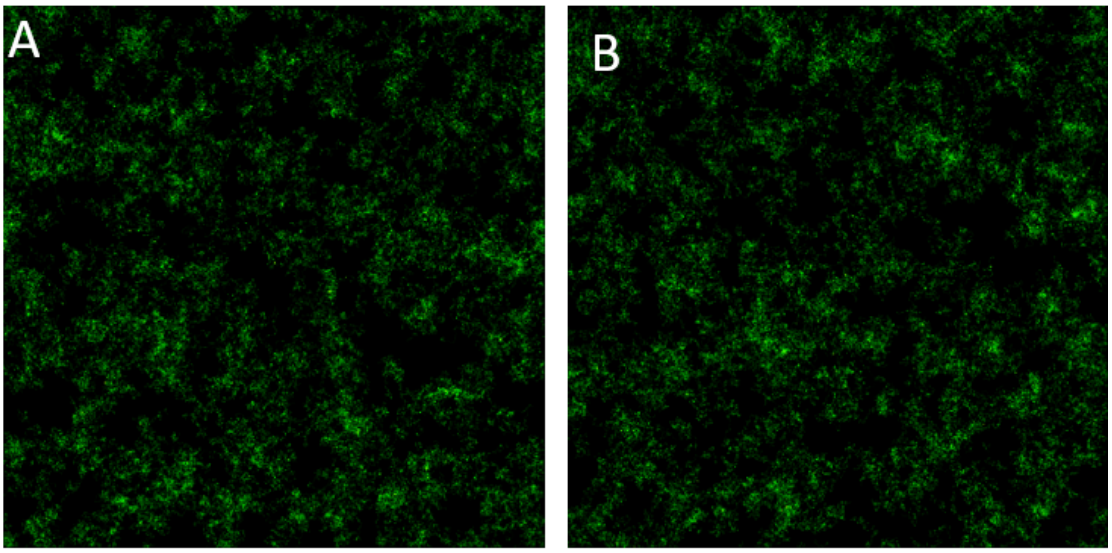



Figure 3.20: Population heatmaps (2.2.2) for C1 general movement with a similar population trend to observed real data for assessment of phenomena impact. A,B) Cumulative motion heatmaps after application of the entity removal rate.

**GPCR (C1) population trend** To assess the impact of secondary effects upon general population progression, we first apply our Brownian general movement hyperparameters and then parameters to add or remove population members. One of the observed trends from the quantitative real data analysis was that of a C1 gradual population decline (Figure 3.6). We defined a model with a probability of removal, gradual negative population change has little effect on analogous C1 movement patterns over time (Figure 3.20). Population reduction also leads to an associated distance travel reduction with variance being due to fluctuations in population removal and random Gaussian jump distance. It is reasonable to define restriction models without population change initially since it has been shown to have low immediate impact.

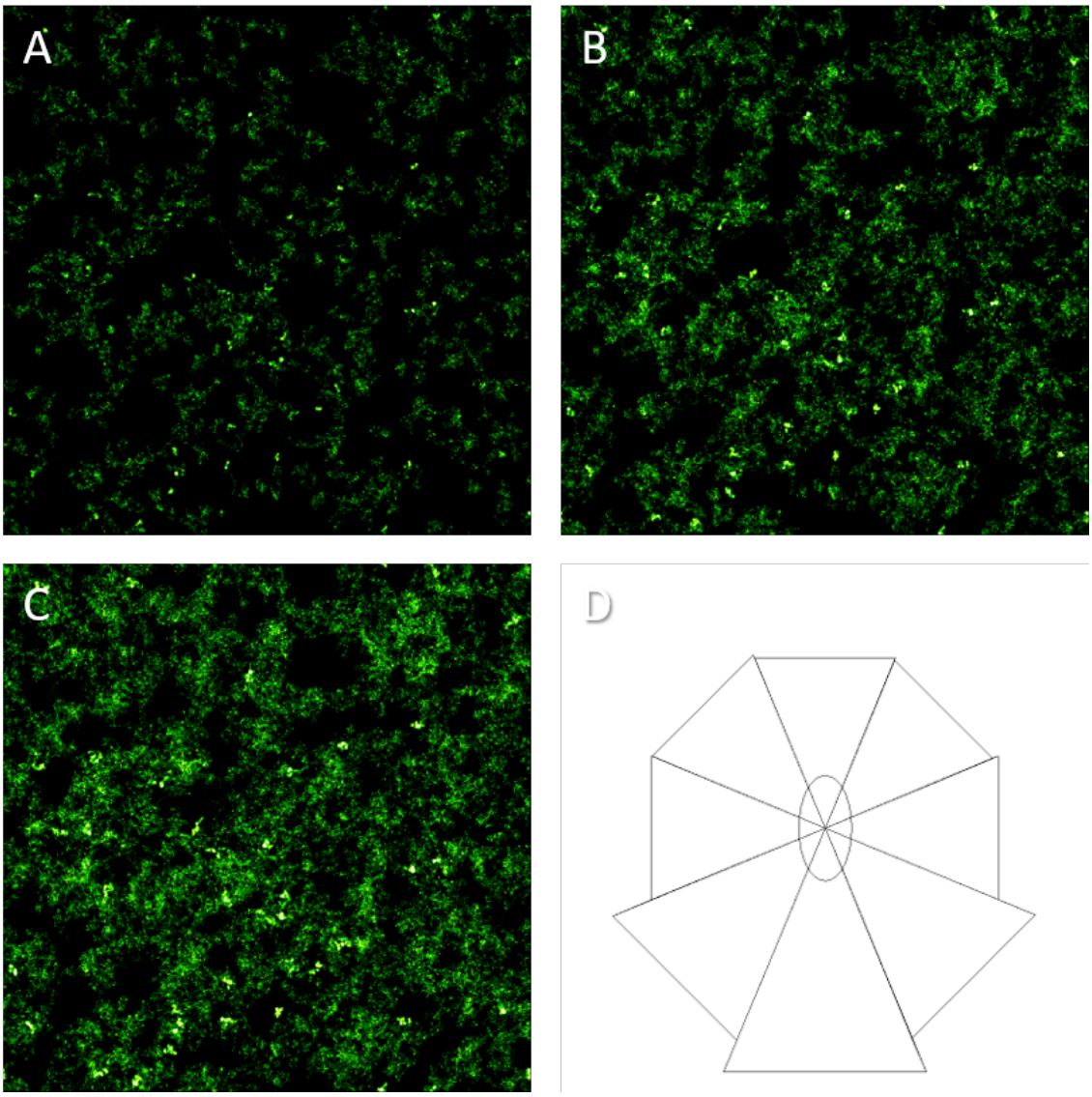



Figure 3.21: Movement heatmaps (2.2.2) for an active motile population of 360 entities and a 40 entity sub population of rear biased shivering immobile entities can be introduced at the same time to produce the above images after different time points. A)50, B)200 and C)400 time increments. D) Entire 400 increment period as a turn diagram taken from active population turns

**GPCR (C1) immobile sub group** The incidence of immobile population members has been noted in the literature as 11%, some smaller heat zones can possibly be attributed to them [54]. By creating a normal C1 movement group of 360 members and a smaller sub population of 40 shivering members we can attempt to replicate the phenomena in a model (Figure 3.21). Indeed, small hot-zones of extremely localized movement can be seen to occur across run generation, the sub population members adding visible differences at 50, 200 and 400 time increments. Additionally, the rear biased shivering effect even in such a small proportion has the effect of creating an overall rear turn bias similar to that shown in observed data. Despite being a smaller sub population, their behaviour still affects quantitative measures across the entire model. Repli-

cating this trend supports the idea that in the real-world GCPR system such a small immobile shivering sub population can cause some of the patterns we have noted.

**Model output: Attractive area vs restrictive**

With a reasonable definition of generalized Brownian motion focus of our work can shift to explanatory environmental hot-zone effects. We have shown that a low motility sub population can affect movement and turn bias via shivering (tracking or naturally occurring). Low motility does not explain smaller zones of co-localization or larger hot-zones and overlap between C1 and C2. As such we sought to implement representations of cytoskeletal deflection and catchment areas within the framework model. Hoping to reproduce both hot-zones and localized starvation the implementation allows for strength, size, and placement variation in both cases.

**Attractive zone model results** We can attempt to replicate co-localization hot-zones and local movement starvation by adding areas of attraction to the general background motion of our C1 model. Small but dense hot-zones can quickly be created if applied with high attraction strength. Locality starvation seems to increase with attraction strength, entities rapidly pass through the area of effect to the centre of attractive zones. With lower attraction strength, local starvation becomes less pronounced, but size, shape and effect more closely mirror the hot-zones observed in given GPCR and G protein data. It should be noted that all models are run for the same number of increments as observed data. In this timeframe attractive zones do not seem to significantly impact overall motion patterns beyond local starvation and hot-zone emergence (Figure 3.22). The attractive zone-based models suggest that the observed real-world *in vitro* phenomena are not necessarily created by purely restrictive interactions. Rapid attraction and compression lead to similar localised movement patterns, more successfully than purely restrictive models thus far.

We can again add attractive areas to the generated C1 motion model in order to address the problem of hot-zone size. While parameters for attractive zone pull strength, the value that dictates entity catchment and retention, can create larger areas at lower values the hot-zone morphology and local movement starvation associated are not as pronounced. Two possible implementations were applied: a central permissive area where entities could behave normally once

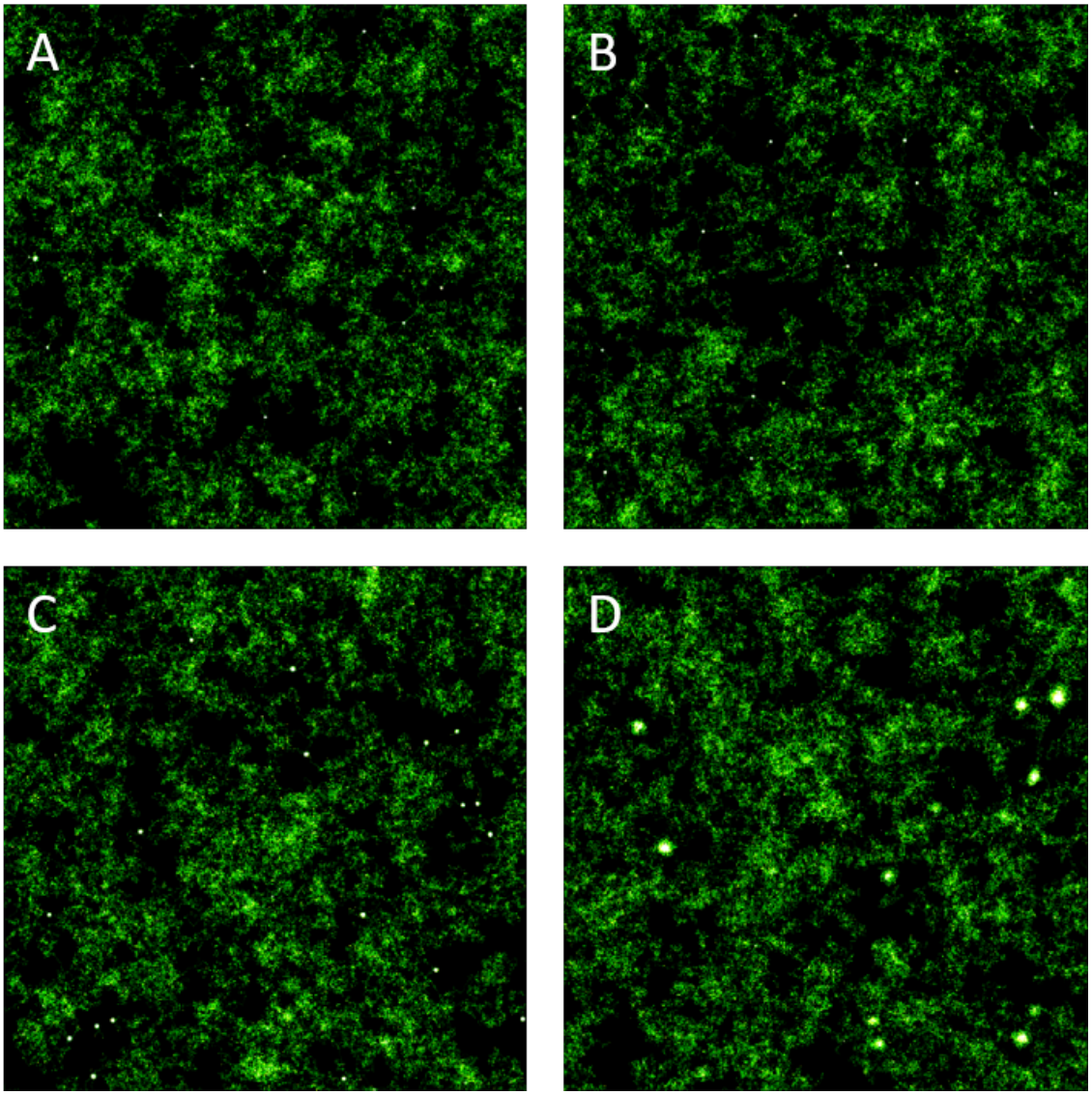



Figure 3.22: Movement heatmaps (2.2.2) for models using Brownian diffusion with parameters for C1 similarity, with attractive areas of various strength but similar size variance. Captured entities have their direction modified towards the central point by increments of 80 (A), 60 (B), 40(C) and 10 (D) % per jump.

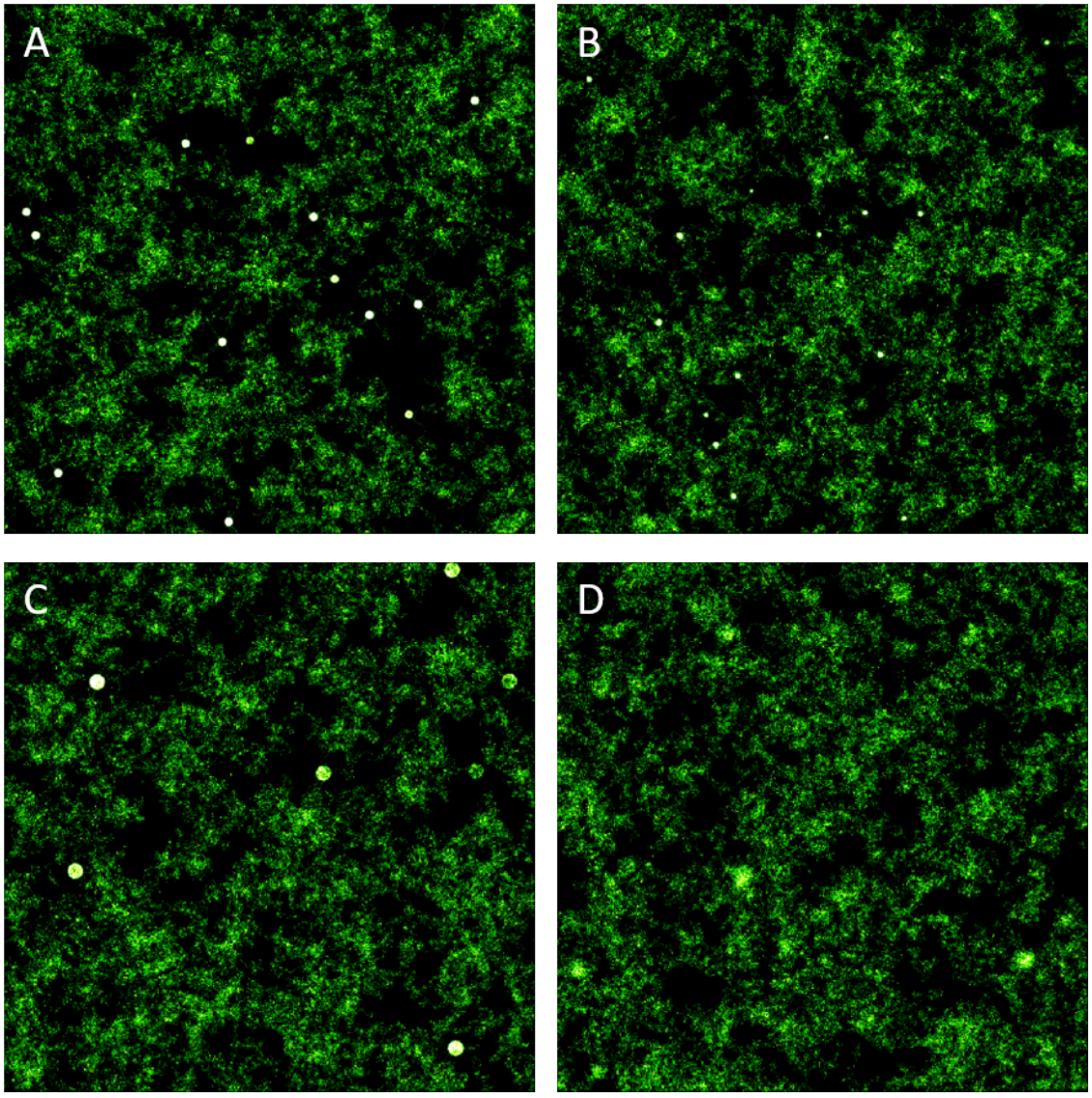



Figure 3.23: Cumulative motion heatmaps (2.2.2) for the C1 motion model but with the addition of attractive areas with permissive central environments and varying attraction strength. A) 60% strength and 10 pixels (px) of permissive space, B) 20% attraction and a 20% strength falloff to create a gradient falloff zone, C) 40% attraction and 20px permissive, D) 5% strength 10px permissive 1% falloff.

within an attractive outer area, and attraction strength falloff as entities were further from the centre (Figure 3.23). A larger permissible area within the centre of an attractive zone creates larger hot-zones but the shape is uniformly circular and does not compare well with observed data. A falloff approach creates more natural shapes at the expense of easy size definition. A combination of very low attraction, falloff and a permissive central area seems to create hot-zones like the locality starvation examples observed in GPCR and G protein data.

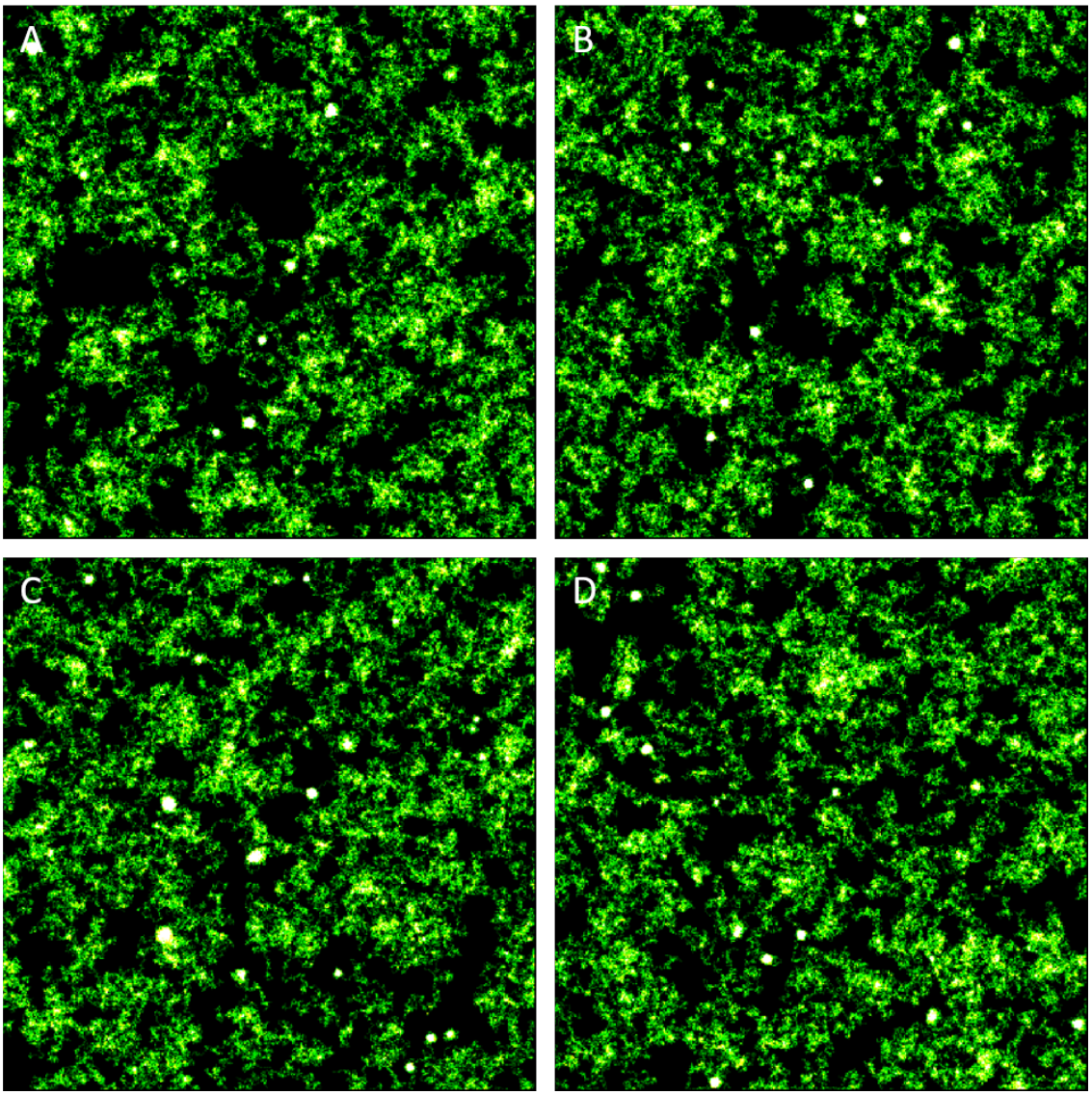



Figure 3.24: Given a range of input parameters including variance the modelling libraries can generate cumulative motion heatmaps (2.2.2) for the C1 model with the inclusion of widely variable attractive area parameters settings. A-C) We can generate *en masse* distinct separate runs to allow for stochastic effects.

**Attractive higher ranges** We observe the effect of higher numbers of attractive areas with higher size and strength variance. Increased localized starvation and global movement starvation can be observed with higher numbers of attractive zones; as entities are trapped, general movement diminishes (Figure 3.24). With a similar increase in numbers but lower strength smaller individual zones are not as clear, patterns blend into high movement areas. In observed real data, small zone distinctness suggests high immobility or area constriction greater than that we applied.

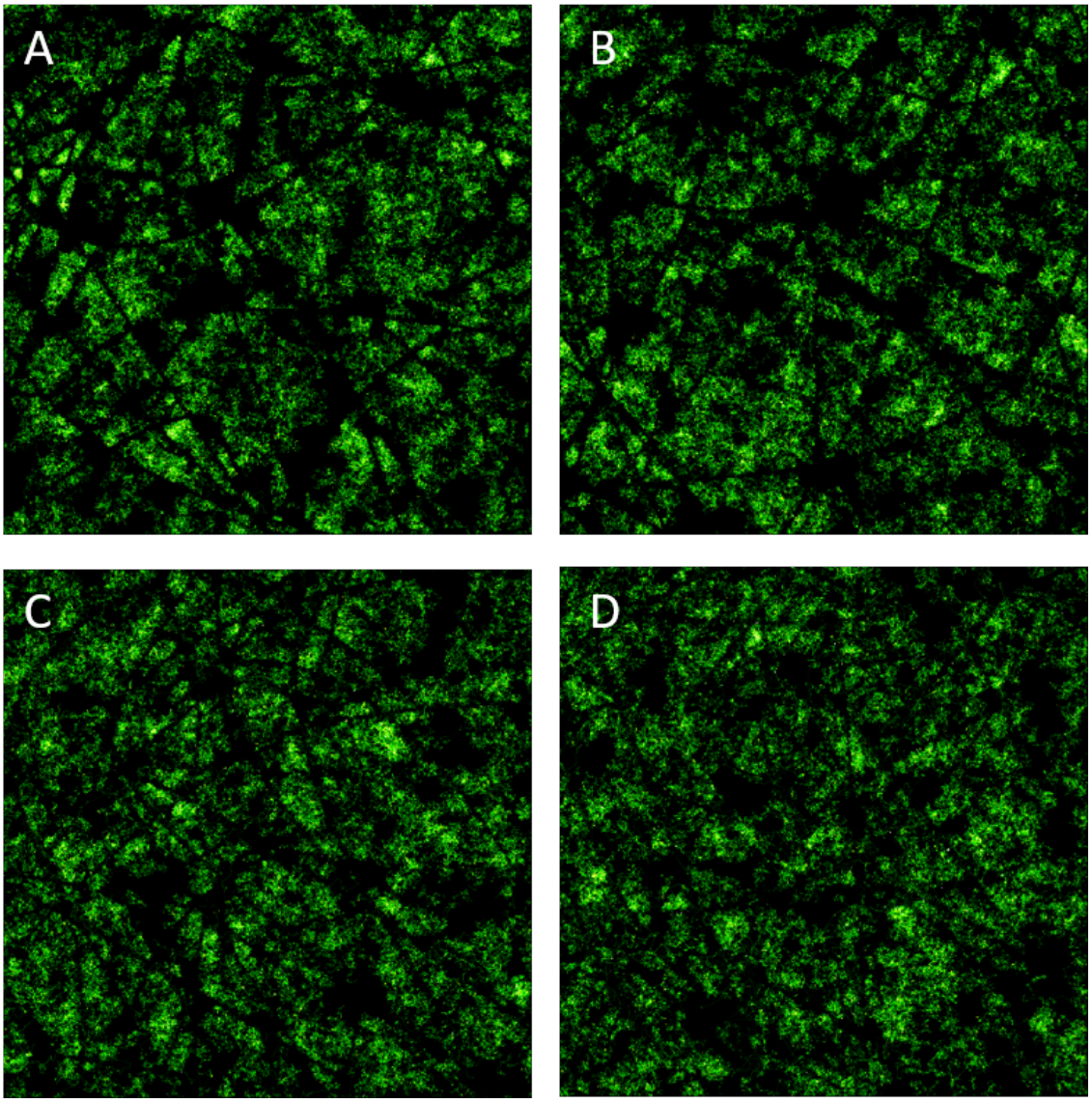



Figure 3.25: Cumulative movement heatmaps (2.2.2) created over the entirety of a model run using C1 based general motion profiles with the addition of deflective curves to investigate cytoskeletal restrictive areas. Varying deflection likelihoods have been applied A)90, B)80,C)60 and D)40 to constrict and compartmentalize entities.

**Deflective curves** The introduction of multiple cubic curves can represent an overlapping cytoskeletal mesh in the environment, variable permissiveness then possibly creating hot-zones from which no entities can escape. To replicate observed real-world trends, a range of deflective strengths were applied to the defined C1 motion model (Figure 3.25). It is possible to create "hotter" areas via deflective curves with general confinement and areas of motion starvation. However, it is difficult to identify the same geometric patterns identified across observed C1 sets. The created hot-zones are brightest and most pronounced at the highest deflection rates and not spherical. Additionally, introduction of these deflective areas adds a clear pattern of reduced movement where they run. Extreme local permissiveness differences could account for a lack of these clear motion breaks in the real observed data, consistent with rapid micro-environmental change but increased difficulty. It is important to note that while hot-zones from deflective catchment do become clearer at longer timescales, these are not close to the real set time frame. In the given timeframe such strong barriers create the clear movement starved path patterns that don't clearly appear in real data images.

**Hybrid model** One possible explanation for the lack of clear barrier movement effects in the observed real-world GPCR and G protein data is that of combinatorial effect with other environmental parameters. We therefore applied both attractive and deflective hyperparameters in a single hybrid model.

Adjacency of attractive zones to restrictive barriers within the model does increase intensity and modify hot-zone shape (Figure 3.26). Barriers intersecting attractive zones also create greater outline perturbation in a similar manner to observed patterns. Movement starved areas are still clear around hot-zones and overall distribution is not clearly changed except in the visual presence of deflective barriers.

#### Turn comparison

We can improve POM by first modelling with hyperparameters for explanatory visual pattern effects and then comparing with other metadata such as turn preferences. In the case of the limited models discussed so far both C1 and C2 general motion preferences are entirely random

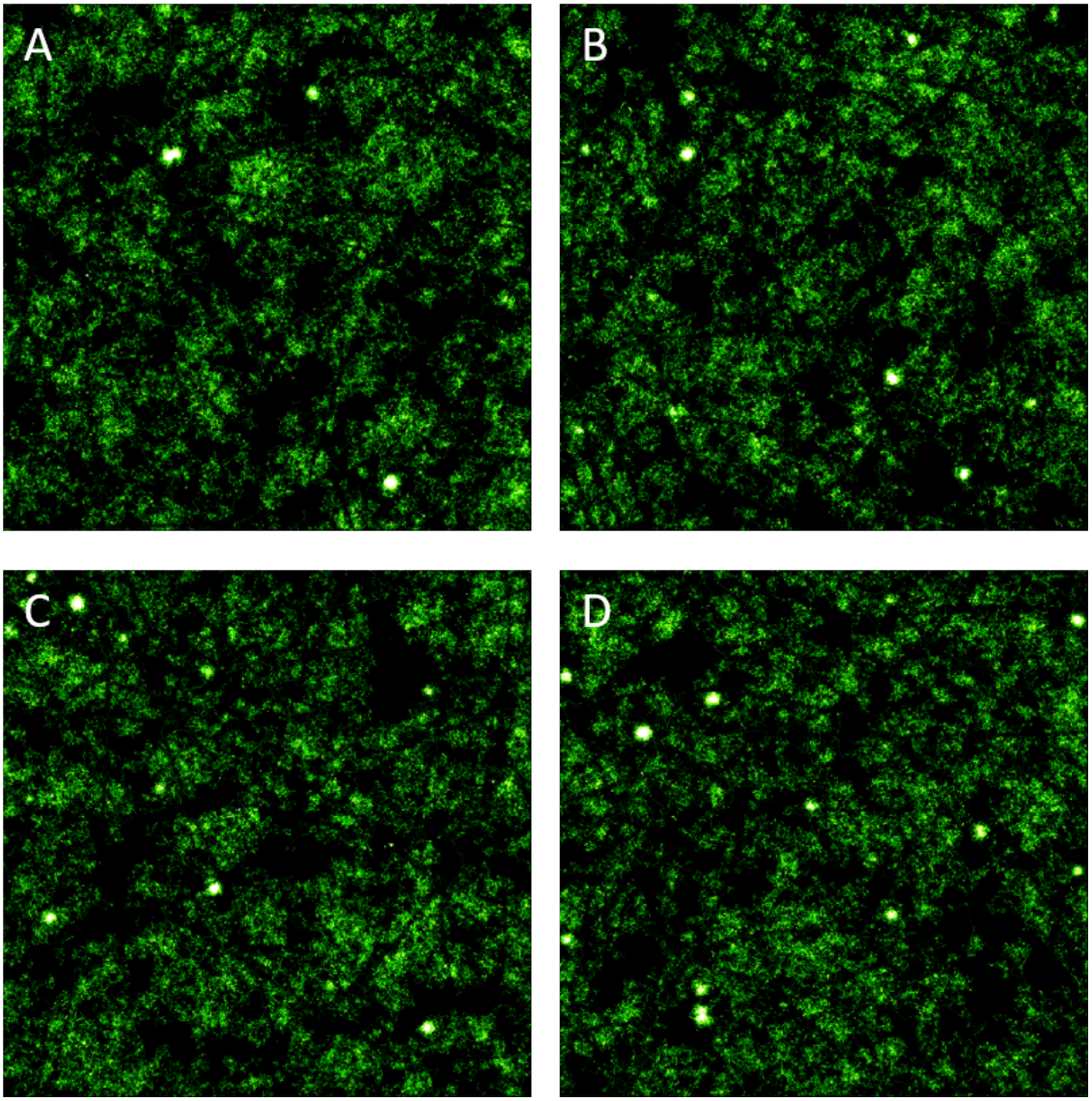



Figure 3.26: By mixing deflective curves and attractive zones with 40% deflective strength and 10% attractive strength respectively we can create repeatable and scalable combined effect cumulative motion heatmaps (2.2.2) over many separate runs.

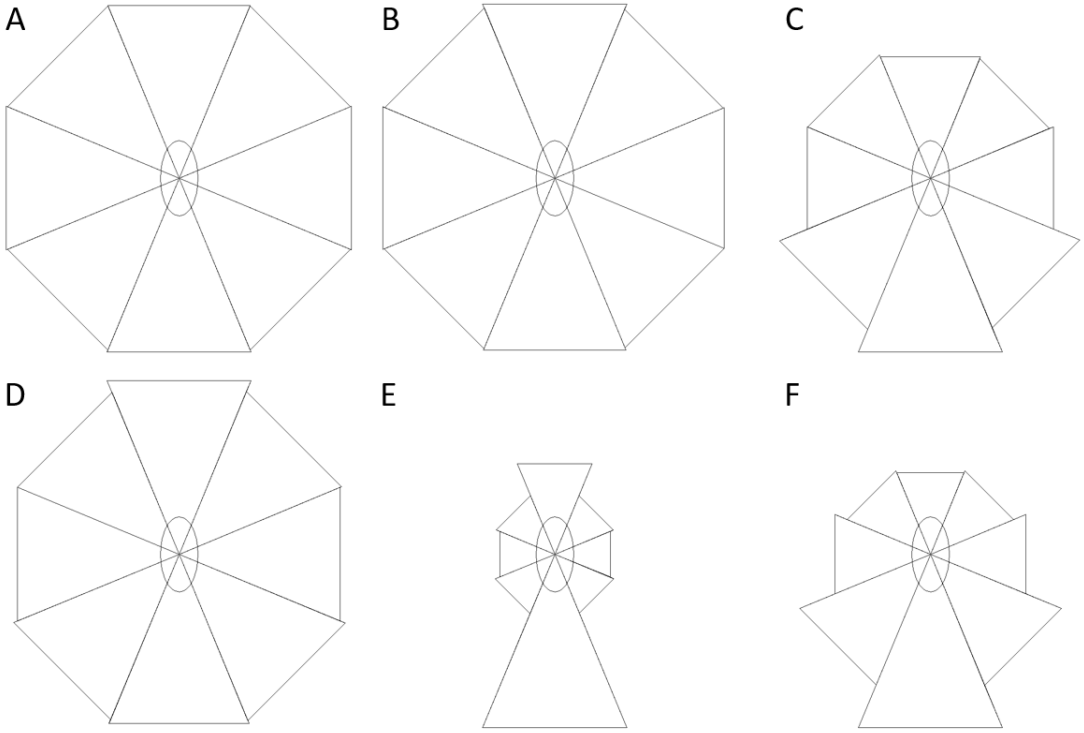



Figure 3.27: By measuring turn preference across a population over the entirety of a track set or model data run we can create relative turn diagrams for visual comparison. The larger a section the higher preference for turns in that direction relative to an entities previous direction. A) Base C1, B) base C2, C) C1 with immobile population, D) deflective barriers, E) attractive zones and F) real observed GPCR and G protein across all available data sets.

as expected with a Brownian motion implementation (Figure 3.27). We previously showed that the inclusion of an immobile shivering sub population creates a similar but less pronounced turn profile to real-world GCPR. We can observe from turn diagrams that deflective curves lead to a slight forward bias, possibly a product of deflection angle suggesting that a more reflective barrier might be appropriate (Figure 3.27). Attractive zones create both forward and rear bias, if we assume forward to be the initial trajectory modification of attraction the rear bias might be reasonable evidence for the enhancing effect of activity within a larger hot-zone.

## 3.4 Discussion

### 3.4.1 Tracks

Using our Framework we have been able to identify several recurrent patterns within the provided GPCR and G protein positional data some of which were discussed within the literature [54] (Figure 3.16). A combination of quantitative statistical measures (Figures 3.6, 3.7 ) and visual metadata such as movement and population heatmaps (Figure 3.10) with qualitative observation was recorded. We highlighted hot-zones and begin to understand the dynamics that lead to their emergence. Travel distance, turn rate and population size correlate with observed surface area but does not dictate the overall turn distribution of a population within the observed system.

As time progresses movement seems to decline (Figures 3.6, 3.7), lower ongoing activity might indicate experimental agonist introduction and drop off or an indicator of a cyclical behavioural phenomena in G proteins. A large disparity of total movement might be attributed to tracking inconsistency or other outside interference. However, the pattern is consistent across available data sets with a slight drop, possibly an indicator of population-wide effects such as a receptor specific stimulant.

The literature suggests that hot-zones exist at least partially to force GPCR and G protein together for enhanced binding and communication [6, 55, 56, 57]. Following the co-localization narrative, areas of high movement density should be common across C1 and C2. We have identified that smaller hot-zones may also be immobile population members (Figures 3.14, 3.15).

Therefore, overlapping stationary hot-zones may indicate a method for differentiation between low motility single members and small restrictive areas confining multiple population members. Further, this comparison is complex, neither C1 solely interacts with C2 nor C2 solely with C1, the hot-zones may be self-sorting for micro-environmental specific co-localisation or have different active time effects for different pairings (Figure 3.13). We should also be aware of tracking mismatches or video misalignment, one detection channel being a very small distance off centre can create a mismatch, patterns of mirrored mismatches should still be reasonably localized in this case.

Within the narrative of confinement, a rear bias (Figure 3.8) to turns could be caused by a boundary based hard bouncing effect or bungie like snap back. Either would be environmentally driven, possibly relating to confinement or fence and picket analogies outlined in literature [6, 55, 56, 57]. It is also important to remember the effect of tracking bias; track loss and fragmentation is present in any application. We once thought stationary entities may be interpreted by tracking as shivering; we tested it and found shivering tracks are the result of tracking shivering entities. It is unlikely tracking bias caused the rear turn bias.

There are a few general common observable patterns across all sets, caused by the method of measurement and the targeted medium (Figure 3.12). Activity occurs on and around the cell membrane but that it is not a true two-dimensional environment. The clear lines of activity reduction can be cell edges. Movement can occur outside the clearly defined cell locality. However, it is not importance to the communication processes we seek to define, in fact it should generally be ignored. As with all observed real-world data, interaction and extensive environmental effects have been present prior to measurement. An observed environment is a product of ongoing and prior behavioural emergence from complex unknown prior interaction. Therefore, observation becomes a task of identifying salient trends occurring within explainable time frames.

High movement hot-zones have been confirmed as present with varying size and entity retention levels across time, with both small and large zone possibly indicating confinement (Figure 3.10). Population positions over time suggests large hot-zones are more likely indicative of multi entity co-localisation. Heatmap representations also highlights the presence of greater general movement of C2 over C1 populations. C2 populations showing greater distance and turn num-

bers, closely tied to average larger active populations.

Possible avenues for further investigation were also identified. In all data sets the C1 active population showed continual reduction as time increments passed. Both C1 and C2 showed a marked trend down in the last increment likely due to tracking but C2's slow downward movement trend suggests we may be observing part of a possibly cyclical population behaviour or decline. In addition, we were able to identify a common recurring pattern of movement starvation around hot-zones. Whether starvation is due to a strong attractive process or barrier condition is currently unclear. There is some evidence that it may be a slowing barrier with variable paucity between C1 and C2 populations. In set 643 C1 we were also able to identify a large hot-zone that didn't appear to conform to the overlap trend already established. Identifying whether it is a co-localization of several hot-zones or larger construct could lead to definition of a new hot-zone type. However, we should remember that although recording conditions are the same, some data variation is inevitable in biological systems; possibly interesting phenomena can be outlying data as well as indicators for further hypothesis exploration

### 3.4.2 Models

We can create small shivering areas, analogous to non-co-localization hot-zones previously observed in the GPCR and G protein data by adding a smaller 11% present immobile sub population(Figure 3.17). The shivering sub population could indicate a real-world phenomena, entities rapidly moving back and forth caught by an attractive zone or restrictive pocket. Large low attraction areas suggest an explanation for localized movement starvation and the patterns of observed larger co-localizing hot-zones (Figure 3.22). While the turn bias observed suggests that the initial attraction method is not led as in the model, rear bias suggests that bouncing within a more permissive central attractive area may be present. While deflective boundaries do not seem to be a strong single explanation for hot-zones, movement segmentation does bear some similarity (Figure 3.25 ). It is possible that deflective boundaries help to compartmentalize general movement, once localized weak attractors bring receptors and G proteins together in a more concentrated mass.

### 3.4.3 Conclusion

Initial visualization of GPCR and G protein sets successfully highlighted several key movement patterns: hot-zones, general Brownian movement and starvation were all present along with clear areas of GPCR and G protein co-localization when C1 and C2 were overlaid. We were able to examine the spatial relationships between GPCRs, G proteins and their environment. Using our approach, we can now suggest explanations for several major real-world *in vitro* phenomena by combining distinct models of movement patterns; we were able to replicate patterns within representative models. General background movement patterns were made of population wide Brownian motion and areas of movement starvation often near catchment zones. Also, attractive zones and shivering sub populations both produced hot-zone patterns similar to those observed variable morphology. Shivering sub populations also reproduced a rear turn preference across the population.

However, there are some interactions and environmental factors described in the literature that we could again include in future models. For example, lipid-protein complexes and nanodomains could be included affecting population member movement, it is thought that the arrangement of such constructs can strongly influence confinement [9]. Similarly, confinement conditions can be transient and impact protein and receptor movement and behaviour [9]. Implementation of hyperparameters to represent either might be important. While we currently have a cytoskeletal confinement model, the structure does not change across time within a run. Therefore, representations for cytoskeletal structural change change over time may improve representation for more nuanced models.

Existing simulations might also be further examined in a more detailed metadata representation. If we can improve the reliability of our representations and the quality of information derived from real-world systems, we generate more in depth patterns for comparison and validation of representative models. Analysis improvement allows us to increase model complexity and in turn provides opportunities to investigate real-world interactions, helping us understand the biological system. To progress and improve analysis, it is reasonable to next use the framework to examine outside bias, improve heatmap directional representation and automated quantitative comparison tools.

# 4 An expanded micro-environmental view: methods for further pattern identification

## 4.1 Introduction

Previously, we used our framework to generate mechanistic hypotheses for observed population-wide movement patterns of both cancer cell (Figure 2.11) and G protein-coupled receptors (GPCRs) (Figure 3.16). Across these applications gaps were identified regarding the detailed construction of meta patterns such as strands or hot-zones over time, and how location dictated entity directional selection. Therefore, the next steps aim to improve our framework by enabling the characterisation of directional trends within micro-environmental patterns, their construction and morphology over time. We observe micro-environmental effects by further exposing localised population behaviour.

Movement heatmaps were modified to display directional preference, allowing observation of population movement dynamics on a smaller scale, more environmentally specific. Furthermore, we want to test if populations could be split into multiple sub-populations based upon representative patterns, e.g. subsets interacting or not with specific environmental effects. Therefore, track filtering was also introduced with real-world digitised tracks split according to manually selected metrics into subsets for comparison.

In addition to observing recorded environmental effects, we suggested that pre-existing environmental conditions such as lattice paths or varying environmental density likely affected the populations we observed. Such pre-existing conditions would be difficult to differentiate and fully explain without a matching set representing an environmental baseline. However, we can focus upon change within observable intervals. Isolate changes within images generated for each consecutive observed window rather than displaying cumulative movement. This will ensure

that trends are continuous and not a reflection of strong early effects being carried over with cumulative measures.

### 4.1.1 Summary

We extended our modelling and analysis by developing new tools: directional movement heatmaps, time phasing and population sifting upon gathered metric data. Applying them to the cancer set reveals that strands may well coalesce from forging behaviour interacting with pre-existing least resistance paths. Also, cancer cells seem to retain some adherence to these paths even when entering less restrictive strand like areas of movement. In the GPCR and G protein system, we identify that rear turn bias across the population is likely indicative of hot-zones. Visual inspection of directional selection shows rear dominated movement is pervasive across identified hot-zones. We also observe hot-zone coalescence and disassembly over time.

## 4.2 Methodology

### 4.2.1 Pattern identification

Often, when presented with a real-world *in vitro* set, we have limited examples. One solution is to use representative models to expand sets for machine learning. Therefore, in the case of both GPCR and cancer cells with relatively small datasets, it became essential to maximise the depth of possible analysis. Consequently, we increased available visible patterns for observation, primarily focusing on entity directional choice within micro-environments.

Micro domains and localized environmental effects are expected, movement-based heatmaps allow us to visualise their effects on general activity, where addition of a directional element potentially improves this approach. Similarly, a common issue with analysis of gathered real-world *in vitro* sets is the difficulty of estimating prior interaction and environmental variables. Accounting for interaction prior to observation and effects upon a population can be very difficult. Splitting observations by phase can help us simulate unseen prior parts of a biological system or isolate and clarify the stages we can observe. Finally, sub-populations can have pronounced effects on overall observed behaviour. Filtering real-world tracks into sub-populations for input can test the relationship between subsets and patterns. To summarize, we developed tools to

reveal patterns with greater topographical, chronological and population specificity.

To minimize the introduction of possible confounding parameters, previous hyperparameters for the representative models were used for both cancer 2 and GPCR 3 comparison. There is also a significant limitation in the current application of these new heatmap representations as it is very difficult to generate representative numerate analysis. Hot-zone and lattice path construction in particular is not just a question of movement density but also morphology, with complex problems requiring very detailed multi variate representation for definition.

#### Directional heatmaps: localised directional preference

Implementing directional heatmaps allows us to observe turn preference relative to a position within the observed environment over time, where population members most often change their general directional preference.

Directional heatmaps seek to overlay directionality upon existing movement density values (Figure 4.1). When normal movement heatmaps increment a movement density value at the position of population members, directional also registers their heading. Direction is totaled at stated intervals and displayed as arrows for each section of a grid, colour is assigned relative to a cap derived from the strength of a dominant movement direction. Directional heatmaps therefore show the preference of a population to make a turn in the direction of a given arrow at every position when observed.

There is a risk that non-dominant movement angle information would be lost with a dominant directional angle display. Supplementary heatmaps are therefore generated for each of the eight angles being observed. We can then compare dominant angle displays with angle isolated heatmaps to gather additional insight if relevant. Normal movement heatmaps are also still generated to provide general movement density information. Ordinarily movement is gathered on a 2x2 pixels scale, to accommodate arrows more pixels are needed so grid cells are scaled up to a 4x4px minimum. Data could be gathered at a smaller scale then positional multiplication and scaling used to display properly, as is the case with the built-in visualizer tool. However, file and

image sizes become difficult to manage for a small increase in detail.

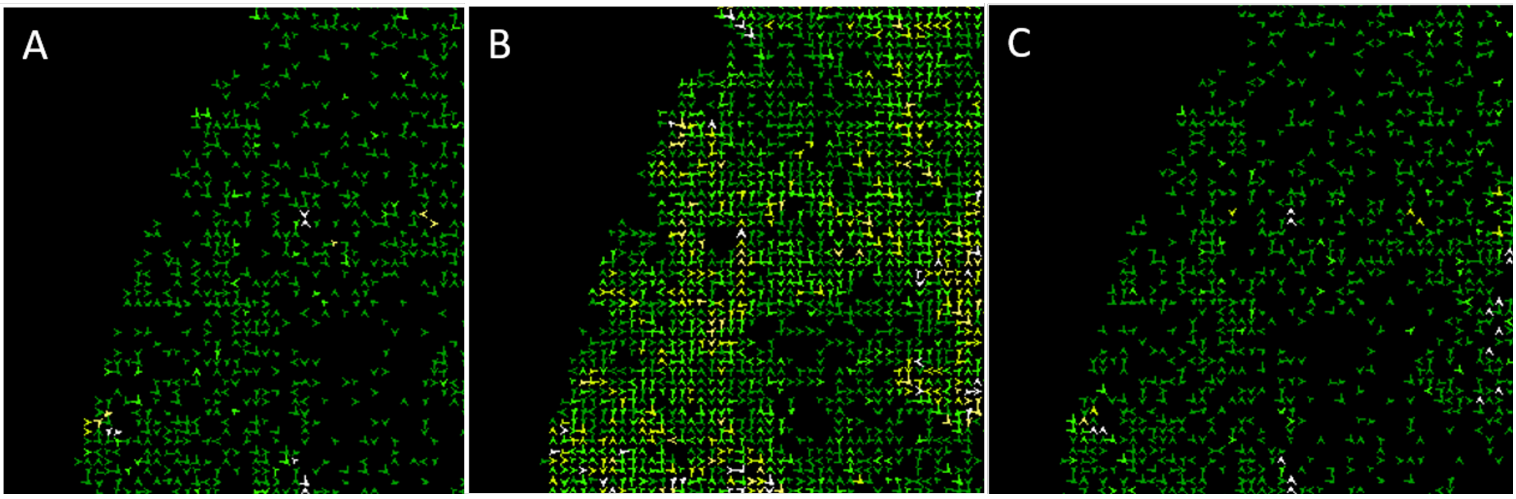



Figure 4.1: Directional heatmaps (4.2.1) where triangles indicate the actual direction of travel for entities (A,B) and relative to previous turns (C) at jump (B) and increment (A,C) time steps.

There are currently three main implementations of directional movement heatmaps. They use the same representation, but differ in arrow meaning and capture timescale (Figure 4.1):

1. **Absolute directional maps** show arrows indicating the direction of preferred movement relative to environmental orientation, that is, if an arrow points somewhere then movement is preferred in that direction at that position (Figure 4.1 A,B). Showing absolute direction allows us to see the reaction pattern of entities relative to any environmental effects deflecting or reversing them.

2. **Constant absolute turn representations** display absolute direction captured at every smaller jump step. Each time increment and known position is connected by a number of intermediate jump steps within a data run. By registering the presumed steps we loose some accuracy but include information that may represent missed or obscured movement.

3. **Relative directional heatmaps** also only register at known turn steps but relate back to relative turn circle diagrams, an arrow pointing up means directional preference of sector 0 (337.5 to 22.5 degrees) in relation to their previous direction, down suggests a reverse not southern preference. They are intended to show the continuity of directional choice relative to previous heading, detecting patterns such as conservation of direction or rapid/random directional change.

**Phasing: isolating short and long term effects on a system**

In our case, a time phase refers to a sequence of population movements between two set time points. Previously, data was gathered in a cumulative manner and presented for interpretation at the end of a run. With phasing we record and then re-set counts and measures so that our data represents behaviour between two chosen time points; the time phase. We can isolate and observe each phase without patterns from previous movement.

Most of the numerical metrics gathered on an analysis run represent the actions of a populations' entities between given time phases. Cumulative totals are auto generated by simple addition of ongoing phase values. However, movement heatmaps are both spatially large and information dense enough that merging them has a noticeable computational and therefore time cost. Previously it was more effective to simply increment movement values and display cumulative maps at each time frame. Addition of heatmap cleaning between time frames requires a compute increase but allows us to observe the effects of environmental interaction in a more nuanced manner.

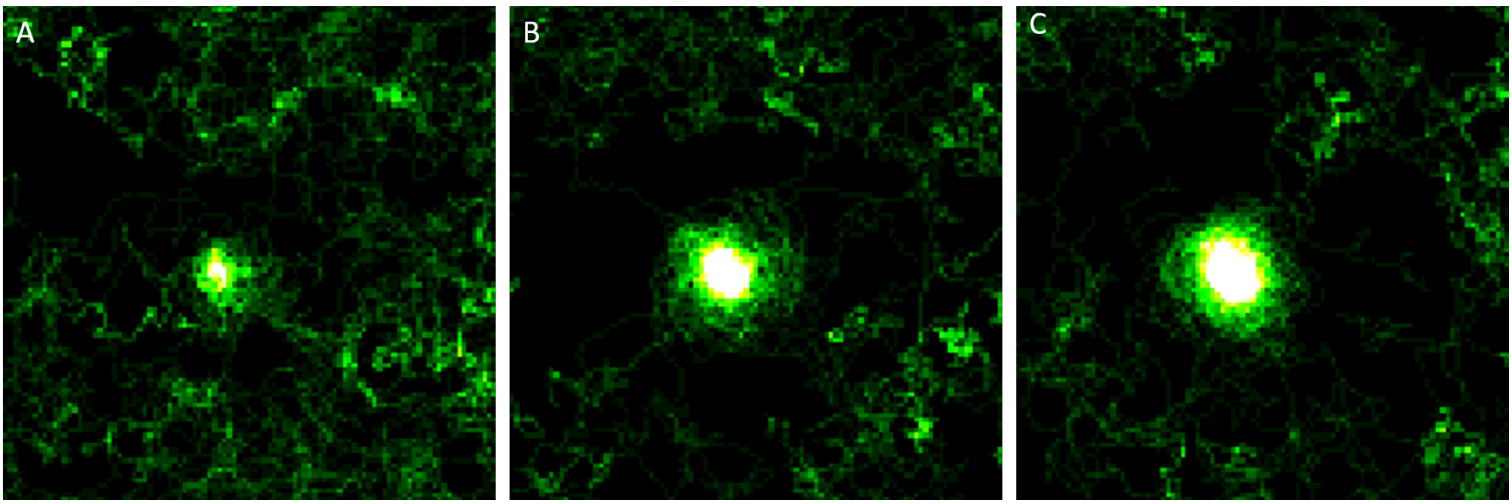



Figure 4.2: Heatmaps (2.2.2) displaying population movement over time in an environment with weak attractive areas through three distinct non cumulative time frames

By separating the movement values for phases we can assess the growth of environmental effects such as *attractive zones* without noise from previous interactions (Figure 4.2). In the case of attractive zones, each subsequent time phase displays increased movement within and gradual movement starvation outside as more members of a population are captured. We can create a circumstance where a single model addresses formative steps and the resultant long-term pattern presented by population movement.

**Sub-populations: filters for population behaviours**

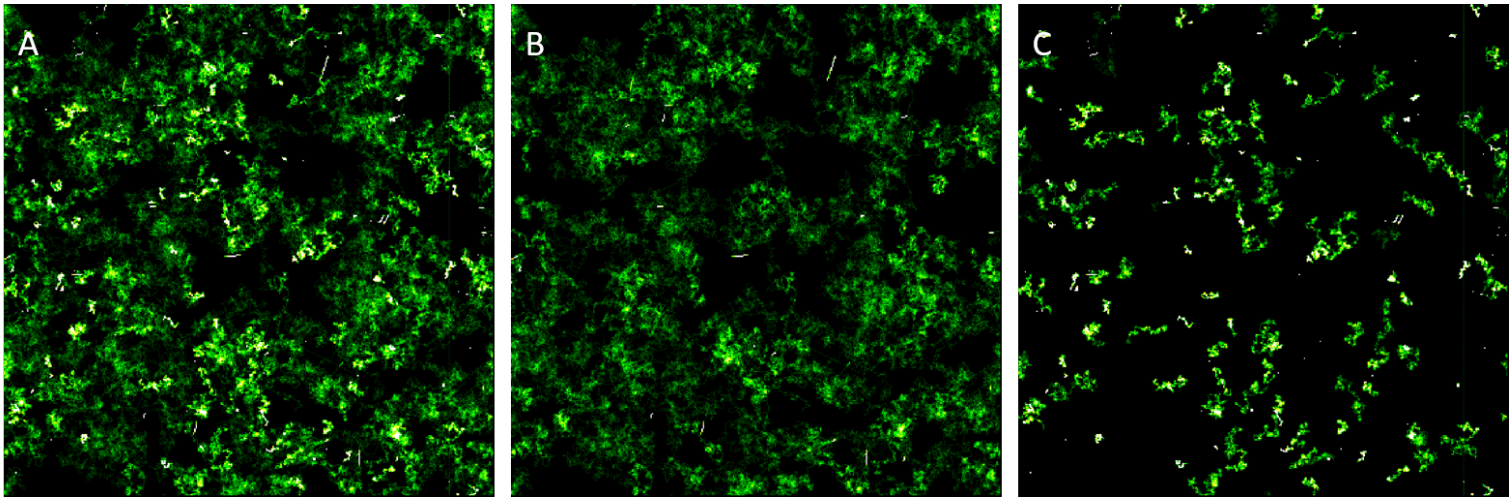



Figure 4.3: Movement heatmaps (2.2.2) representing behaviour over time of an initial random movement with stopping sub-population (A) and the population split roughly in half by their preference for and against forward movement and directional selection (B,C)

Populations can often be divided into subsets such as affected and unaffected members, or members displaying a specific behaviour and those that don't. This can be caused by an underlying biological sub-population due to clonal heterogeneity in the case of cancer cells, possibly affecting motility and replication [80]. In the case of GPCR and G protein the effects of highly localized interaction of environmental variables like an attractive or restrictive zone may define subsets. In the GPCR and G protein, we have observed a strong rear bias in both receptor and G protein populations and reproduced similar trends with a small shivering sub-population simulation. We therefore, attempt to split the population based upon this rear bias to isolate each.

Within the framework we can implement a tool that marks entities throughout a run based upon fitting any of the specific metric criteria generated. Once marked, a population can be filtered inclusively or exclusively, input trajectories are sorted into two new movement subsets. Comparison of filtered sub-populations supports the relationship between causal separation metrics and observed patterns; whether due to interaction with other entities or environment. After defining and splitting a population they can be input and visualised by analysis tools in the same way as normal unfiltered data sets (Figure 4.3).

#### Model Hyperparameters

Across the chapter we reuse the model definitions in the previous chapters. The corresponding hyperparameters are therefore the same, but the visualisation tools differ. Cancer (Tables 2.2 and 2.3) and GPCR and G protein (Tables 3.1 and 3.2) models therefore reproduce movement patterns even if different stochastic seeds are used.

## 4.3 Results

### 4.3.1 GPCR and G proteins

In initial analyses of the GPCR and G protein system, general random movement could be seen across the biological system. Of note, the greater inclination of G proteins (C2) was to uniformly higher movement density while GPCRs (C1) created movement clumps suggesting greater compartmentalization. Near particularly large movement hot-zones, both C1 and C2 data shows a reduction of localized general movement, a starvation effect. High movement hot-zones have also been confirmed as present with varying size and constraint strength. Both small and large hot-zones can be consistent across C1 and C2 but population placement suggests large ones are more likely G protein-receptor congregation points, important to co-localization and biological processes.

Rear bias is likely a true signal and perhaps an effect of entity binding or reflection, directional maps should help localize the phenomena. As noted previously 3, small sub-populations can bias trends. Therefore, identifying where rear biased patterns occur, filtering based on the bias and comparing movement heatmaps may reveal the causal interaction.

#### Micro-environmental turn preferences

We expected to localize and understand the interactions that produced observed rear bias and hot-zones by applying directional heatmaps. Representative model analysis should also improve, more detail allows a greater scope for comparison of resultant patterns, .

**Real data results** Absolute, constant, and relative turn directional heatmaps can be segmented to focus upon a specific hot-zone of interest with surrounding movement (Figure 4.4). With examples of both small and large hot-zones in focus, we can differentiate them with directional heatmaps showing preference at turn steps. Smaller hot-zones are characterized in the absolute directional turn map as generally turning towards the centre of movement as might be expected. The relative directional heatmap displays pronounced rear bias within the smaller hot-zones, supporting the narrative of bias by these extremely restricted. For both absolute and relative directional visualizations, the larger hot-zone is clearly denser than nearby movement but seems to have almost random patterns as entities approach a small area of centrally biased movement. As with previous observation directional starvation is also clearly present. There are lower levels of dominant directional selection outside of hot-zones suggesting local populations can co-localize within.

**Representative model results** General Brownian motion shows reasonable spread without pronounced hot-zones, there are also local changes in direction consistent with expected randomness. Combination of multiple population member movement over time creates areas of general random directional motion. By adding a small sub-population of shivering low motility population members small hot-zones appear. Shivering hot-zones show general preference for centrally pointing directional selection and a greater prevalence of dominant turn direction than other surrounding movement. Since there are fewer active entities an immobile sub-population also reduces overall environmental exploration like observed movement starvation.

As with normal movement heatmaps, the restriction of available space via deflective boundaries shows clear areas of motion starvation that can be identified as the deflective curves. Where population members deflect, clear directional preference trends point away as they bounce off, motion also runs alongside area walls. Caught between two boundaries, a small hot-zone is created with movement primarily toward the centre. However, hot-zones are not as clearly defined or as present as in other model results. In contrast, the attractive zone implementation creates clear hot-zones and localized movement starvation without extended observable restrictive paths.

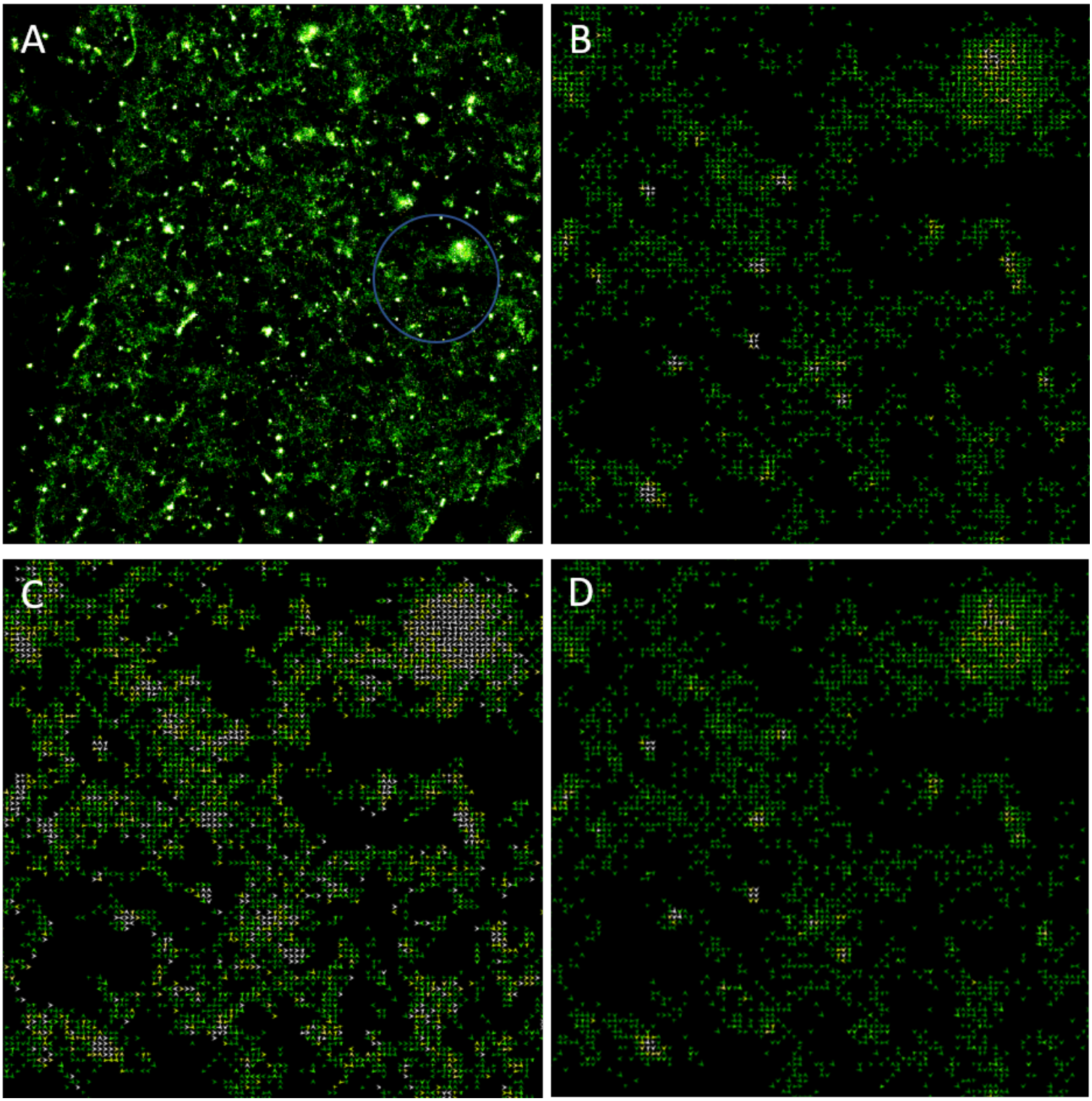



Figure 4.4: GPCR and G protein run TC641 C1 visualisation comprising of a normal movement heatmap(A) (2.2.2) and directional heatmaps (4.2.1) for absolute turn (B), absolute constant (C) and relative turn (D) subsections of the initial run (A)

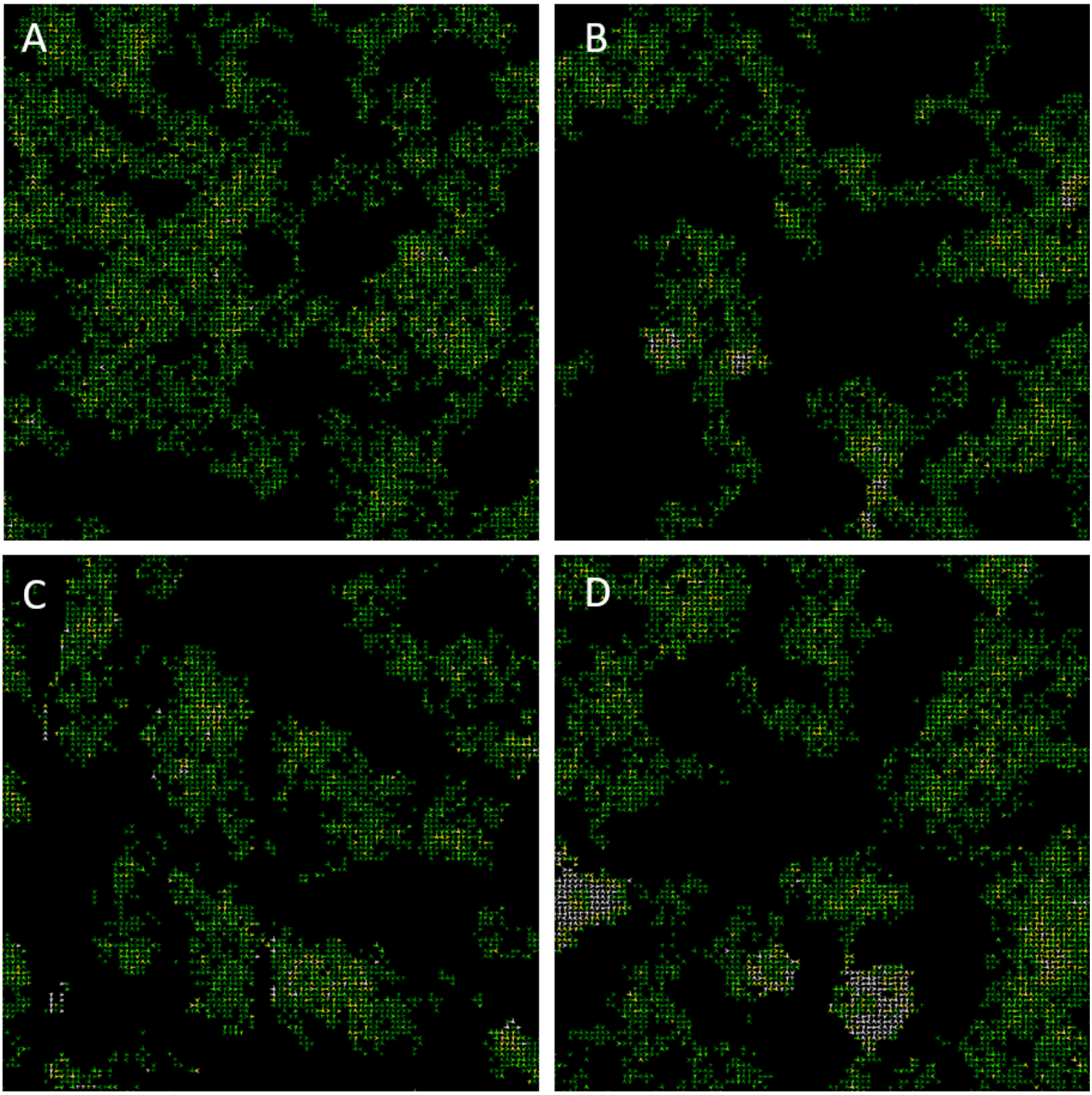



Figure 4.5: Absolute directional subsection representations of results for C1 movement (A), immobile sub-population (B), deflective curves (C) and attractive area (D) models. In absolute direction heatmaps (4.2.1) at turn the arrows represent dominant chosen direction of local population members at each turn choice increment.

Movement within attractive zones also seems to be undirected and somewhat cyclic in nature. Attractive zone implementation creates the cyclic effect via gradual centrally biased trajectory modification, a calm unrestricted area in the centre allows resumption of random movement similar to some of the observable hot-zones in the real-world *in vitro* data sets (Figure 4.4).

With relative display, the model of general Brownian movement shows random directional spread and environmental exploration, confirming that any subsequent patterns are likely due to the added interactions (Figure 4.6). Immobile sub-population models create clear hot-zones and movement starvation. Deflective boundary implementation models display reverse preference along hard small zone barrier intersections, more permissive space leads entities to run alongside with extended directed micro forward turns. Again, highlighting the implementation of attractive areas, hot-zones show strong forward led micro turns around a central more random permissive area.

#### Short and long term system effects

Time phased based observation allows us to visualize movement that has only occurred between two selected time points. For the purpose of this section each time phase is made of 100 time increments and 3 sub jumps per increment over 400 total increments as in the GPCR and G protein set. The top-end cap for heatmap gradient is also the same across all visualized runs to aid effective comparison, each shade in a figure represents the same value within a gradient with white being highest movement density and black 0. It is expected that patterns with long term slight effects will become less prevalent but that cumulative and static interactions should be consistently visible across each phase.

**Pattern persistence and hot-zone morphology in real-world set** Over the course of 100 time increments there is enough space for general movement to generate and display patterns similar to purely random movement. Uniformity between visual patterns across phases is therefore likely representative of population members with consistent movement, environmentally or behaviourally driven in any area.

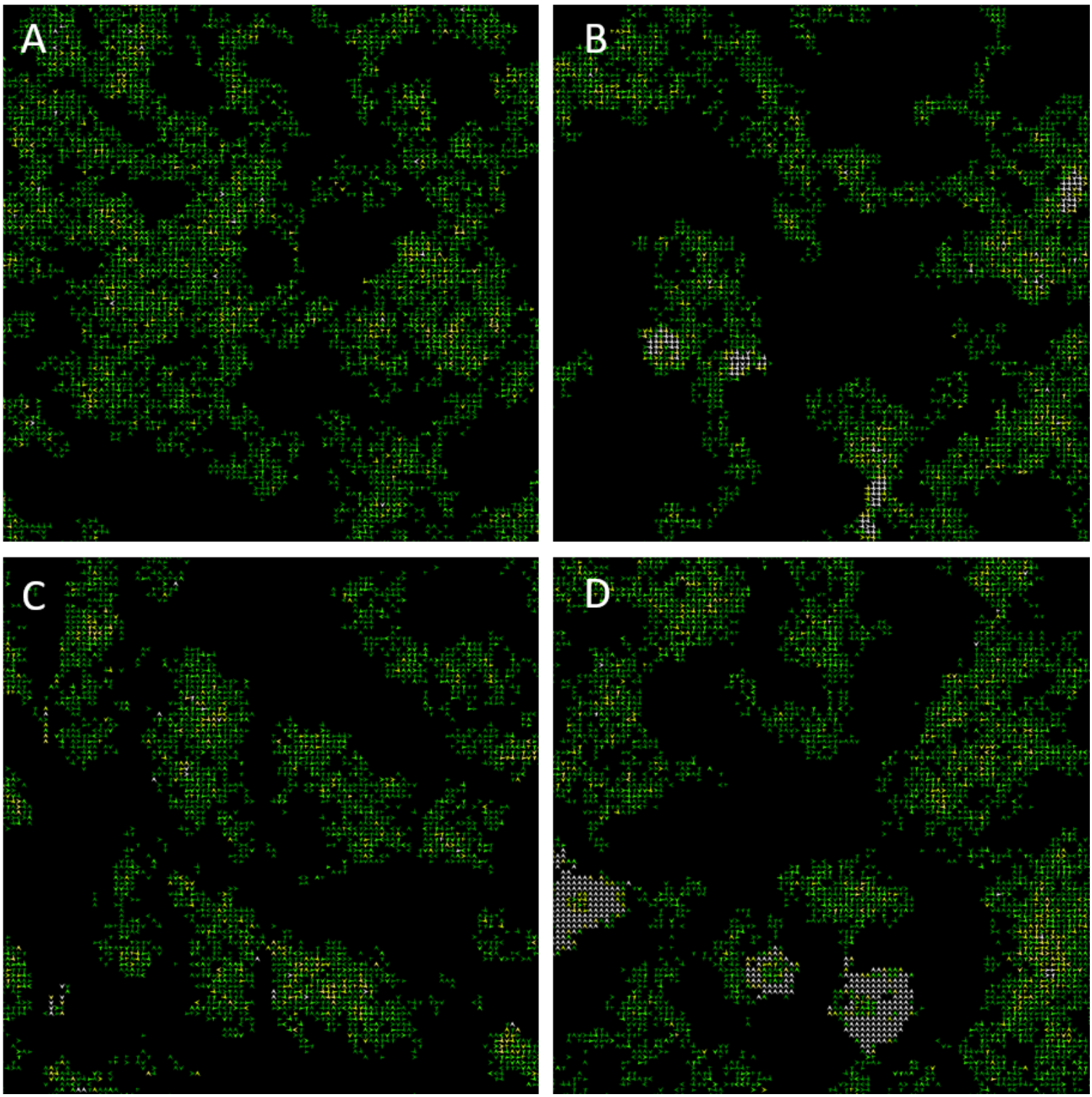



Figure 4.6: Relative directional subsection representations of results for C1 movement (A), immobile sub-population (B), deflective curves (C) and attractive area (D) models. In relative direction heatmaps (4.2.1) at turn the arrows represent dominant chosen direction of local population members at each turn choice increment in relation to their previously chosen vector, north facing arrows mean micro turns and south facing reverse favoured modification.

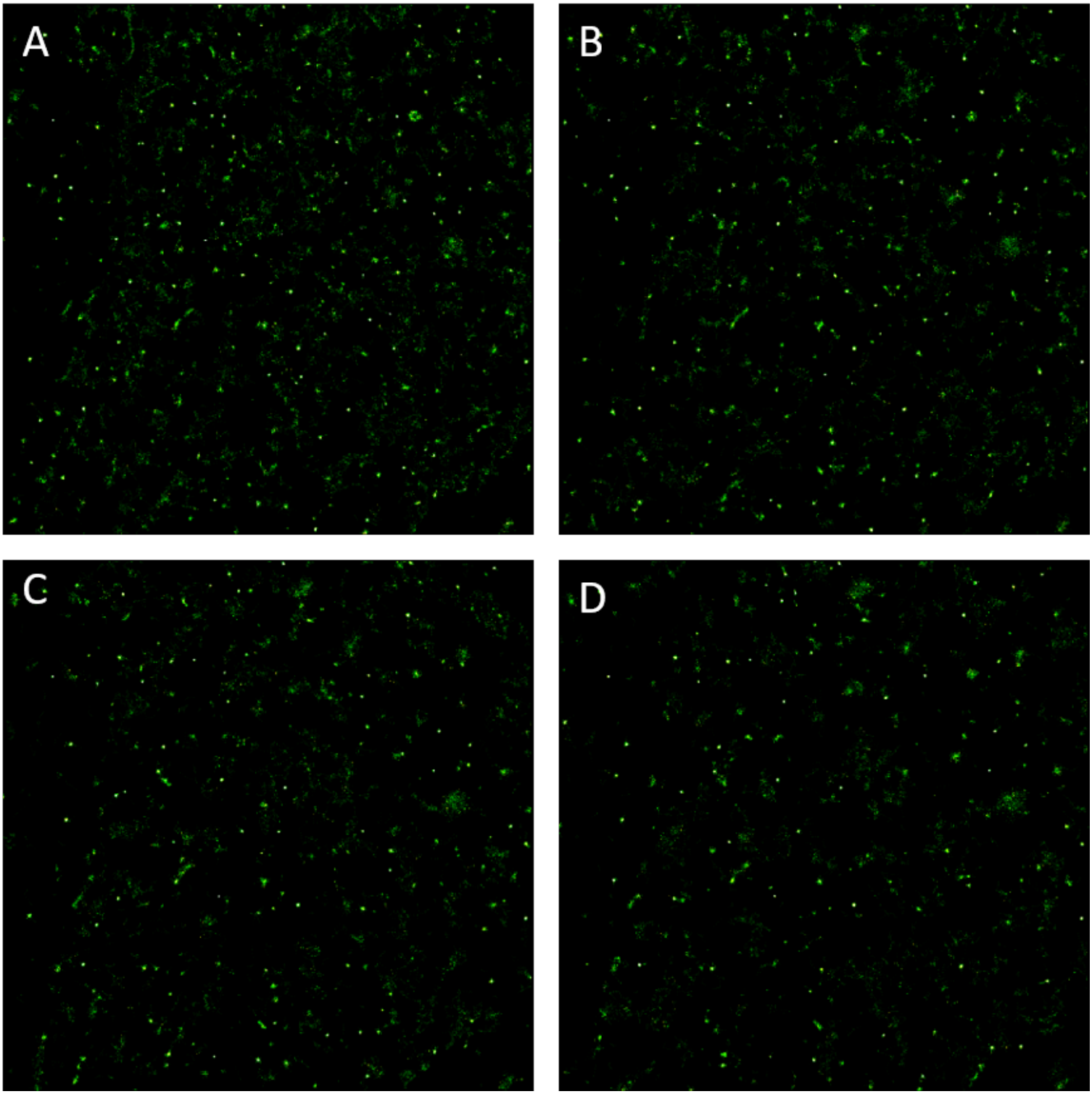



Figure 4.7: GPCR and G protein data set TC641 C1 population movement heatmaps (2.2.2) measured within 4 contiguous stages with movement only occurring between each, 0-100 (A), 100-200 (B), 200-300 (C) and 300-400 (D) time increments were visualized

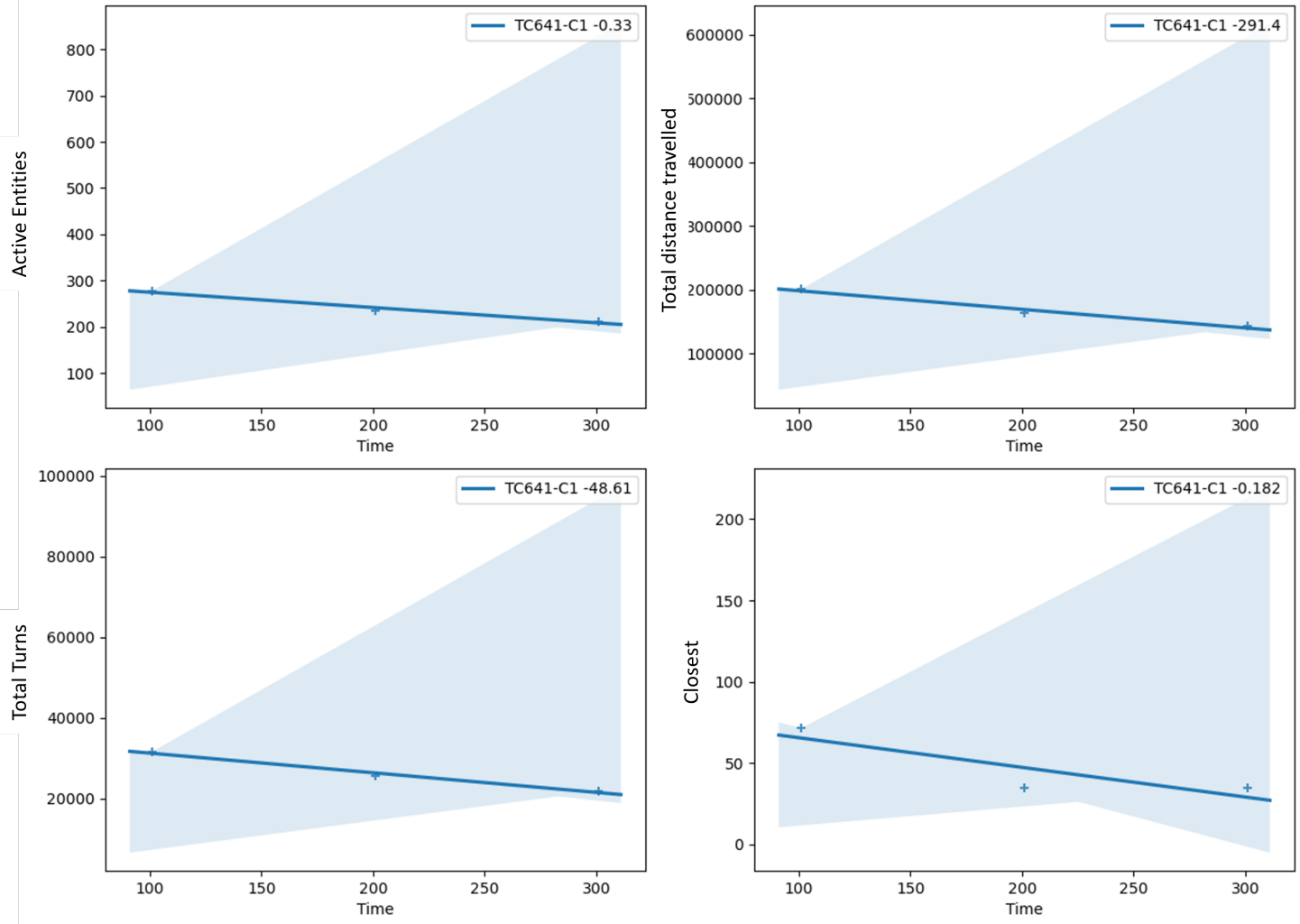



Figure 4.8: GPCR and G protein data set TC641 C1 active entities, total distance travelled, total turns and summation of entity adjacency across the population within each phase of 0-100 , 100-200 , 200-300 and 300-400 time increments.The number by each set label is the slope of the regression model result line and the darker area is a 95% confidence interval for the regression.

Recurrent hot-zone location across time phases supports the consistency without specifying if they are caused by long term immobility or constraint. The shape of larger hot-zones can be seen to change in several distinct patterns: random movement within a roughly uniform area as might be expected for constrained co-localization over time, coalescence and dissolution. Interestingly, there are several examples of larger hot-zones that seem to come from smaller hotter areas and disperse over contiguous time phases (Figures 4.7,4.8). Additionally, there appear to be larger strand-like patterns that persist across runs, possibly a similar phenomenon to that which creates hot-zones or adjacency of several active restrictive boundaries in the real-world *in vitro* system.

**Immobile sub-population results** One of the models generated for comparison with GPCR and G protein system consisted of both Brownian background motion and a small sub-population of shivering entities (Figure 4.9). When phases are applied, we observe consistency from the position of immobile sub-populations and ongoing unchanged presence of subsequent hot-zones. More transient features such as specific background movement patterns vary across phases, they do however stay generally localized where greater concentrations of population members are travelling. Entities can maintain locality across phases but not specific shapes. This indicates low relative area exploration by the background Brownian motion for our representative models.

**Deflective boundaries and representative models** When we implement strongly deflective curves representing restrictive boundaries, we are usually able to observe clear paths of movement starvation in heatmap representations (Figure 4.10). During time phase observations, the lower prevalence of general background movement obscures the strand pattern. We can identify several isolated pockets of movement that are likely constrained but even over time it does not lead to distinct hot-zones, highlighting a difficulty in the restrictive model. Where boundaries are strong enough to restrict movement out, they necessarily also reduce entity ability to enter a space. General population movement may be prevented in some areas, however, without addition of more population members larger hot-zones become an aberration of initial placement and consequent catchment in the model.

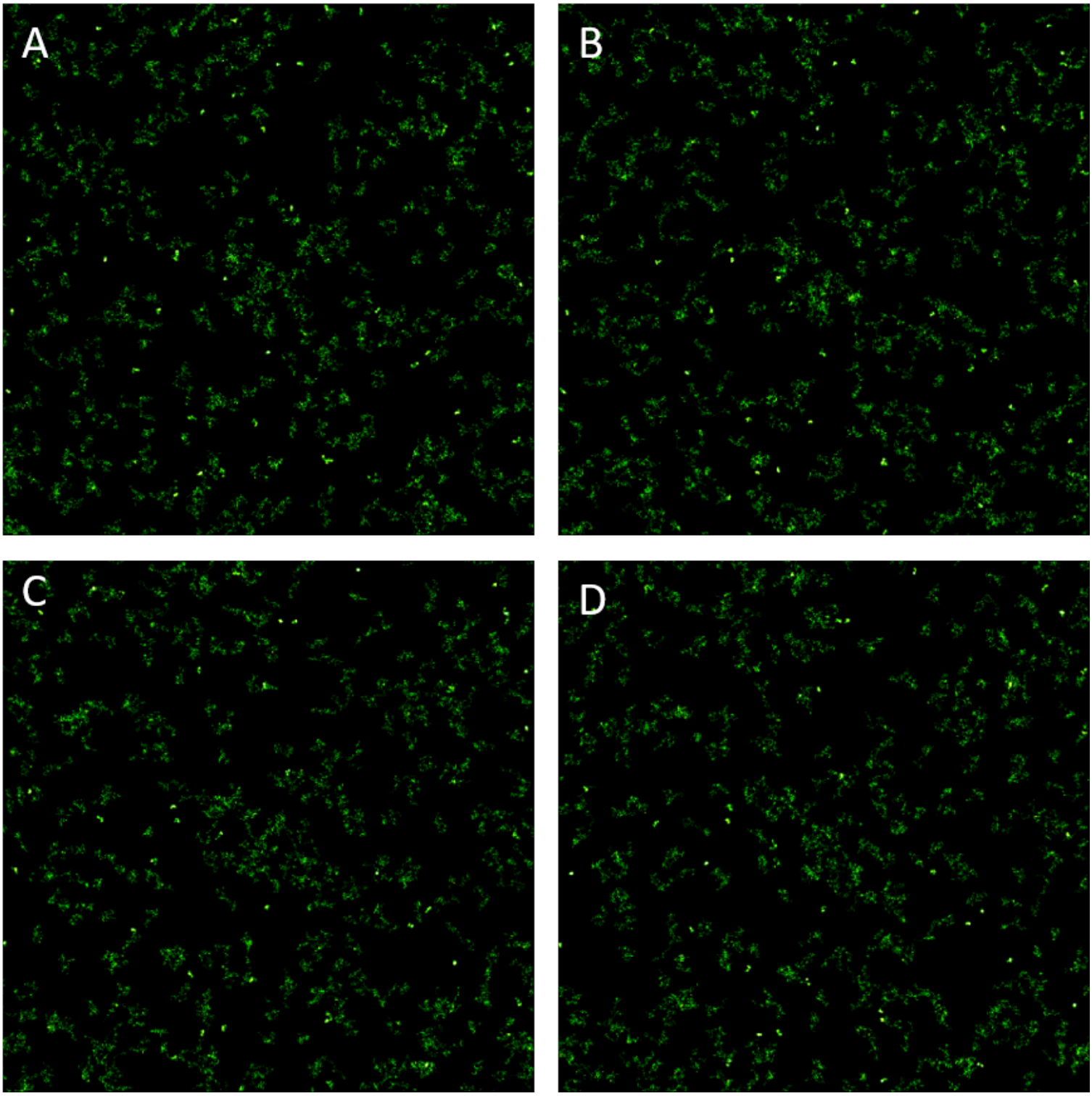



Figure 4.9: C1 background Brownian movement model with added immobile sub-population representation movement heatmaps (2.2.2) measured within 4 contiguous stages with movement only occurring between each, 0-100 (A), 100-200 (B), 200-300 (C) and 300-400 (D) time increments were visualized

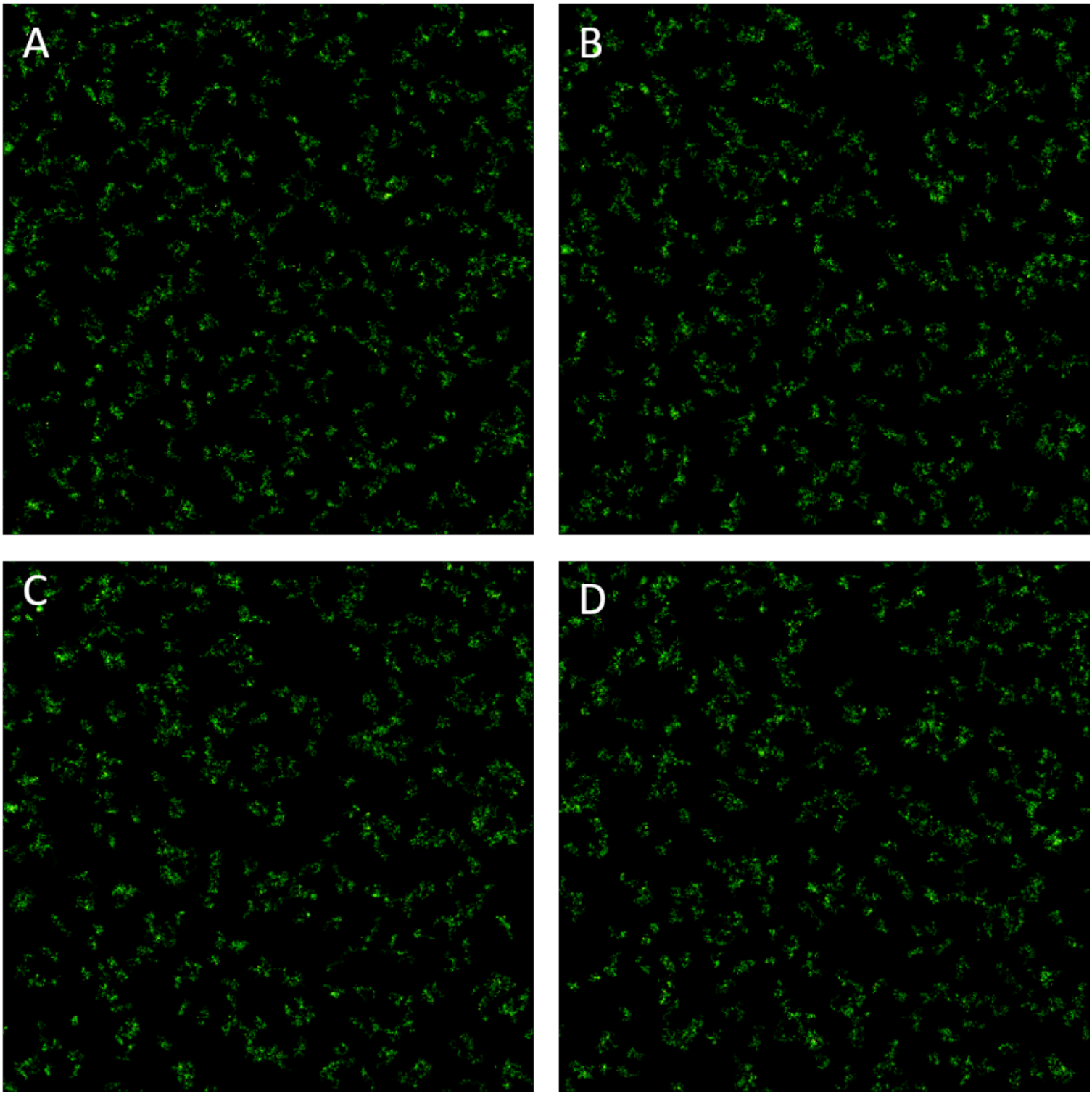



Figure 4.10: C1 background Brownian movement model with added deflective curve representation movement heatmaps (2.2.2) measured within 4 contiguous stages with movement only occurring between each, 0-100 (A), 100-200 (B), 200-300 (C) and 300-400 (D) time increments were visualized

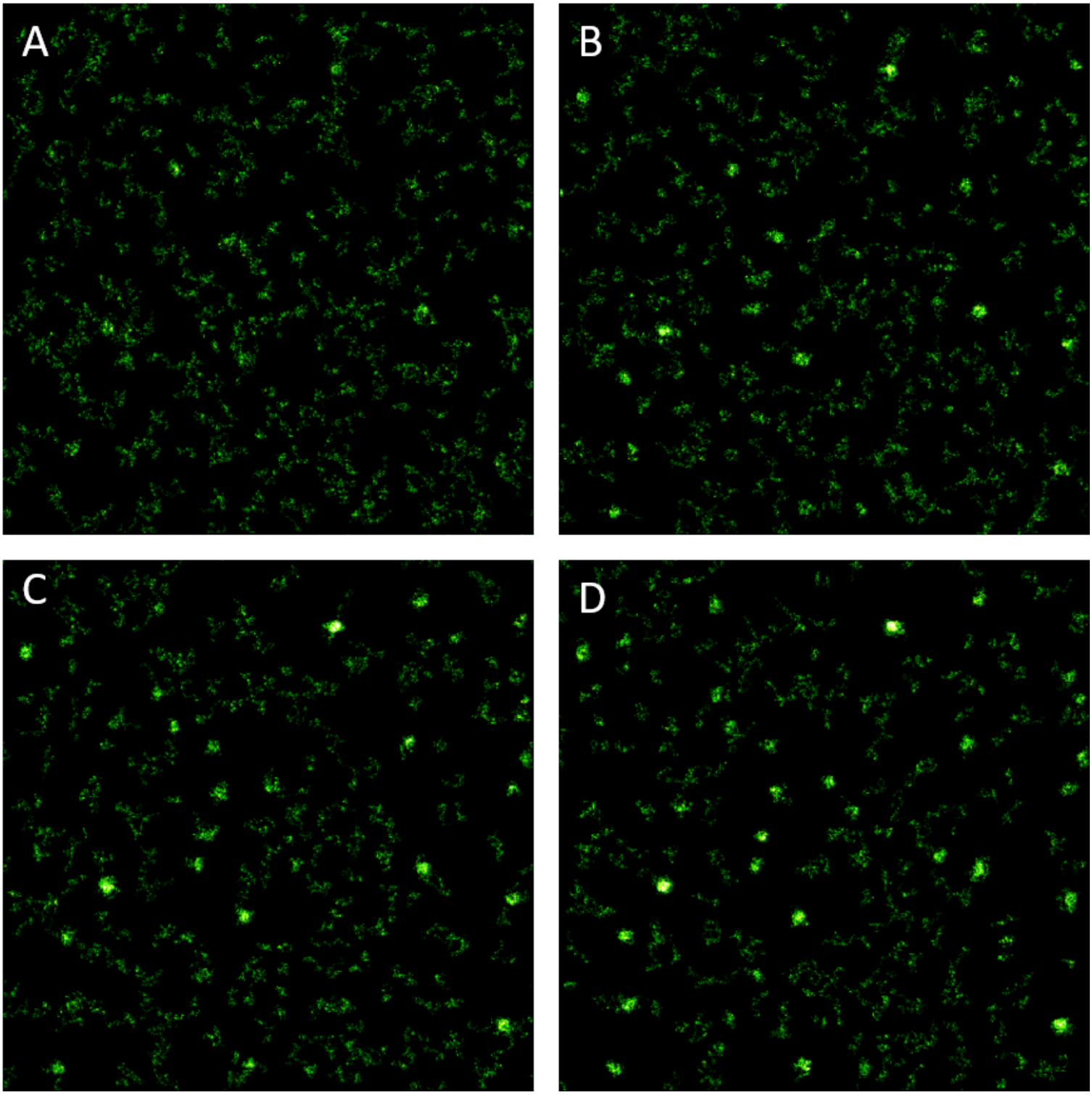



Figure 4.11: C1 background Brownian movement model with added attraction zone representation movement heatmaps (2.2.2) measured within 4 contiguous stages with movement only occurring between each, 0-100 (A), 100-200 (B), 200-300 (C) and 300-400 (D) time increments were visualized

**Attractive zone representative models** Previously, we added a representation for zones of entity attraction as an alternative explanation to restrictive boundary theories outlined in literature [56, 58]. While we found exploratory models implementing them to be more compelling, they still caused some difficulty in definition due to the interaction of attraction strength and morphology: too strong an attraction variable leads to very small dense hot-zones, too weak and they don't appear. However, size was not the only morphological aspect of hot zones; neither small or large entirely represented *in vitro* observations. Time phases allow us to visualize the life-cycle of strong attractive micro domains and identify possible comparable points (Figure 4.11).

Across each time step, attractive zones coalesce from general movement, pull in surrounding entities, and exist as clear areas of heat within dark surroundings. We can also observe the ability to create local movement starvation patterns over time without overly affecting general simulation area movement density, general background movement remains clear throughout. Also, with the addition of phases, heterogeneous shapes can be observed as part of the coalescence process, mirroring more closely hot-zones in the real data source but lacking later dissolution.

**Sub-populations**

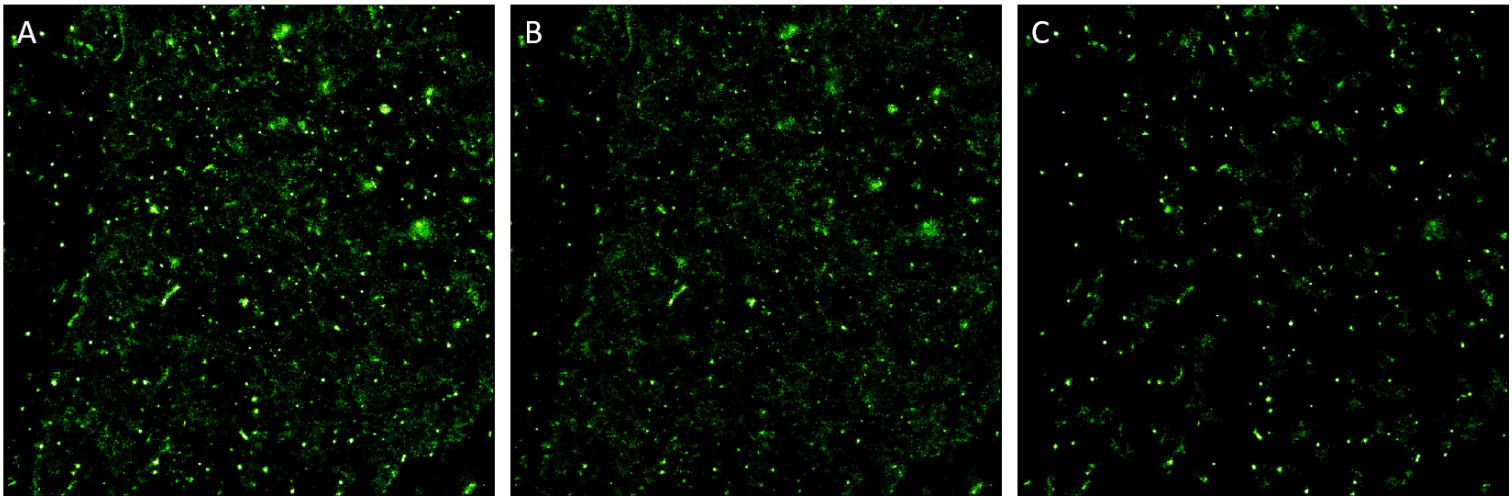



Figure 4.12: Movement heatmaps (2.2.2) for GPCR and G protein set TC641 C1 over the entirety of the set time-frame (A), the population was then sifted based upon number of rear turns 4381 tracks with fewer than 5 rear turns (B) and 395 with greater than 5 (C) and movement metrics generated

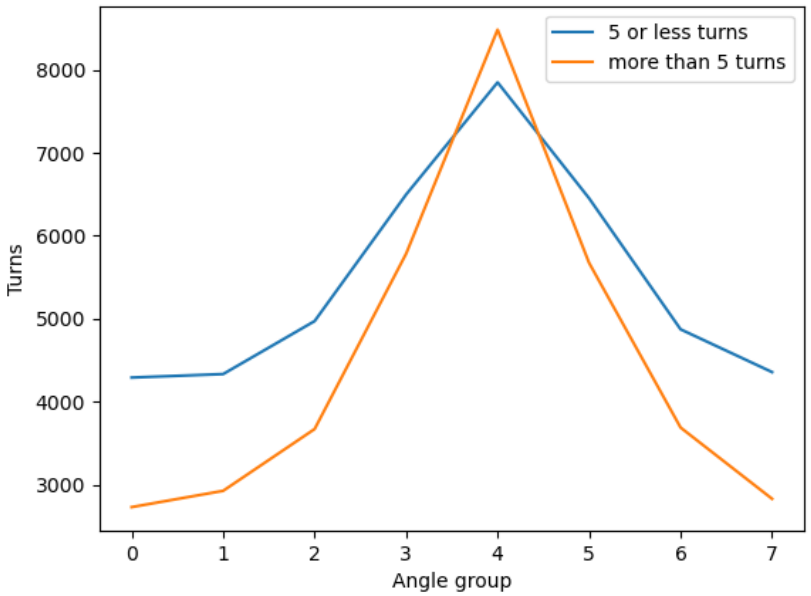



Figure 4.13: GPCR and G protein set TC641 C1 post sifting into tracks based upon number of rear turns 4381 tracks with fewer than 5 and 395 with more (Figure 4.12)

Once a hypothesis is developed for movement patterns, we can filter a population into subsets that conform to the trend and those that don't for comparison. In this case we assume that the rear bias is sourced primarily from hot-zones, as supported by directional heatmaps. Therefore, the population of a GPCR and G protein data set was divided based upon a broad turn trend definition into members with differing rear turn amounts and the results visualized (Figure 4.12).

Tracks, and therefore real entities, with greater rear turns do primarily dwell within hot-zones. However, it is not possible to say they only exist there as hot-zones are still clear across both new populations. Larger hot-zones and general background movement are better represented in the lower turn sub-population. As with the findings of directional heatmap analysis, small hot-zones are more clearly rear turn biased than large ones; the rear bias is also more pronounced in the higher turn set (Figure 4.13).

### 4.3.2 Cancer cells

Our original analysis of cancer motion identified forward preferred and striated strand-like movement patterns over time. Both built in lattice paths and entity-based path forging models showed

similarity and possible explanatory application. Lattice paths were more effective for creating clear movement strand visual patterns through areas of general movement. Path forging behaviour creates more organic and stochastic patterns. A combination of the two may be present in the real-world environment since both are supported in existing literature [124, 92, 80].

### Micro-environmental directional effects

Notice should be taken of directional motion within areas of general, seemingly random and populated movement patterns, differentiating whether these attractive strands continue or disperse (Figures 4.14,4.15).

### Patterns in real data

**Short runs** For the shorter time frame cancer sets (A-D), strict adherence to strand-like movement was observed. With the directional heatmap visualization, clear movement strands are still visible (Figure 4.14). Further, the relative turn heatmap displays widespread dominance of upward facing arrows, in a relative diagram this means forward facing directed movement is pervasive. Visualization of continuous absolute direction preference shows an even stronger following pattern, due to the large number of jumps between time increments. Population members are clearly following strand-like paths. A hub of paths can also be seen at the intersection of several strands where movement becomes more undirected, the general area is more permissive. When movement does become more generalized a forward preference of direction is less clear but still present, there is also rapid coalescence into strand following when leaving generalized areas.

**Long run** As initially observed, the longer 181 time frame cancer set E displays a more general movement pattern with some observable strand following still visible. It was suggested that due to population growth patterns and comparison with shorter time frame cancer sets, that it was captured at an earlier extended time frame. When we study set 5 with directional heatmaps, previously observed trends of widespread general movement with some strand following behaviour remains (Figure 4.15). There are subsections of movement containing a mix of well-travelled and differentiated paths, there are also more general movement areas. Similarly, relative directional

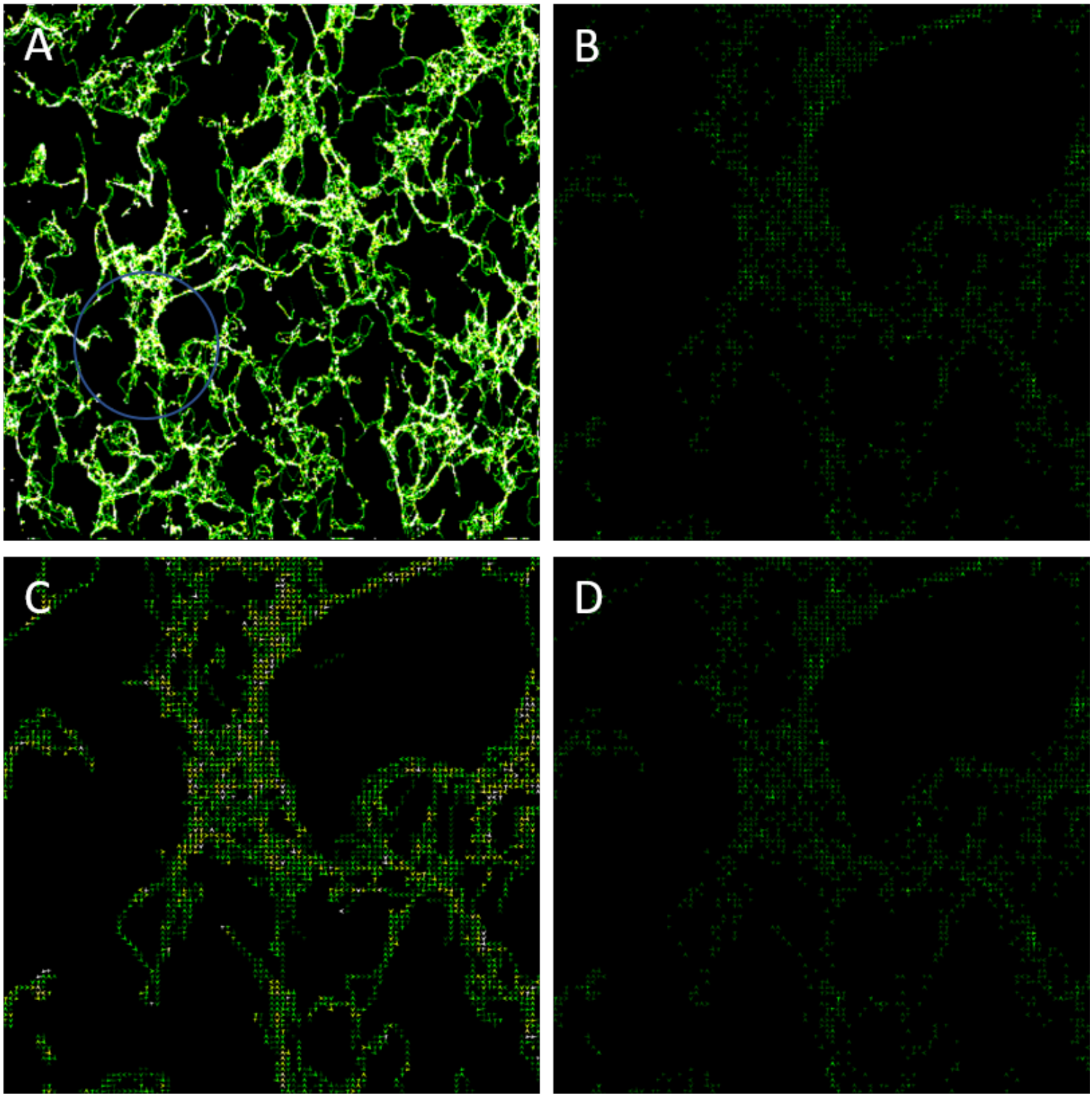



Figure 4.14: An 84 step cancer movement files is visualized as a movement density heatmap (2.2.2) (A) and then a sub section displayed as absolute chosen direction (A) and relative directional preference (4.2.1) (D) at each larger time increment along with absolute heading direction at every jump step (C)

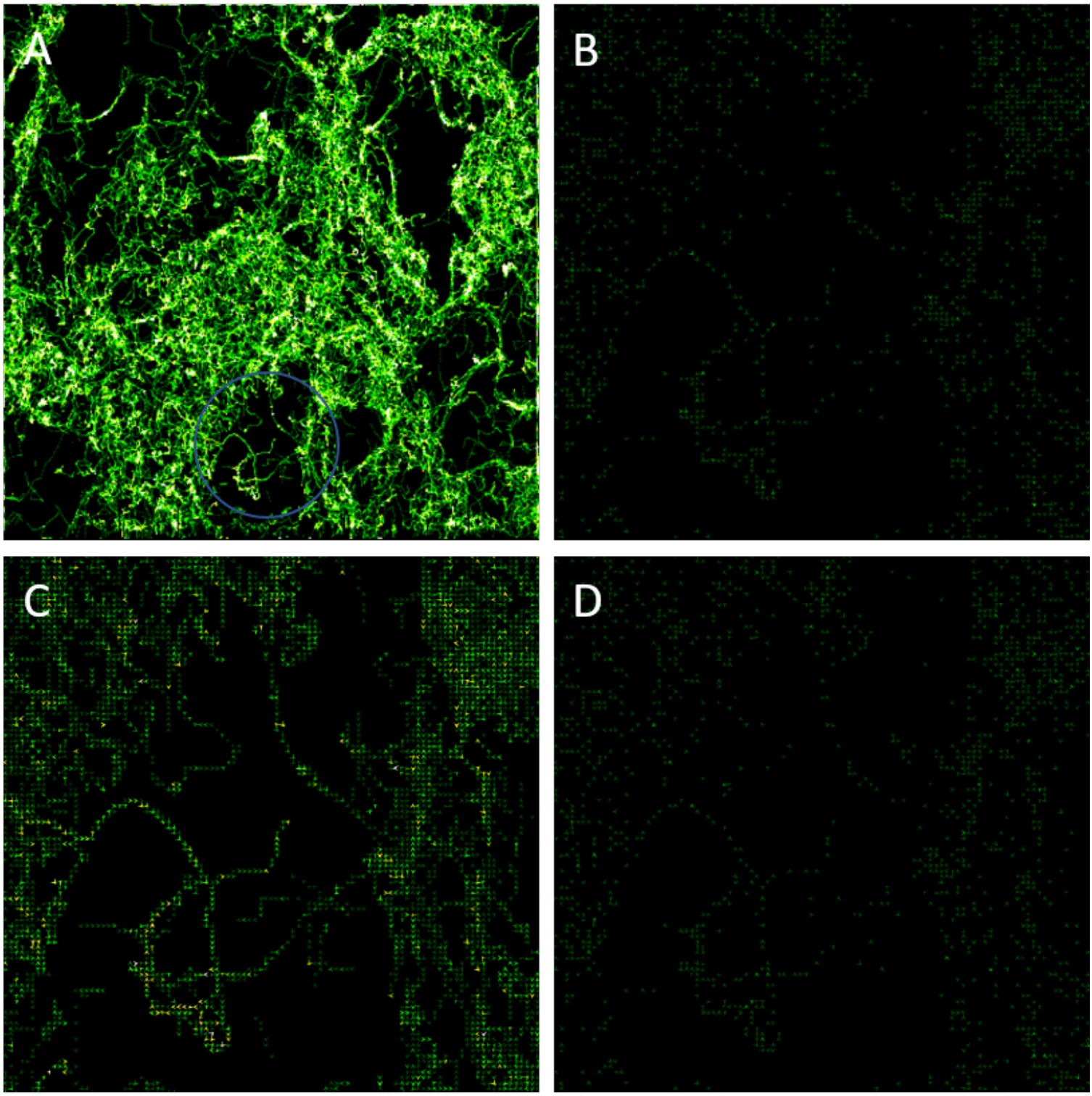



Figure 4.15: An 181 time step cancer movement file is visualized as a movement density heatmap (2.2.2) (A) and then a sub section displayed as absolute chosen direction (A) and relative directional preference (4.2.1) (D) at each larger time increment along with absolute heading direction at every jump step (C)

maps suggest that forward preferential movement is still dominant in the strands and in more general areas, the amount reduces as an area becomes more travelled. There are two examples of general movement that can be observed, one displays close knit but still distinct webs of strand movement and the other a less directed random area.

**Comparison with representative models** We can represent population travel and directed preference with a constant absolute directional heatmap from a general cancerous movement model (Figure 4.16). The forward biased pattern our general movement model represents is derived from overall population behaviour in observed real-world system. Therefore, replicating the trend creates similar movement without the required environmental drivers. While general movement areas are similar, the same shared strand cohesion is not present, movement strands are of random thickness and density due to random coincidental following.

The directional trend approach to visualizing lattice paths of least resistance suggests that such paths are too stark. They lack biological variance as shown by complete uniformity of direction and entities are siphoned away from general movement. However, lattice paths do create a strong strand pattern even when combined with general movement. The path forging and following implementation leads to the opposite issue, strands are difficult to discern and movement in areas of multiple entities becomes more undirected and general. We can observe some crossover of both phenomena by combining path forging and lattice path representations with constant absolute turn visualisation. Path following is still present and clear with the directional map suggesting some cut through of population zones. Equally, zones of undirected general movement suggest concentrations of population members also exist. Perhaps coincidentally, more striated general movement areas and patterns are also visible, a possible effect of chance or overall starvation of movement by entity attraction.

With relative turn diagrams we expect following entities such as cancer cells to show high forward preference (Figure 4.17). General motion models with the replicated turn trend show a strong forward preference, in an empty environment the only parameters to change that pattern are low random chance and entity to entity collision. Similarly, for lattice path following models, entities are drawn in and continue to travel along a very direct forward driven path, consistent

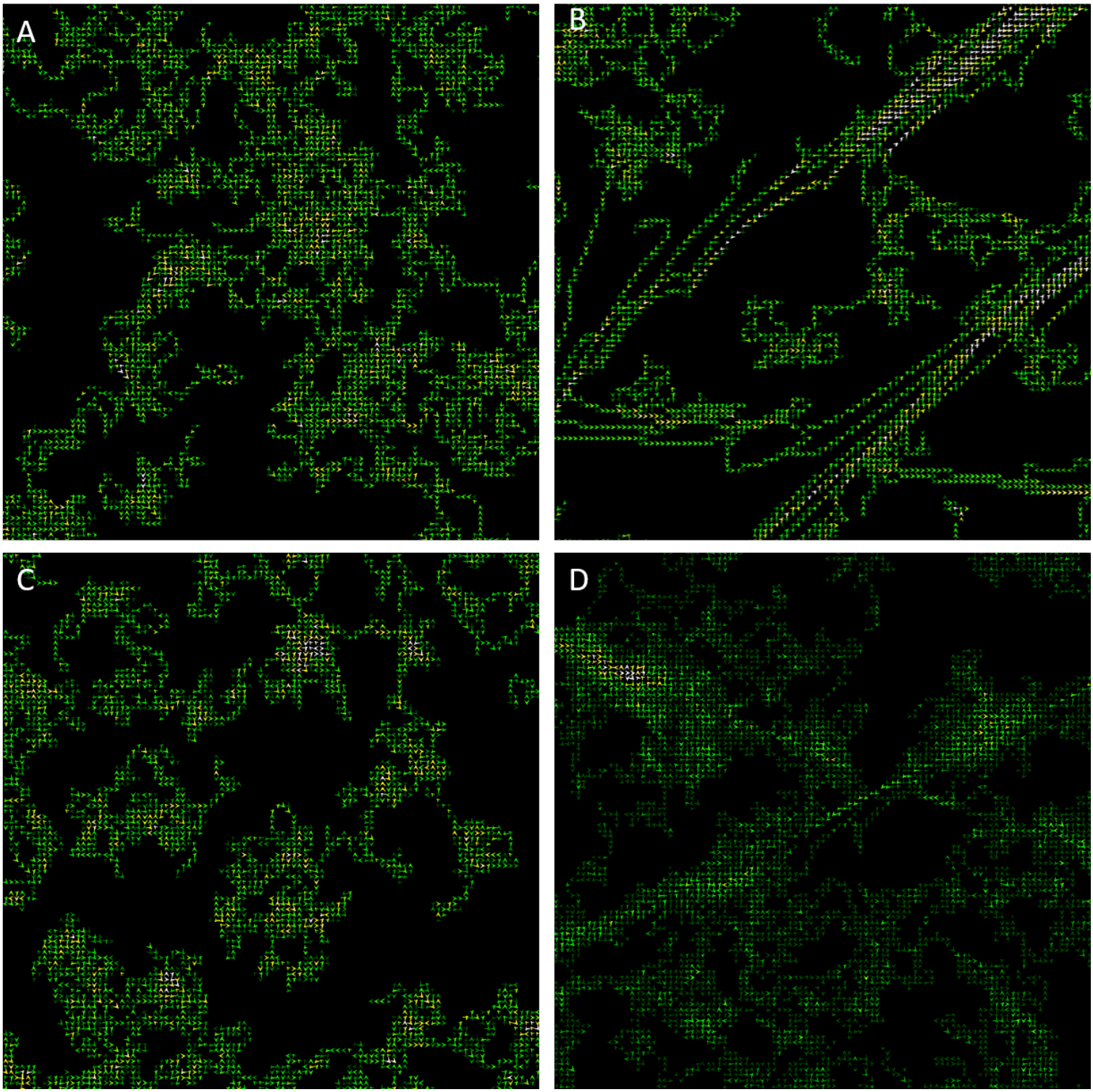



Figure 4.16: Absolute direction added movement heatmaps (4.2.1) where arrows indicate dominant movement direction at a location for general movement cancer (A), lattice attractive path (B), following (C) and hybrid following attractive lattice (D) models.

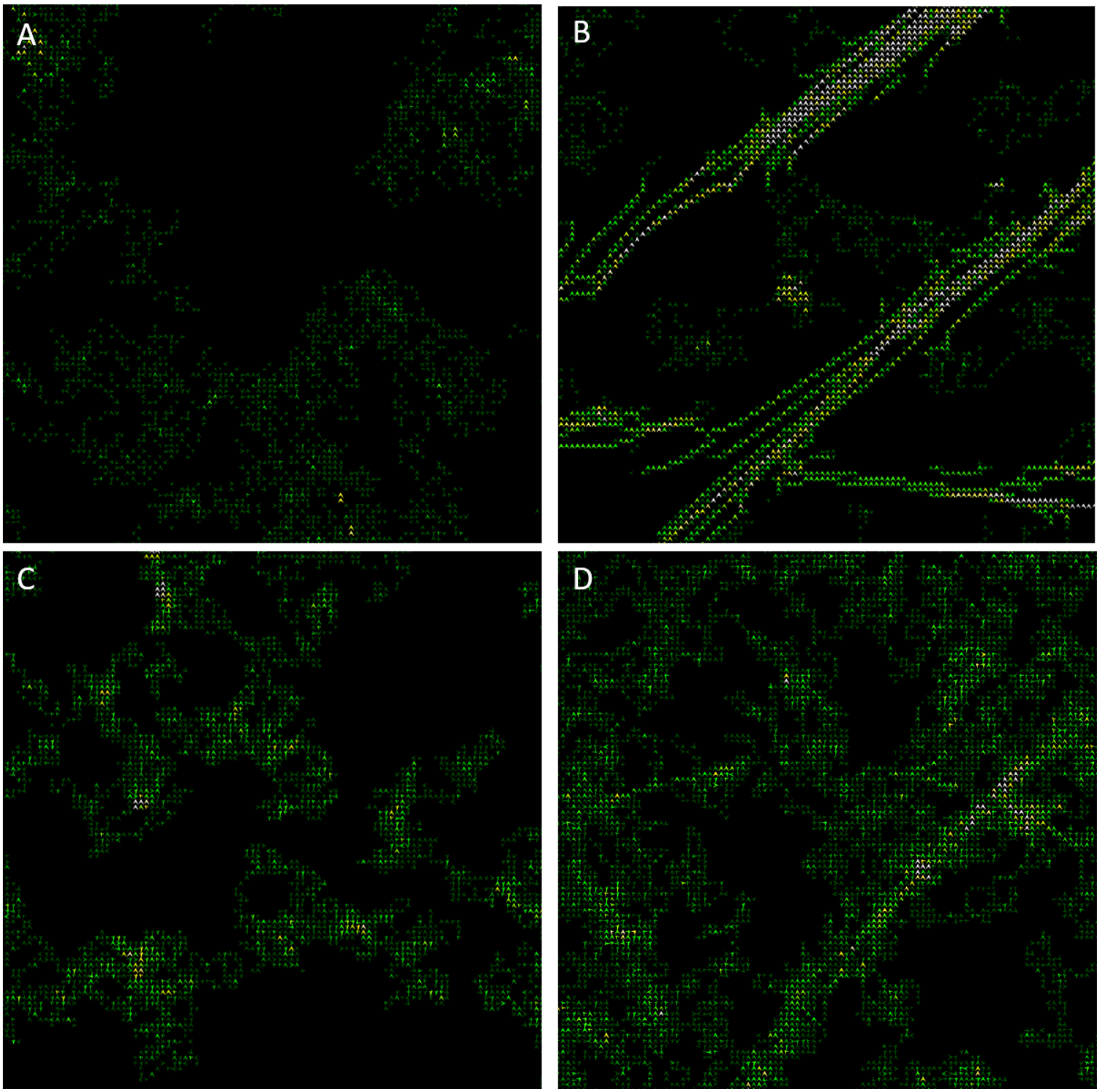



Figure 4.17: Relative direction added movement heatmaps (4.2.1) where arrows indicate chosen new direction relative to previous vector from north as 0 for general movement cancer (A), lattice attractive path (B), following (C) and hybrid following attractive lattice (D) models.

movement in any direction registers as forward movement. Path forging and following models show a more pronounced preference than general movement. The addition of least resistance following leads to entities consistently following previous paths compounding the foreword trend over time and eventually ending in highly localized hot-zones. Combination of broader lattice paths and forging again shows promise, clear strands run through general entity movement zones with strong dominant directionality remaining. We also observe increased general movement at path intersections with some undirected general movement permissive areas.

#### Short and long term pattern persistence within the system

In observations thus far, the primary possible effects identified for time isolated observation of the cancer data sets are pre-observation environmental effects. The larger set E may be a precursor with low strand cohesion progressing to the strongly established cancer environment A-D. Therefore, application of heatmap time phase restriction should highlight consistent trends and strand coalescence over time.

**Pre-tracking micro-environmental modification** Due to population growth patterns and comparison with data sets A-D, E may be captured at an earlier extended time frame. Comparison can therefore occur not only between phased model results and a real-world set but also across data types, short and long.

**Short sets** As with cumulative visualizations, the distribution of movement across the entire simulation space is clearly strand-like for the short runs A-D (Figure 4.18). We can also observe that these strands are not just strong short-term effects with time phase specific representation. They do not just bias visualization via a single interaction but lead to consistent long-term movement capture. Variance in exact movement patterns does occur across time but the broad strands of common movement remain, likely several population members traversing around the network. Greater specific heat consistency can be observed in general movement zones, fitting the narrative of population density leading to general environmental degradation. With more cells present, a greater proportion needs to leave to drastically change movement within a single

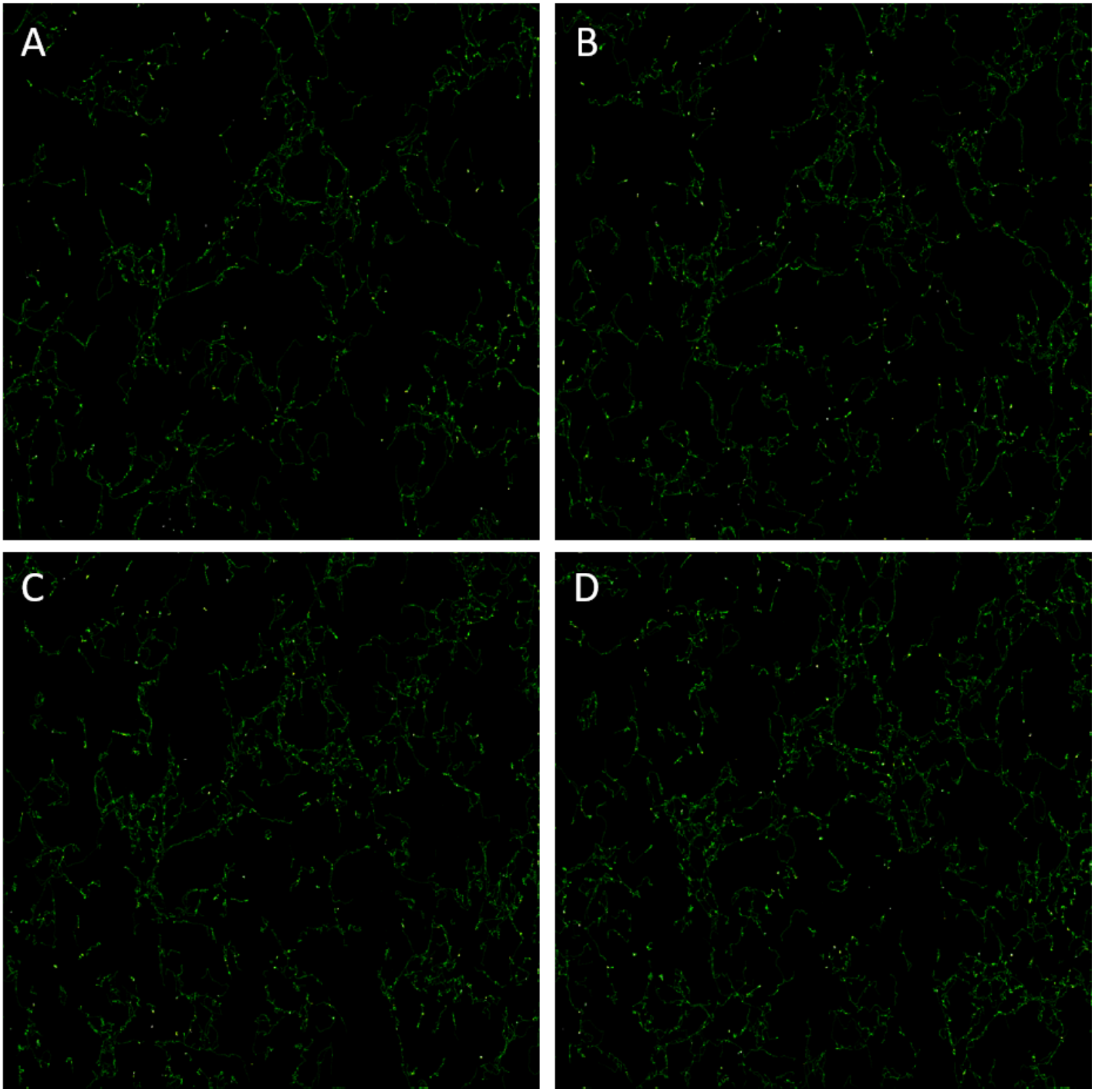



Figure 4.18: Heatmaps (2.2.2) representing population movement density over time for the first short 84 time increment cancer cell data set, heatmaps are representative of 0-20 (A), 20-40 (B), 40-60 (C) and 60-80 (D) increment subdivisions respectively

time phase of 20 increments.

**Long sets** Several interesting new overtime development patterns can be observed in the longer cancer set E with the time phase specific heatmaps (Figure 4.19). In the upper centre of heatmaps what begins as a concentrated group of cells spreads out into several more strand-like movement patterns over time. Conversely, and in line with our time frame hypothesis, general movement across the simulation space seem to coalesce from random to more strand like paths as time progresses. Even dense undirected pattern areas appear to become more defined; the question remains of whether the primary cause is the exploitation of built in lattice paths or simple natural least resistance path forging in permissive environments.

**Representative background movement models** We can visualize the normal extent of population movement without external environmental effects by separating a set of general cancer model heatmap movement representations into several smaller time phases (Figure 4.20). Movement is almost entirely disparate, commonality across time being based upon low general movement variables allowing entities to remain in a similar area. Therefore, the greater the localization of population members the higher the correlation of movement observed across phases. It also provides a good visual representation of serendipitous convergence for comparison with other models and real-world sets with no known environmental effects.

**Path of least resistance models: Lattice paths** With similar disparate general movement across phases to the first replica model, lattices add a pattern of common movement similar to that observed in the real-world sets (Figure 4.21). Paths become more pronounced as time passes and more entities are captured, they also broaden with greater activity. However, again, path shape is not consistent with the real-world strand phenomena and general movement zones, even less so. Uncaptured entities also seem to remain within their own area of influence creating small stark turn circles of movement in each short phase.

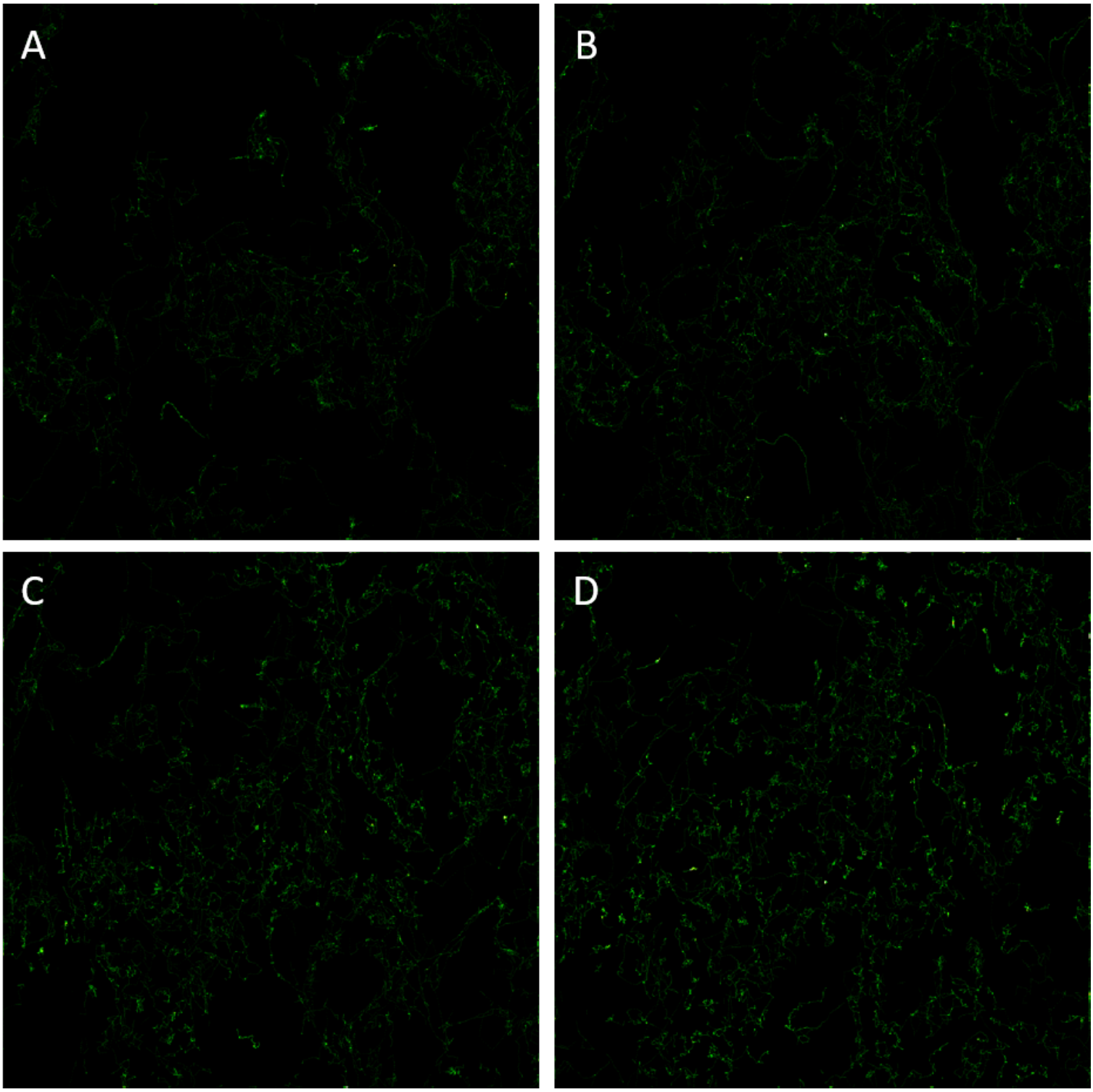



Figure 4.19: Standard heatmaps (2.2.2) representing population movement density over time for the longer 181 time increment cancer cell set, heatmaps are representative of 0-45 (A), 45-90 (B), 90-135 (C) and 135-180 (D) increment subdivisions respectively

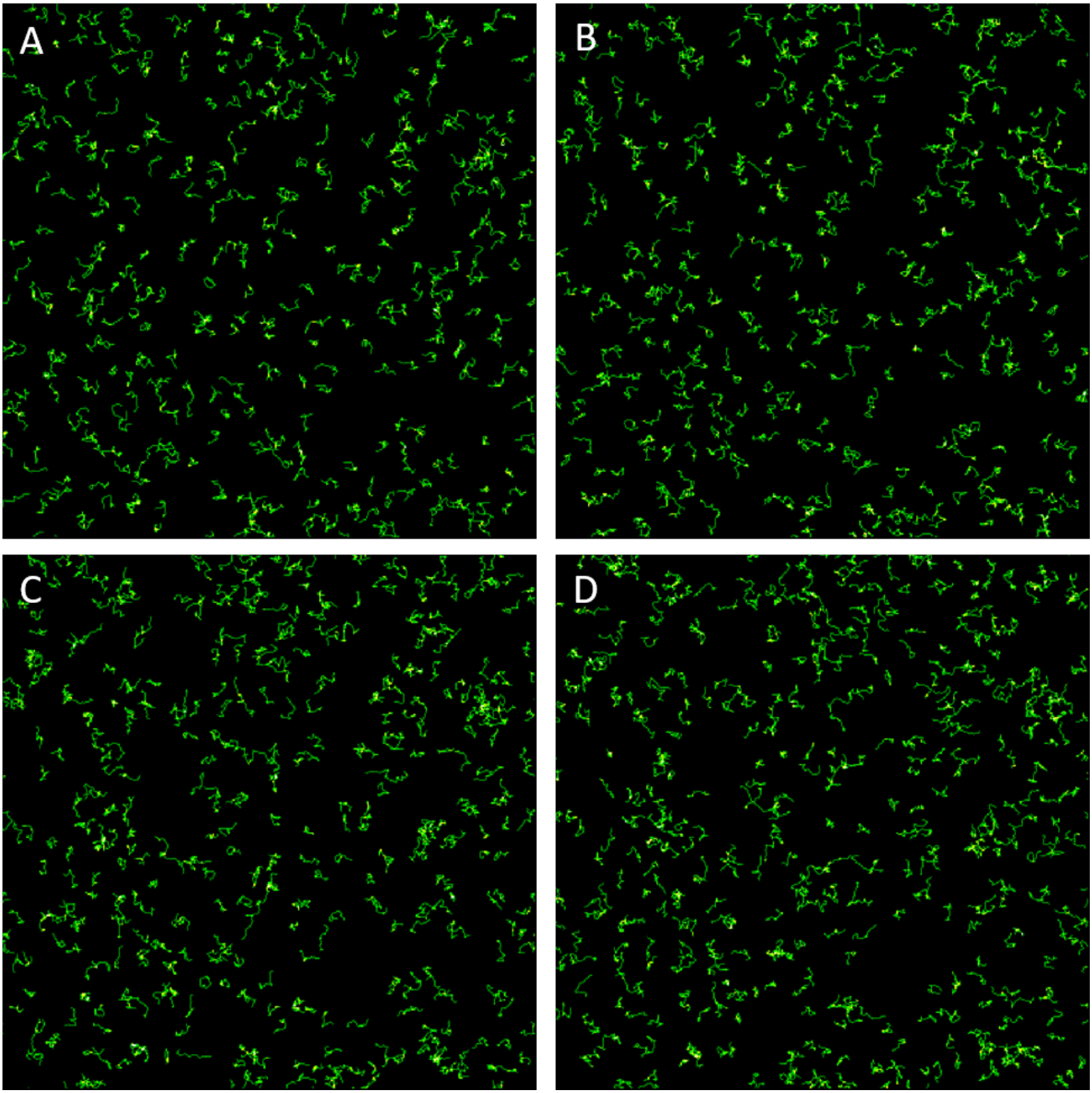



Figure 4.20: Heatmap (2.2.2) representations of movement density over time within set time phases of a model replicating general movement trends such as travel distance, population number and turn preference at 0-20 (A), 20-40 (B), 40-60 (C) and 60-80 (D) increments.

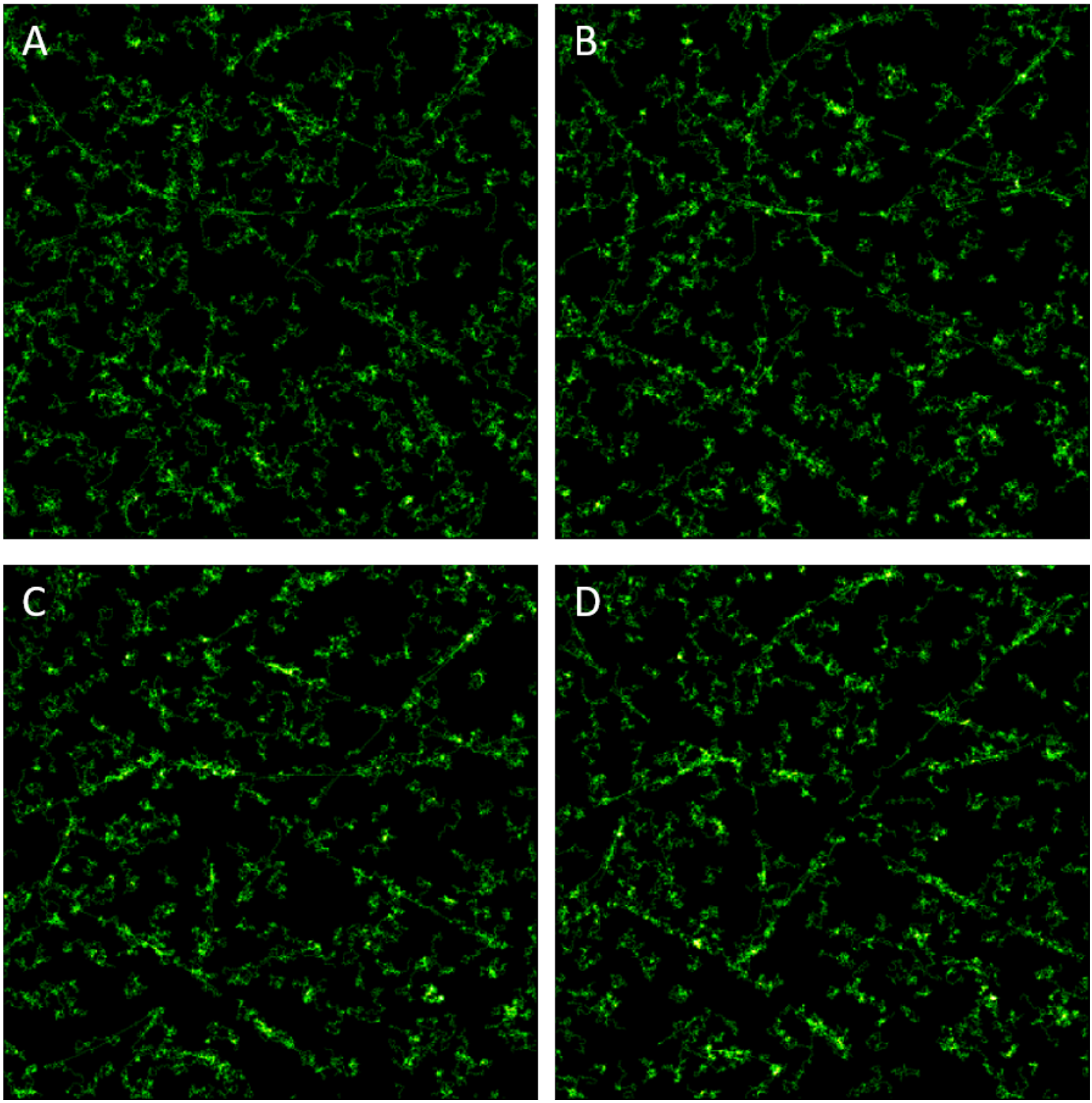



Figure 4.21: Heatmap (2.2.2) representations of movement density over time within set time phases of a model with broad and strong attraction cubic curves at 0-20 (A), 20-40 (B), 40-60 (C) and 60-80 (D) increments.

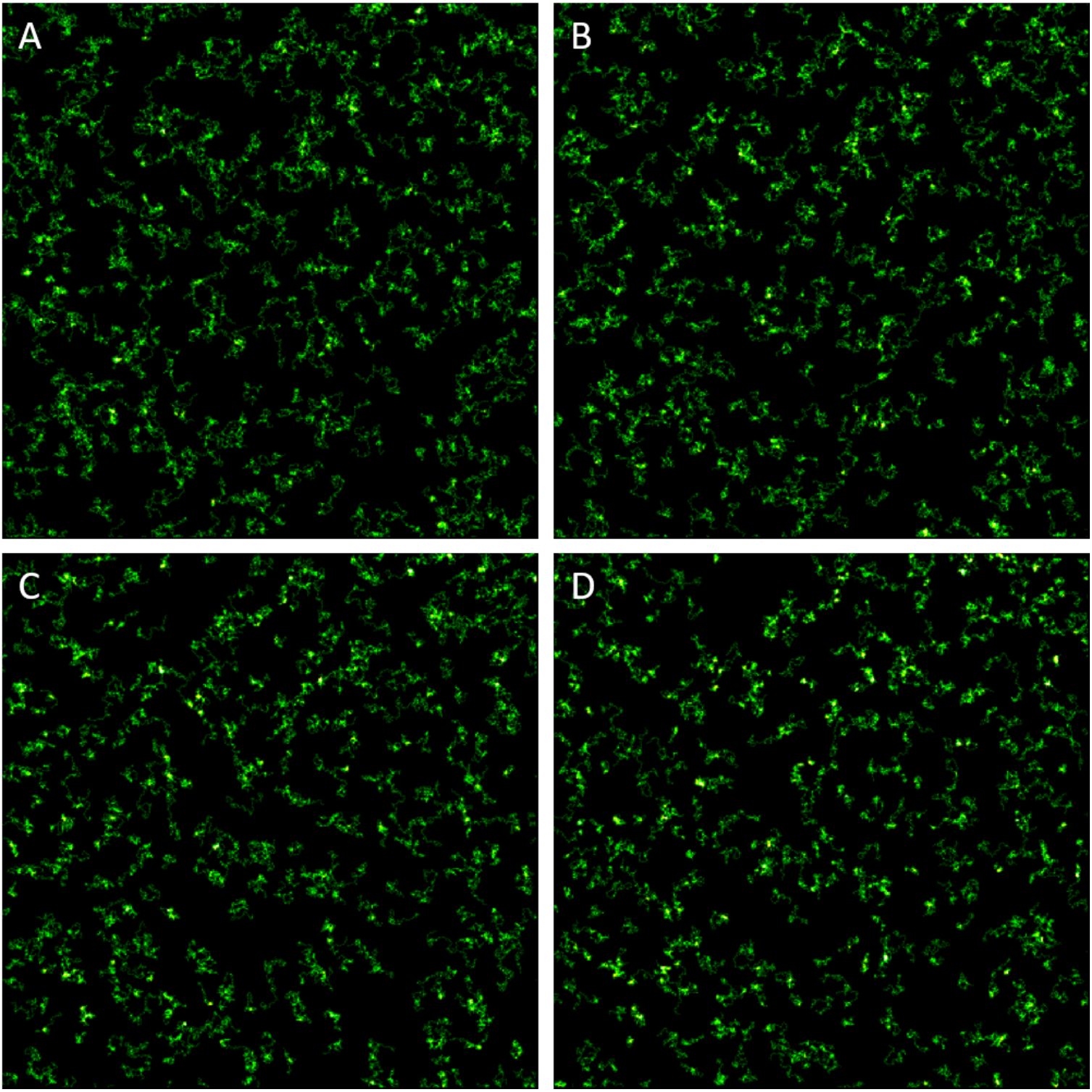



Figure 4.22: Movement heatmaps (2.2.2) within time segments of the overall model life-cycle Combining both lattice attractive paths and path forging following behaviour: 0-120 (A), 120-240 (B), 240-360 (C) and 360-480 (D)

**Least resistance and forging in representative models** Combining lattice paths and path forging behaviour replicates some recognizable patterns while requiring longer timescales to accommodate effective path forging, (Figure 4.22). Consistent with larger timescales, general movement is less sparse, it has not however become general undirected movement. Clusters of entities seem to exist around common strands, reinforcing localization and movement density in those areas. Weaker paths are still recognizable, suggesting that the path forging behaviour has a reinforcing effect, some signs of coalescence over time can also be seen. Across phases paths display stronger consistent pull and local starvation as entities both broaden them and reinforce attractive strength, in line with an exploitation narrative. However, clear defined paths are more difficult to identify.

#### Sub-populations

The larger cancer cell position set E was split into two separate comparable subsets by incrementing the threshold for direction 0 until the resultant sets each contained half the original trajectories. The subsequent sub-populations can then be passed through the framework as comparable data sets (Figure 4.23). The sub-population with randomly preferential turns, low motility and replication appear evenly spread across the simulation space but does not contain as many isolated strands. The second more forward biased sub-population presents most of the movement and exploratory behaviour in line with a narrative suggesting a more motility driven subset.

## 4.4 Discussion

### 4.4.1 GPCR and G proteins

**Micro-environmental turn preferences** Previously we found movement hot-zones (3) similar to those in GPCR and G-protein system literature [6, 55, 56, 57]. We also found a pattern of rear biased turns across the population and suggested that it may be related to hot-zone occurrence. Absolute directional heatmaps for GPCR and G protein set TC641 C1 show a preference for movement towards the centre of hot-zones. Relative maps suggest these turns are reflective from outward facing movement which shows as a rear turn preference(Figure 4.4).

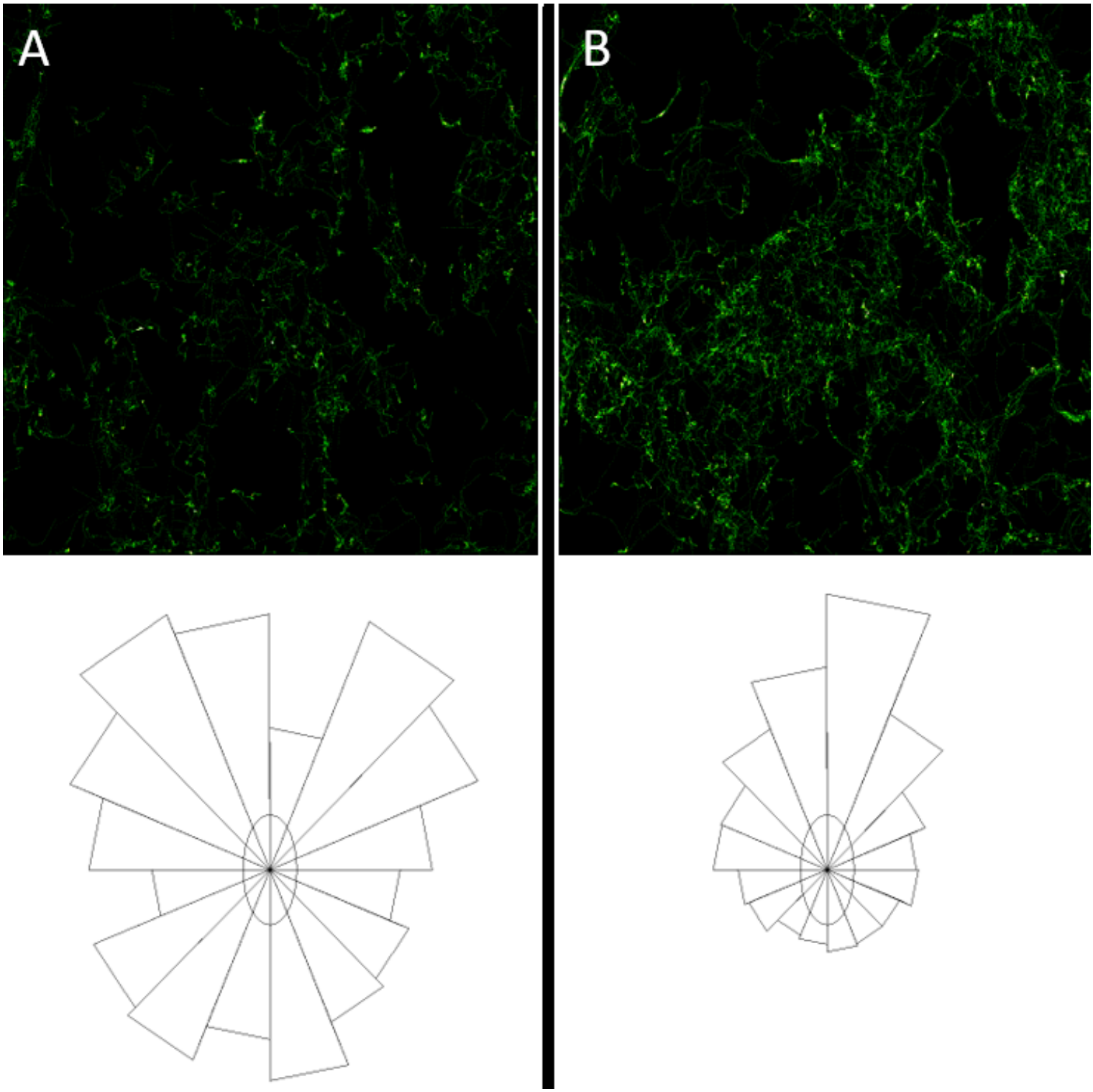



Figure 4.23: Sorted based upon the population preferred trend generates sub-populations with more general turn profiles (A) and directed trend preference (B) visual representations.

Models with a smaller immobile sub-population of 11% created small shivering areas analogous to non-colocalization hot-zones. Shivering immobile sub-populations show similar patterns to real in both absolute and relative directional maps suggesting the population size and movement behaviour is similar to real-world captured entities (Figure 4.4,4.6). Models with deflective boundaries representing cyctoskeletal compartmentalisation [6, 55, 56, 57] did not seem to be a strong single explanation for hot-zone movement. Here we were able to show that deflective boundaries can create some similarity to real-world results, however, only in extremely compressed spaces (Figure 4.6).

We also defined models with large low attraction areas possibly explaining localized movement starvation and the patterns of observed larger co-localizing hot-zones. Now, attractive zones show more undirected movement within, further replicating patterns within observed larger hot-zones (Figure 4.4). In contrast, relative diagrams show that motion in attractive representative models is extremely biased towards forward micro moves towards the centre (Figure 4.4). This differs strongly from real-world patterns which possibly have more bungee like interaction, at a distance from the centre entities are snapped back, a further iteration of attraction could be made. Movement starvation is a clear result of population restriction in all non-general models, supporting the real-world observed phenomena. While larger real hot-zones are clearly differentiated, the pattern of motion might indicate a slight coalescence around a central restrictive point. It is possible that the larger zones are forming or disassembling but some seem to comprise of layers like the attractive zone model. Directional heatmaps also confirm that the implementation of background general Brownian motion is not locality biased until effects are applied (Figures 4.4,4.6).

#### Short and long term system effects

Introducing time phases for heatmap generating metrics lead to several interesting new patterns being highlighted in both real and model data for comparison. The implementation also offered further support for already established suppositions with Pattern Oriented Modelling (POM) based validity. For example, presence of an immobile sub-population on the scale described would not be enough to create movement reduction across the simulation as seen in the real data (Figure 4.9). Similarly, attractive hot-zones alone are not sufficient, substantial portions of the population are being co-localized and constrained (Figure 4.11). Interestingly, many of the

smaller real-world hot-zones attributed to shivering exist within more movement starved areas than their larger counterparts (Figure 4.7). Movement starved areas may be indicative of unsuccessful low capture areas or a larger environmental phenomenon that makes movement more difficult, holding these members in place and excluding others. The starvation observation lines up with some of the C1 and C2 comparative channels in previous chapters concerning possible variant density and its effects; differing or even transient environmental densities possibly reducing movement in specific localities of the real system.

Mirroring real-world observations of larger hot-zones, larger attractive areas across time phases can be observed, capturing local movement, and intensifying over time (Figure 4.11). Again though, the amount of capture is not sufficient to lower general movement density as far as that observed in the real-world phases. So the real-world system still differs from the representative model (Figure 4.7). A model with simple attraction representation and general movement was also not enough to replicate the new slight strand like hotter areas. Consistent general movement without repeating pattern can be shown across all runs where population members are mobile. Therefore, it is likely a result of general travel distance within a hundred increments when added to stochastic Brownian movement within the model definition. Phasing also offers an explanation for the dissonance between deflective curve models and real-world *in vitro* results (Figure 4.10). By introducing time phases, we lose sight of even the strongest deflective curve representations within the model. Despite having a strong effect, similar deflective boundaries could therefore be present in the real-world system but requiring observation over much longer time scales than are available, thus supporting the real-world hypothesis of a restrictive zone interaction.

### Summary

Across the new tool applications some patterns remain clear, hot-zones and the attendant general localized movement starvation is consistent. We can also observe new localised patterns of movement for pattern comparison, possibly due to transient confinement conditions [9] or unknown interactions. Some larger hot-zones remained consistent among time phases but with differing morphology, again possible indicators of a environmental transience not present in our current models. However, we also observed many small hot-zones form and loose cohesion, this supports some of the Calebiro groups recent work identifying and describing transient trapping in single entity trajectories [152]. It may be that our confinement model requires transient catchment

areas, combined with an attractive retention effect such as creation of dimers reducing travel and improving representation. As such the models could be improved with time based deflective curve strengths and shapes along with interacting sub populations with changing behaviour.

We also filtered results on quantifiable data such as turn amount and distance travelled observing whether hot-zones could be easily separated from general Brownian diffusion patterns (Figure 4.13). However, simple metrics such as these were not sufficient and allowed edge cases to be captured.For example, since turns are a raw metric and not the proportional dominance of turn direction, there is space for fragmented tracks to register incorrectly. Because results can be skewed by many factors, a further area for improvement would be the inclusion of metrics such as maximum square displacement or image analysis techniques such as neural networks [152].

### 4.4.2 Cancer cells

**Micro-environmental movement patterns** We previously observed that strands were probably generated by least resistance gradient following. Now, analysis via directional heatmaps allows us to observe that forward motion dominance is pervasive throughout but more pronounced in strands (Figure 4.14). As suggested, entities are clearly following a path. In general movement areas there is still some forward preference. However, there are also small cyclic sub patterns often observed in model results and other sets for low motility or generally immobile population members. Therefore, areas of general movement patterns may be population centres with no strong directional drive but indicative of another process such as cell splitting.

Path cohesion and congregation seems to increase over time, similarity was observed at different time points across all the generated model categories, path forging and lattice (Figures 4.16,4.17). A combination of several interactions such as path following and forging may be present in the observed cancer micro-environment. Therefore, one possible explanation is that of coalescence over time into strand-like groups of cells with greater adjacency but lower generalized exploration. Much like other self-organizing systems, initial exploration of an environment is then exploited with convergence on common paths [90, 92].

Differentiation between general movement areas can also be observed, multiple small strands making up a larger area of dense undirected movement (Figures 4.14,4.15). Strands could be a large common area of individual movement that creates undirected areas as movement density increases and medium structure degrades. In similar modelled circumstances, possibly significant separation seems to be occurring (Figures 4.16,4.17). Regarding general movement, the combination in models of both forging and lattice paths seems to create clear strands. However, we observe general movement without the level of localized starvation that either forging or paths cause. Indeed, general similarity seems to be greatest across both absolute and relative heatmaps when comparing cancer data sets A-E to the lattice following and forging hybrid model.

### Short and long term effectors

Short length cancer sets A-D and most model phases are only 20 increments long, this can create greater consistency over time than comparison to the longer 45 increment set E (Figures 4.18, 4.19). The general movement phase heatmaps are invaluable for assigning proper weight to serendipitous coalescence and emergence of patterns. Path forging is very difficult to implement in isolation, with longer jump phases entities tend to look at a preferred area then jump through, and past, rather than following the direction of least resistance or attractive path (Figure 4.21, 4.22). We can improve future models with a more nuanced path forging implementation.

Again, the narrative of built-in lattice paths of reduced resistance being exploited by path-forging mobile cancer cells seems to be prominent. The observed coalescence in real-world system and path forging lattice model results supports the sequence of events. Observation of real-world phases also supports the suggestion that the longer time frame set E is taken from an earlier population time point than the other cancer cell sets. Interestingly, hot-zones can also be seen in the real-world phases starving general movement, bright zones suggest a substantive sub-population of immobile entities.

**Summary**

Directional heatmaps allowed analysis of more sub-environmental patterns and cross comparison of strand shape and strength. A possible explanation found in literature is that different confined environments tend to lead to different cancer cell behaviours [140]. The presence of general non directional movement around areas of strands indicate either a change of cell type or environmental constrictive force. It is not clear whether cell movement or degradation of the environment could be a direct cause. Similarly, application of phased heatmap representation further highlighted possible strand coalescence over time. A limitation of the current model definition is the lack of a local density representation, without one it is difficult to define and observe the effects of localised permissiveness beyond our lattice implementation.

Sifting has difficulty separating entities effectively upon only quantitative metric selection (Figure 4.23). We may have identified a proliferating low motility population and a high motility gradient following subset, consistent with population heterogeneity and possibly morphological differences [142]. However, resultant movement pattern distinctiveness is still unclear and separation via existing metrics very difficult. Again, analysis would benefit greatly from improvements in track classification and population separation.

### 4.4.3 Conclusion

By adding tools for more specific primarily directional and temporal analysis, we have improved the framework and enabled further important observations. In the cancer data set, we have supported observations suggesting that some show the entire process from random movement to coalescence around constructed paths of least resistance. Cells are also clearly following directional trends even within less well-defined strand-like movement areas. They may be moving along an in-built lattice of least resistance, but also likely forge those paths over time. Within the GPCR and G protein sets, directional heatmaps have highlighted a strong correlation between rear bias and restrictive or attractive hot-zones. Also, where previously model support for a restrictive zone relationship was difficult, time phase observations suggest an explanation; boundaries may be extremely transitory. Further improvement should highlight further patterns of causal interaction.

We still continue to encounter issues with numerate comparison and selection of visual pat-

terns. We can generally identify visual patterns and assume similarity by the occurrence, but it is very inexact. We have also identified issues with population filtering when we wish to separate clear visual patterns, validation of representative models and in turn automation of model definition should be improved by the inclusion of such methods.

Therefore, we will next develop comparison and classification based upon complex visual data using artificial intelligence methodologies. An additional sub-population filtering tool based upon visual metrics will also be developed.

# 5 Artificial neural nets for movement pattern classification

## 5.1 Introduction

Both model to *in vitro* movement pattern similarity comparison and separation of patterns based upon individual behaviour have been identified as areas for potential improvement. Therefore, we aim to develop a novel workflow to classify visual patterns for model/system comparison and filtering populations into subsets of track patterns.

Our framework enables the analysis and visualisation of population member movement over time in cancer cell and G-protein-coupled receptor (GPCR) biological systems. This is achieved by creating representative visual patterns and in turn developing insight into environmental interactions. However, important mechanisms have proven difficult to characterise, e.g. in the GPCR and G protein system, the differentiation between GPCR hot-zones, co-localization, general movement and static population members. Similarly, cancer systems have displayed complex relationships between random movement and following sub behaviours, e.g. a combination of lattice following and path forging representative models generating results representative of real-world sets. Therefore, we need to enable more targeted analysis through automated differentiation of behavioural types. To break down a biological system into multiple sub systems for comparative modelling, we first need to be able to identify such behavioural sub sets and analyse them separately.

An ongoing difficulty with comparison between representative models and real-world results is validation or definition of a similarity metric; defining how closely a model mirrors real data observation. Thus far, qualitative visual comparison is difficult to validate (Chapters 2,3,4). Furthermore, comparison with quantitative metrics such as travel distance, population size or

turn preference over time has not been directly representative of emergent movement patterns (Chapter 4). Therefore, there is a need to develop a more robust approach for visual pattern based behavioural differentiation and model comparison. Artificial Intelligence (AI) seems to be ideally suited for such a task.

There are many pre-existing approaches to track analysis and classification, but we require an approach that needs minimum manual feature definition and can be generalised across biological systems. In relation to GPCR tracks the Calebiro group recently discusses [152] approaches such as back propagation neural networks, random forests, standardised maximum distance and propose their own recurrence matrix approach. One study discusses the advantages of a graph neural network, surmising that they need a larger amount of input data than classical methods, but perform well and can be reliably trained from simulated data [153], with the framework we have an abundance of simulated training data. Convolutional neural network (CNN) approaches have also been shown to outperform gradient boosted and random forests in trajectory classification [154], albeit only slightly. However it was again noted that features did not need to be defined but learned from the dataset. Therefore we chose neural net approaches which are generalisable across trajectory classification and as applicable to similarity quantification. We should also be careful since the approach is more sensitive to inclusion of patterns from outside its training set and performed poorly there [154].

Neural nets, deep machine learning techniques, present a promising area of growth and can be applied to image interpretation. Google's Tensor-flow platform in particular has been on the rise and claims to be a powerful experimentation tool for research with widespread support [155, 63]). We aimed to conceptualise and apply a general proof of concept approach to explore neural net applicability; defining similarity between visual patterns and segmentation of populations. We present here a process that shows promise for further exploration; a method for quantifying the real-world to representative model comparison and a separate but similar approach to identify subsets by track visual profile.

### 5.1.1 Summary

Trained neural nets are used to generate similarity metrics for GPCR and cancer representative model to *in vitro* results. We found most similarity between attractive area representative models

with GPCR and hybrid forging/lattice path models with cancer results. Both sets are also filtered to create subsets of entities based upon visual patterns: GPCR and G protein split into compressed (GPCRShiver), Brownian (GPCRBrown) and fragment (GPCRFrag) sets, cancer was separated into complex (CancerTurning), simple (CancerDirect) and fractional(CancerFrag) sets. Filtering net transference was also investigated with cross application of neural nets to the set they were not directly designed for; some interesting insights but both types of nets are more applicable to their designed set.

### 5.1.2 Artificial neural nets (ANN)

As a field experiencing considerable growth, the application of artificial neural nets (ANN) have been extensively discussed for decades [114, 115]). Unfortunately, general conceptualisation is still difficult but broadly analogous to interconnected biological neural nets, each neuron processes signals and sends them to their connected neurons. While individual neurons are unaware of what an image is, specific regions of the brain correlate to different visual recognition patterns. In a computational implementation, neurons are nodes made up of input and output relationships, each node taking an input, transforming it and passing it on.

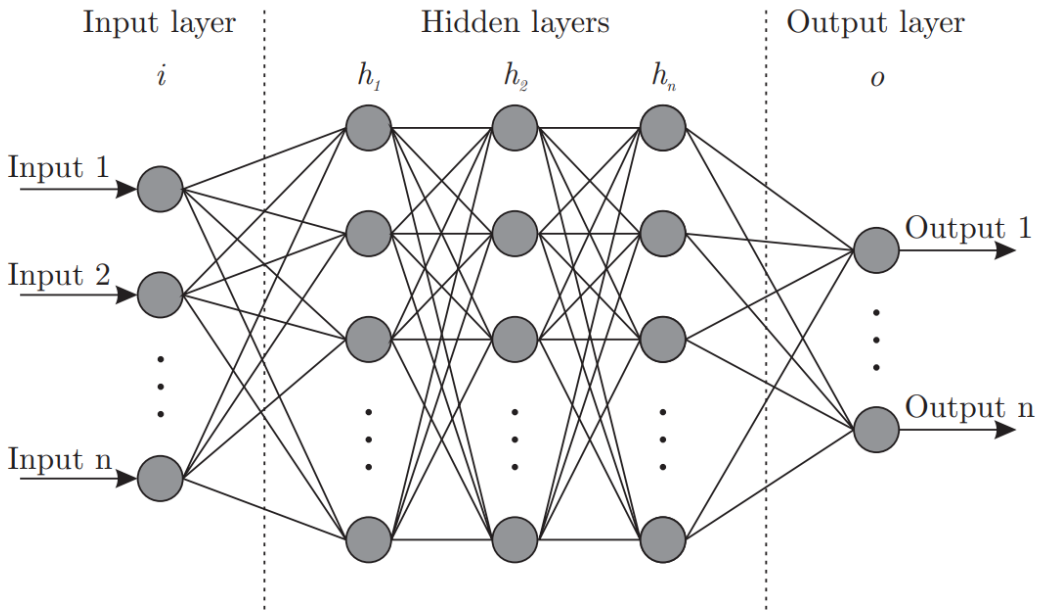



Figure 5.1: The structure of an Artificial Neural Network (ANN), input, intermediary hidden layers and output taken from Bre et al [156]

The transformation of an input is based upon the neuron's internal workings, most nodes are

connected with differing weights. Network nodes are often arranged into hierarchical hidden layers between initial input and final output sets (Figure 5.1). The aim of $n$ intermediary hidden layers is to modify the input values before they reach the output layer so that a requested result is produced. While layer and overall network design, such as neuron inter-connectivity or layer number, is often predetermined and static the weights and function of a network are trained and dynamic, needing to be recorded to repeat results. Training data is given and passed through the network; in our case the data is classified, and the success rate is used to evaluate distance to desired result. Neurons in the network are modified based upon score accordingly, over time and iteration this process can result in a network that accurately and repeatedly separates data types by input class.

Layers can have targeted and expected functions within a network but exactly what is happening at each point becomes obscured. Sometimes conceptualised as a dense decision tree made of black boxes with floating highly variant weights, it is unwieldy at times. It becomes easiest to evaluate nets by output and results. Often within tutorials, the reasoning for values is given as being "*one that works*", the causal relation often seeming intractable. Classification, prediction, and clustering of data are all common applications. We used *TensorFlow* to implement ANN's, a flexible platform that could be integrated into this or a similar framework approach.

### 5.1.3 TensorFlow

The *TensorFlow* platform was developed by Google Inc. and benefits from important community support and library development [155, 63]). Designed to be easy to use, its pipelines have been applied to a wide range of established case studies including MRI imaging and text classification systems. Like Agent Based Models (ABM) and many other natural computation paradigms the mantra of serendipitous emergence can lead to exciting new emergent applications or extended periods of obtuse inactivity. *TensorFlow*, written in C++ and usually interfaced with via Python, differs from Java, the Framework code base, but can be integrated at a code or results level.

### 5.1.4 Convolutional networks

Convolutional neural networks (CNN) are ideal for our visual application as they were designed for visual learning and image classification via feature identification and environmental correlation. Areas of the network attempt to address different sections of the given grid; designed after our understanding of spatial recognition in the visual cortex, breaking images down into sub sections and evaluating them for important features. After training, sections with strong identification correlation might have a higher weight than surrounding areas. All three of the pre-designed neural net models we applied utilised variations of the convolutional approach [157, 158, 95]). As with most algorithms, the developers are attempting to provide best results in reasonable compute time with limited data. As usual for such specialised approaches, they perform variously depending upon the actual input data. The calibration and training of models becomes a case of pre-developed model selection, followed by parameter definition and attempted training within the limits of available compute resources.

## 5.2 Methodology

Training neural nets to recognise features of a model or real-world *in vitro* sets over an entire run can be difficult and sometimes unreliable. Compounding this, the required size of training and validation datasets can be generated by the framework but without the magnitude of real-world sets required to effectively evaluate results. Therefore, we should consider the similarity results a proof of concept within a framework like pipeline. However, for single track classification we do have sufficient real-world examples to generate more stable models.

### 5.2.1 CNN pipelines

One of the key features in our use of the *TensorFlow* platform was the ability to apply pre-designed and weighted nets from previous research. Many patterns common to images such as curve and edge detection can be transferred; a net applied with training on only surface areas, this also improves model stability. As such we used several existing convolutional networks: *VGG19* [157], *InceptionV3* [158] and *MobileNetV3* [95] where they seemed most applicable, *VGG19* performed best for bi-modal classification and *MobileNetV3* for multi-modal.

The stochastic variance throughout definition and training does however make model calibration very difficult. So, we cannot claim best use of each model, only functional use. Therefore, if attempting to reproduce results we recommend applying several of the pre-trained networks and evaluating performance per case.

When referring to classification in the context of neural nets, *TensorFlow* and our available datasets, we refer to categorization of data into known groups. We attempt to train a neural net that can recognise and differentiate between several classes of relevant image by providing known reference labelled examples. Once trained the recognition net can be applied to our real-world *in vitro* systems, to classify them appropriately. Separated sets can mean separation by track which can be used for further analysis in our single track approach. Alternatively, we can separate movement heatmaps as a metric itself; the classification confidence of separation suggests similarity between class training set and applied real-world example, our model similarity quantification.

So, in the model similarity approach we use classification confidence to indicate similarity between real-world images and those generated by our representative models. Similarity between both indicating the strength of representative model definitions.

The single track approach is used for classification and separation of real-world sets; to produce more detailed and targeted pattern representations. For example, in the cancer cell set we attempt to separate path-following and static entities to observe them directly, in GPCR we wish to directly observe hot-zone creation.

Both model similarity and single track classification approaches use the same general pipeline (Figure 5.2). We can create a wide range of images similar to real-world results by leveraging the generalised data generation ability of the framework. Training a neural net ordinarily requires two particularly difficult input requirement: a large quantity of available data and accurate labelling, i.e. the classes of generated data must be known. After identifying the required differentiating sets, a framework model factory run can easily generate large quantities of similar stochastic varied images with known *classes* and apply *labels* for input. Our pipeline splits the

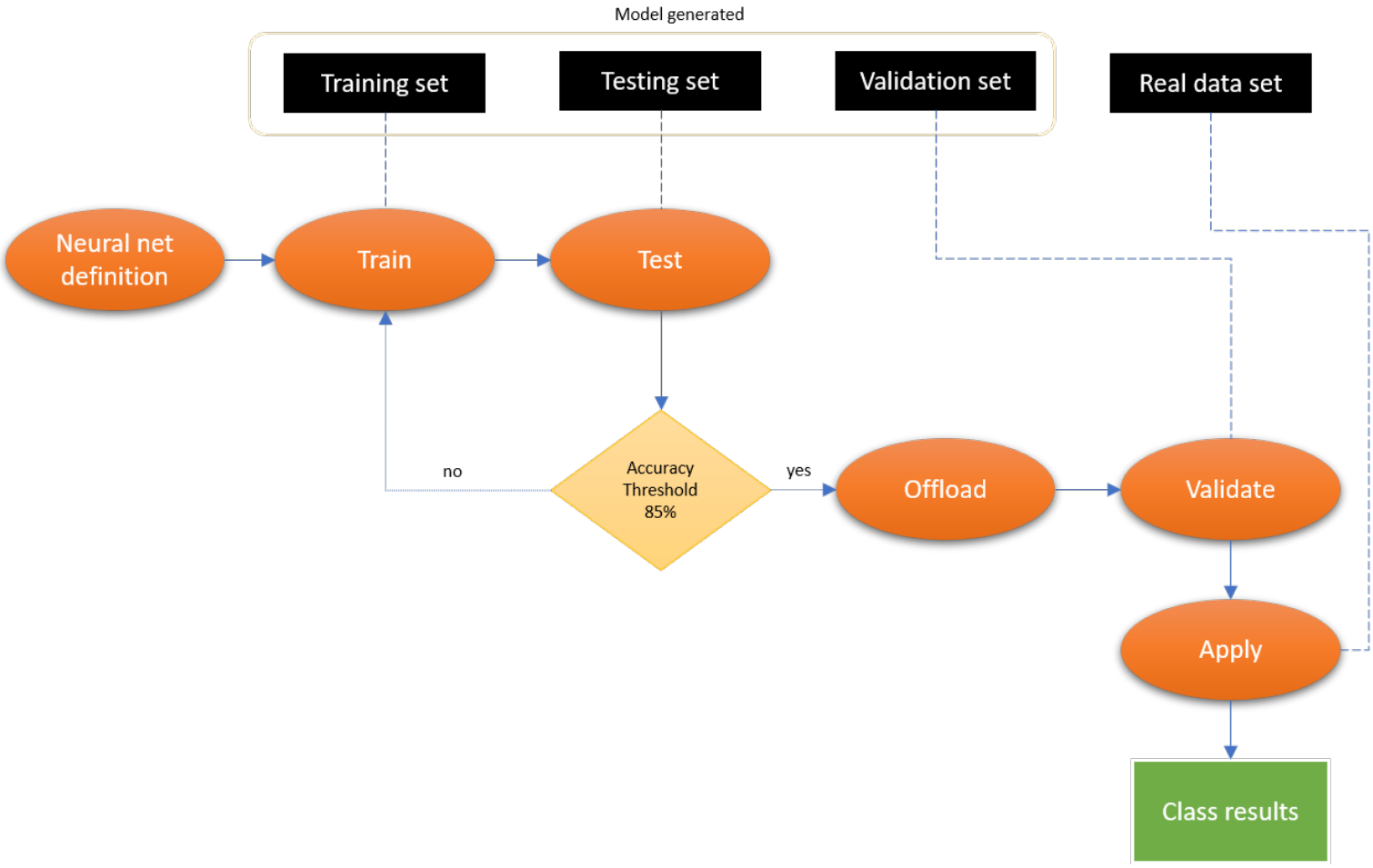



Figure 5.2: Once a neural net is defined, model generated data is used to train across classes of image. Once a net is trained a separate test dataset is used to reduce over and under-fitting as well as ensure applicability. If the accuracy is below a threshold, then a new net is trained with different settings, otherwise the trained net is passed off to be validated with a third generated dataset. Validated nets are provided with real-world *in vitro* datasets and used to differentiate and label classes of data images.

labelled data into training, test and validation sets; neural nets are trained, tested and validated before application to the real-world datasets. If a real-world dataset were large enough, manual classification of a subset of images followed by separation as a validation set would also improve the applicability for more accurate results.

### Model similarity

Nets can be trained to identify and differentiate end of run movement heatmaps into classes. While not infallible, the results produced worked in clear cases and often gave correct results in our test examples. Where training nets to differentiate was difficult, we switched to a grid comparison approach; rather than compare and classify all types of model results, nets made a bi-modal assessment between one class or the other. From this two way model similarity comparison we then assemble a grid of comparisons to generate a similarity confidence metric. The nets performed better when narrowing the field of classes from several to a pair. However, the model similarity single pair approach takes longer to develop and apply since every model type needs training with each other type, greater accuracy at the cost of training and development time. A trained neural net can be applied to real-world data for its classification confidence relative to model trained understanding.

### Single trajectory filtering

Breaking movement down into its simplest representation loses some geographical and emergent pattern information but allows us to turn a single GPCR movement heatmap into 4000 constituent tracks. The automated tools of our framework also make post classification reconstruction reasonable. Therefore, treating datasets individually allows us to filter and analyse sub-populations, break a heatmap into constituent patterns and increase the number of differentiated recognisable features. Resultant separated track sets can also be run through other existing analysis tools. We can then compare movement feature separated sub populations post classification with any other data source within the framework.

We are able to apply externally pre-built CNNs such as *VGG19* [157], *InceptionV3* [158] and *MobileNetV3* because of transference; the concept that relevant pattern recognition can be

retained and applied to different problem areas, in our case type of systems. For example, for our single trajectory filtering we can transfer the cancer trained CNN to GPCR and G protein data, then vice versa. This allows us to assess trained net applicability to different problem spaces and the similarities between resultant population separation.

### 5.2.2 Transference

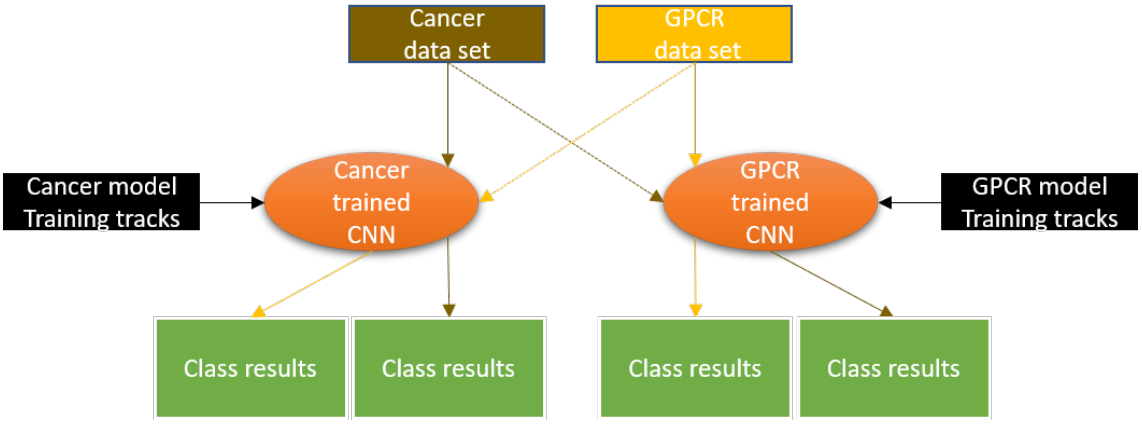



Figure 5.3: After generation of our trained cancer and GPCR separation nets, we can cross apply them to sets from systems they were not originally designed for and leverage transferable properties

While we have designed different training sets for cancer and GPCR and G protein track separation it is also possible to cross apply trained nets to observe differentiation (Figure 5.3). Since much of the time and resource cost is in training and the framework automates analysis, opportunity cost to generate crossover transferable results is very low; any useful new results would be worthwhile for explorative analysis. The GPCR and G protein net is trained for separation on track compression and the cancer set identifies based upon track complexity. We can compare both to explore different possible hypothesised behaviours or to develop better understanding of differentiation in our trained nets.

### 5.2.3 Proof-of-concept

The framework facilitated break down of larger homogeneous sets into individual tracks followed by later reassembly and comparison of produced sub populations. Also, for training, validation, and proof of concept purposes the ability to generate large quantities of model derived simulated data via the framework was very useful.

### Model similarity

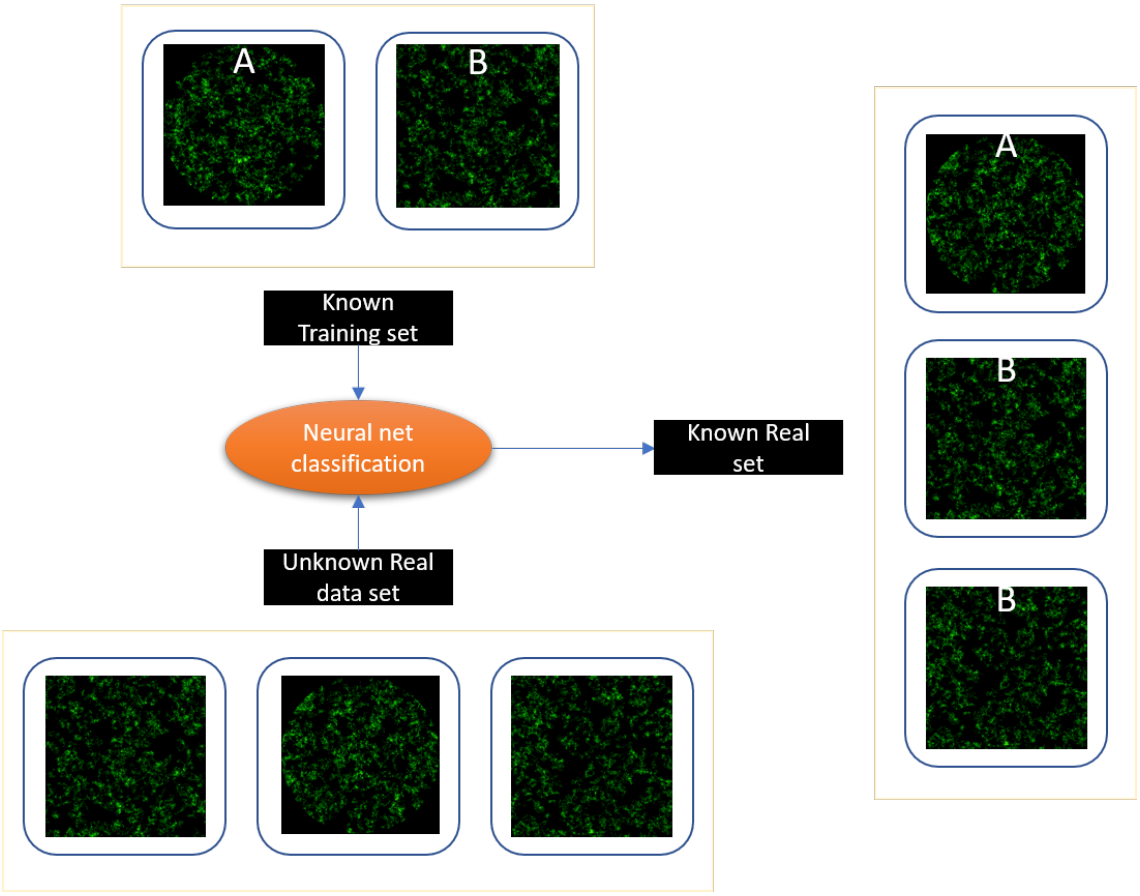



Figure 5.4: Model generated movement heatmaps, training data for bound and random movement is inputted into our classification pipeline (Figure 5.2 along with unknown images. Our trained neural net then classifies unknown heatmaps based upon overall profile.

We were able to input entire run heatmaps for classification by utilising the training, testing and validation based general pipeline we established (Figure 5.2). Our initial proof of concept set used 800 bound and 800 random training cases with 100 non overlapping validation and test sets generated separately for each. The aim was to use a simple, clear, visually differentiated set to test whether such classification could work in a best-case scenario. This is one of many cases there image based classification is necessary, numerically both sets would show identical

| Proof 93% | Bound1 | Bound2 | Bound3 | Random1 | Random2 | Random3 |
|---|---|---|---|---|---|---|
| Bound | 0.76 | 0.63 | 0.82 | 0.08 | 0.09 | 0.03 |
| Random | 0.24 | 0.37 | 0.18 | 0.92 | 0.91 | 0.97 |

Table 5.1: Trained net confidence values for bound and random examples used in the proof of concept model similarity comparison network comparing bound and random movement heatmaps

quantitative measures such as active entities, turn trends and movement over time. An artificial boundary while easy to spot visually is more difficult with less complex metric comparison.

Once defined and trained our proof-of-concept neural net for random and bound movement heatmaps was presented with three newly generated examples of both bound and random sets (Figure 5.1). The larger validation set returned an accuracy value of 93%, where the net was able to successfully identify all the unknown examples but with greater surety for random sets than bound. From the variance of proof values, binary classification could be achieved successfully via this approach.

#### Single trajectory filtering

As a proof of concept for feature separation and demonstration with easily differentiated movement patterns we elected to generate a model with two distinct populations (Figure 5.5). One half of the population moved randomly for the duration of the model, the second half shivered back and forth. Both population behaviours resulted in distinct movement profiles that could be trained against. By classifying and separating the starting model trajectories with a neural net most hot-zones were extracted from general movement patterns. Of the 2000 tracks 30 were mis-classified (1.5%), a reasonable loss rate when attempting to analyse population wide pattern generation but also important. In this case mislabelled population members are recognisable in each of the separated sets in small quantities, in less controlled environments we should recognise the possibility and be conservative with the interpretation.

We visually identify several clear target cases within the real-world set and generate classes of training tracks with the framework. Once training tracks are available we employ our classifica-

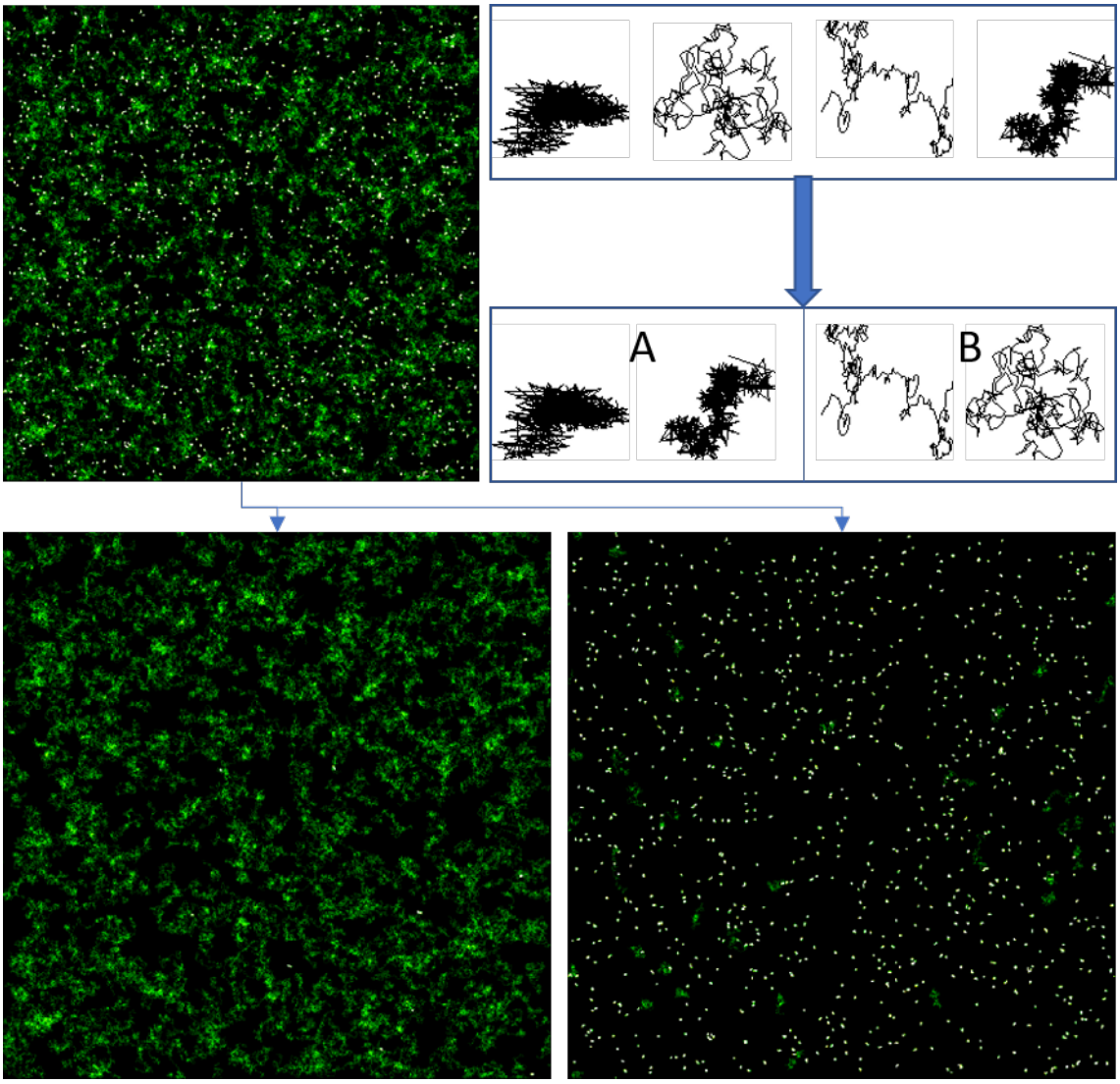



Figure 5.5: Movement heatmaps (2.2.2) for a model comprising of shivering and random movement sub populations. Prior to filtering based on distinct trajectory movement profiles (Figure 5.6 and then the movement heatmaps for resultant separated sub sets.

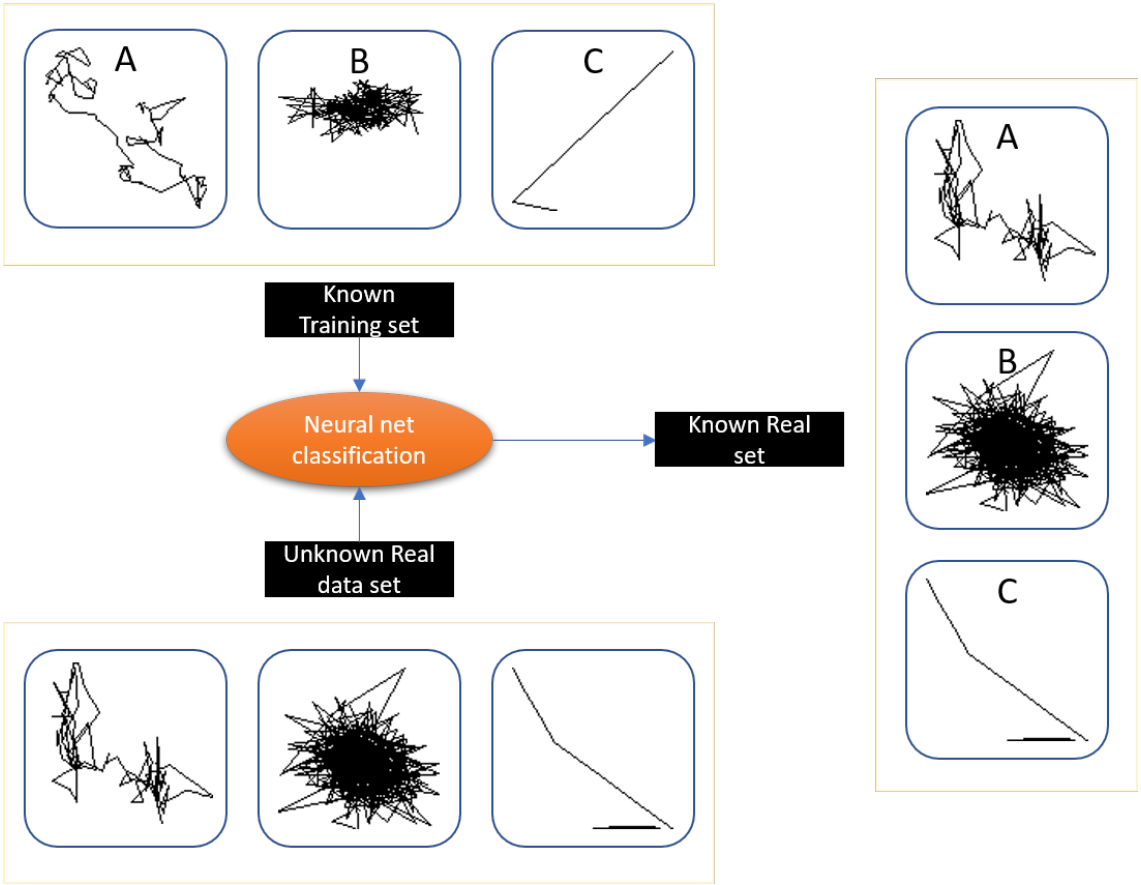



Figure 5.6: Model generated known training track data is inputted into our classification training pipeline (Figure 5.2 along with real-world broken down unknown tracks. Our trained neural net then classifies real tracks based upon movement profile.

tion pipeline to generate an trained neural net and input our real-world dataset for separation (Figure 5.2). In general, we used training sets of 1000 members per class, the amount of available training data was much larger but viable results were generated with a shorter training cycle. After separation, we give the classification list to the framework to split our real-world dataset into sub population sets for visualisation and further analysis.

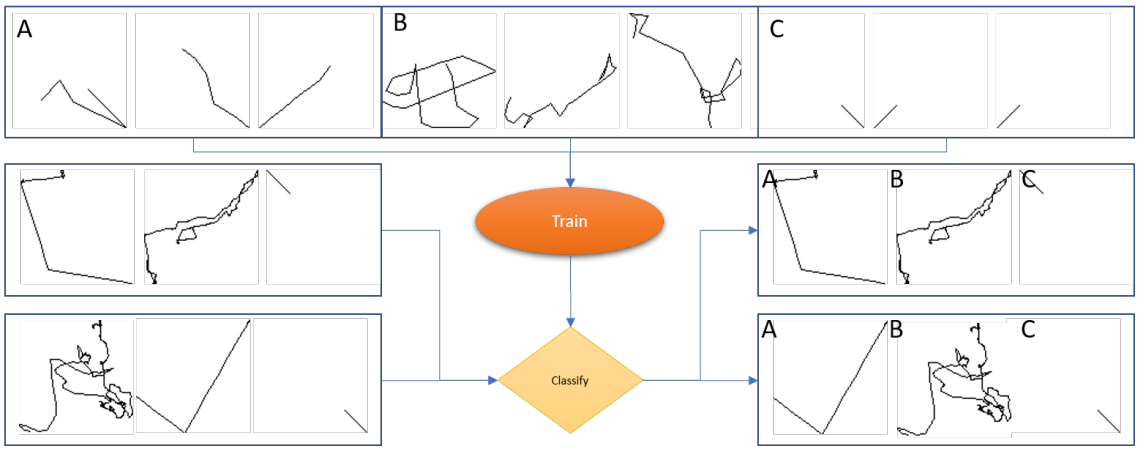



Figure 5.7: Classifications of CancerDirect (A), CancerTurning (B) and CancerFrag (C) generated tracks are fed to a training pipeline (Figure 5.2, the resultant trained network is given real-world input from large and small cancer sets for classification into three initial classes.

**Cancer classes of trajectory** As most cells behave individually in similar manner, track differentiation can be difficult. We therefore chose to differentiate, and train based upon track and therefore behavioural complexity (Figure 5.7). *CancerDirect*(A), *CancerTurning* (B) and *CancerFrag* (C) track classes were defined and generated by the framework. After training and separation, classification between CancerTurning and CancerDirect tracks seems to be primarily differentiated on length of near straight segments or general stretch of tracks. A circular track would be classified as CancerTurning but a curve with compact overall profile despite many micro turns is recognised as CancerDirect. CancerFrag tracks represent an attempt to identify tracking disconnects and resultant shortened movement.

### GPCR and G protein classes of trajectory

Description of the GPCR and G protein dataset thus far has focused heavily upon hot-zones. We expect that differentiation from normal Brownian motion and individual based hot-zone

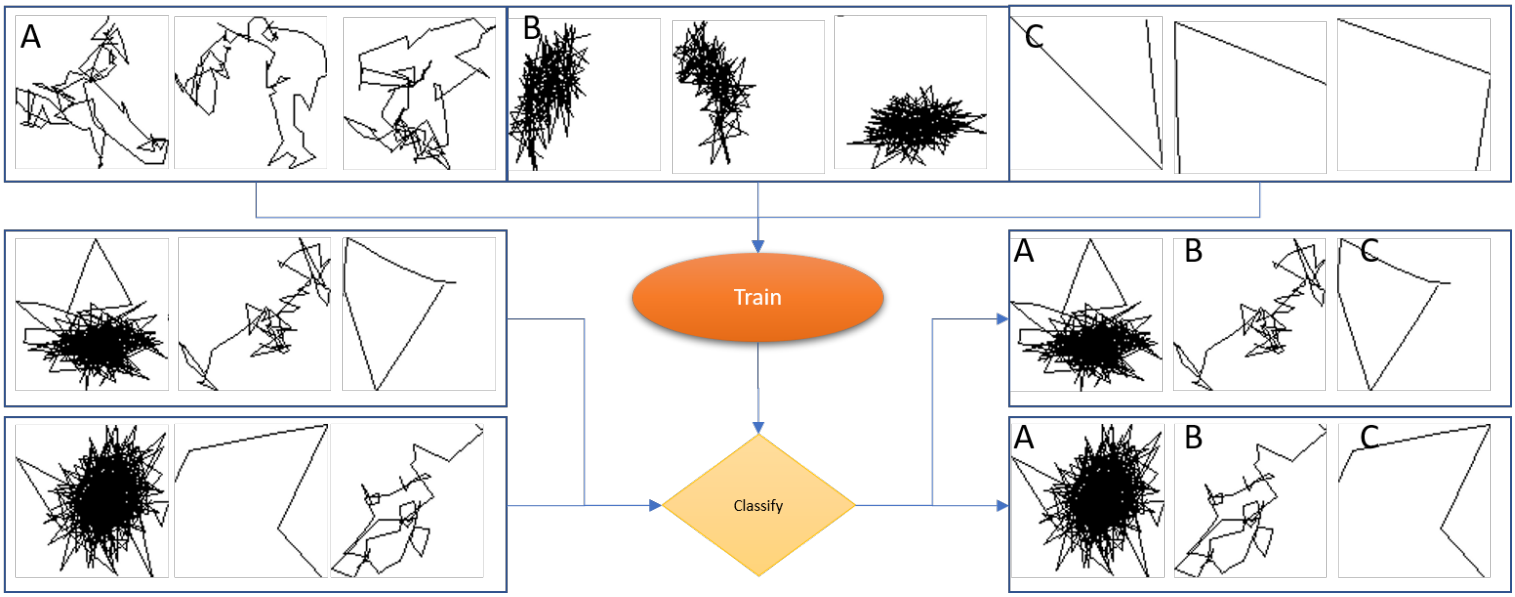



Figure 5.8: Training trajectories based upon general GPCRBrown motion (A), GPCRShiver shivering (B) and short term GPCRFrag movement (C) are used to generate a GPCR and G protein specific classification net to turn unlabeled C1 and C2 sets into groups of known track types.

generating behaviours should improve our ability to target observation and analysis. Manual observation of individual trajectory images once separated from the dataset showed several clear categories. Entity behaviour seemed separated into general movement *GPCRBrown*, highly compressed *GPCRShiver* back and forth movement and fractions where tracking was lost *GPCRFrag*. Training sets for all three behaviours were generated and used to develop a neural net for filtering (Figure 5.8). We aimed to differentiate background and hot-zone movement and possibly separate hot-zone and single entity shivering phenomena.

## 5.3 Results

### 5.3.1 Cancer

Comparison of model and real-world *in vitro* images has provided some insight and highlighted questions regarding striated movement paths in real-world systems: e.g. whether a following and forging behaviour is creating paths over time or if paths of least resistance were part of the environment. We previously implemented models representing lattice attractive paths, path-forging and the combination of both (Chapter 2 2.2.2) with some success. While we registered qualitative visual similarities between patterns it was difficult to define which was most similar

| Cancer set validation 89% | PF | PFL | BG | P | R |
|---|---|---|---|---|---|
| short A | 0.07 | 0.54 | 0.16 | 0.18 | 0.04 |
| short B | 0.17 | 0.69 | 0.06 | 0.07 | 0.01 |
| short C | 0.21 | 0.55 | 0.09 | 0.13 | 0.01 |
| short D | 0.17 | 0.52 | 0.13 | 0.16 | 0.02 |
| long E | 0.25 | 0.09 | 0.21 | 0.40 | 0.06 |

Table 5.2: Confidence classification for full movement heatmap (2.2.2) comparison of cancer set heatmaps with the model generated results discussed in chapter 2 and 4, values represent the classification allocation given by our trained neural net pipeline as a portion of 1 or 100%, higher values represent higher similarity between movement patterns within (Path-forging (PF), Path-forging and lattice paths(PFL), Background(BG), Paths(P), Random(R) )

| Cancer short | PF | PFL | BG | P | R |
|---|---|---|---|---|---|
| mean | 0.21 | 0.60 | 0.16 | 0.24 | 0.04 |
| max | 0.25 | 0.69 | 0.21 | 0.40 | 0.06 |
| min | 0.07 | 0.09 | 0.06 | 0.07 | 0.01 |
| range | 0.18 | 0.60 | 0.15 | 0.33 | 0.05 |
| median | 0.17 | 0.54 | 0.13 | 0.16 | 0.02 |

Table 5.3: Averages and range for confidence classification of full movement heatmap (2.2.2) comparison for short A-D cancer set heatmaps with the model generated results (Table 5.2) (Path-forging (PF), Path-forging and lattice paths(PFL), Background(BG), Paths(P), Random(R) )

without application of CNN's.

#### Comparing real-world and representative model movement patterns

For the model similarity test, we fed training data for path-forging, lattice paths with forging, background, lattice paths and completely random movement models over time to the neural net. Training of a neural net to recognise differences between the cancer related models was possible as a single large classification net with a reasonable training validation success rate of 89%. We then gave the trained net our real-world data heatmaps and quantified the classification choices (Table 5.2,5.3).

The entirely random set presented all the lowest confidence values with the real-world movement heatmaps 0.01-0.06 (Table 5.2,5.3), confirming that movement within the real-world sets is not purely random and following the literature [124, 92, 80]. Across runs our trained net supports the assertion that the models are more representative than undirected purely stochastic model generation.

The short and long sets generate very visually different heatmaps, as such the CNN produces quite different classification associations. Short sets associate most closely with path-forging and lattice path models 0.51-0.69. While the longer set correlates well with lattice paths, 0.4, it has a much lower value than the short set association with combined models (Table 5.2,5.3).

For the short runs path-forging (0.07-0.21), lattice path (0.06-0.18) and background general movement (0.06-0.16) models all have similar association levels within reasonable ranges, they are difficult to differentiate. For the long set, while path-forging 0.25 is closer than background 0.2, combined path-forging and lattice path models 0.09 are low enough that it almost approaches random movement 0.06, differentiation being again, difficult (Table 5.2,5.3).

#### Trajectory complexity classification and population subsets

In the case of both short and long cancer sets, we expect the clear strand features are created by multiple entities working in concert over time. Observational classification and segregation into sub-populations via quantifiable metrics has therefore been difficult, our target pattern is made up of many individual interactions. Applying single trajectory classification might identify the constituent behaviour types, for example, cells that build paths versus those that exploit them. A trained net was generated and applied to both short and long time-frame cancer track sets.

#### Short real-world data set separation

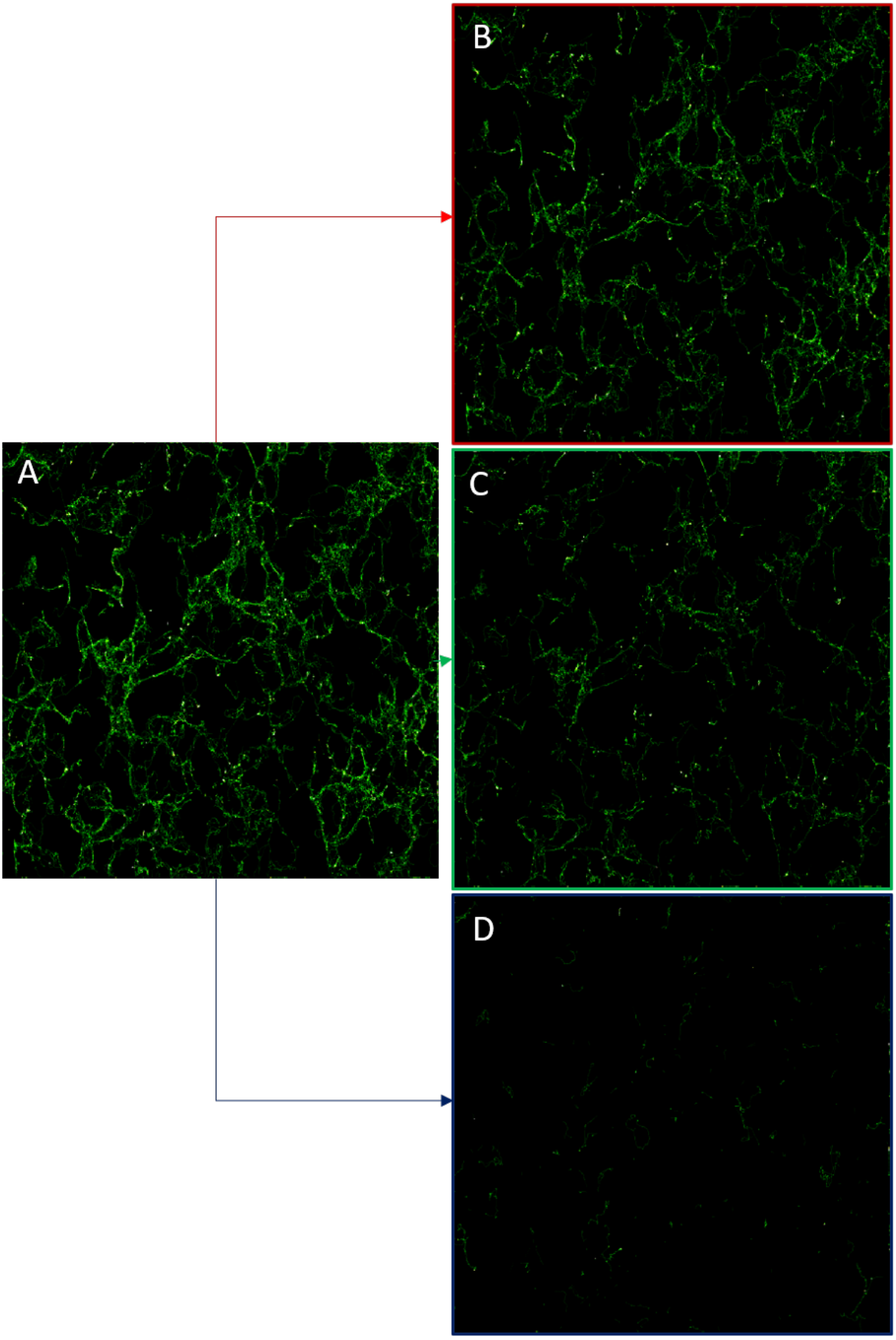



Figure 5.9: The general movement heatmap for short cancer data set A (A) then after separation into CancerTurning (B), CancerDirect (C) and fraction (D) filtered sub-populations (Figure 5.10 B:red C:green and D:blue respectively)

Filtering the first short 84 increment cancer dataset through our newly trained neural net for classification separates a starting set of 2433 tracks into 1047 CancerTurning, 1091 CancerDirect and 291 CancerFrag sub populations (Figure 5.9, 4 tracks were lost. Strands are still visible in both CancerTurning and CancerDirect sub populations, in line with our expectation of population wide participation.

The sets can also be differentiated by the distribution of movement across time: CancerTurning 157,703 movement steps, CancerDirect 120,404 and CancerFrag 30,397. Comparing visual patterns after population separation shows that CancerTurning strands appear broader and brighter; more CancerTurning strand based movement over time. The CancerDirect strand representation while less broad often appears to represent the core of identifiable strands. Therefore, CancerDirect movement may be indicative of exploratory behaviour using straight paths of least resistance or likeliest placement within a population wide trend, the middle of strands.

**Overlaying behavioural placement** Taking the three separated sub-populations and converting heatmaps through grey-scale luminosity based images we can utilise the Fiji software to overlay the sub-populations in an RGB visualisation colour scheme (Figure 5.10). Within the overlaid representation, we see the previous strand like patterns, green CancerDirect tracks rarely occur without nearby red CancerTurning track movement.

Red tracks appear to comprise the majority of strand widening patterns and are developing outlying strands without green motion. Green movement appear to primarily consist of strand core restriction, the straightest tracks are found in movement permissive areas of strands.

**Long real-world dataset separation**

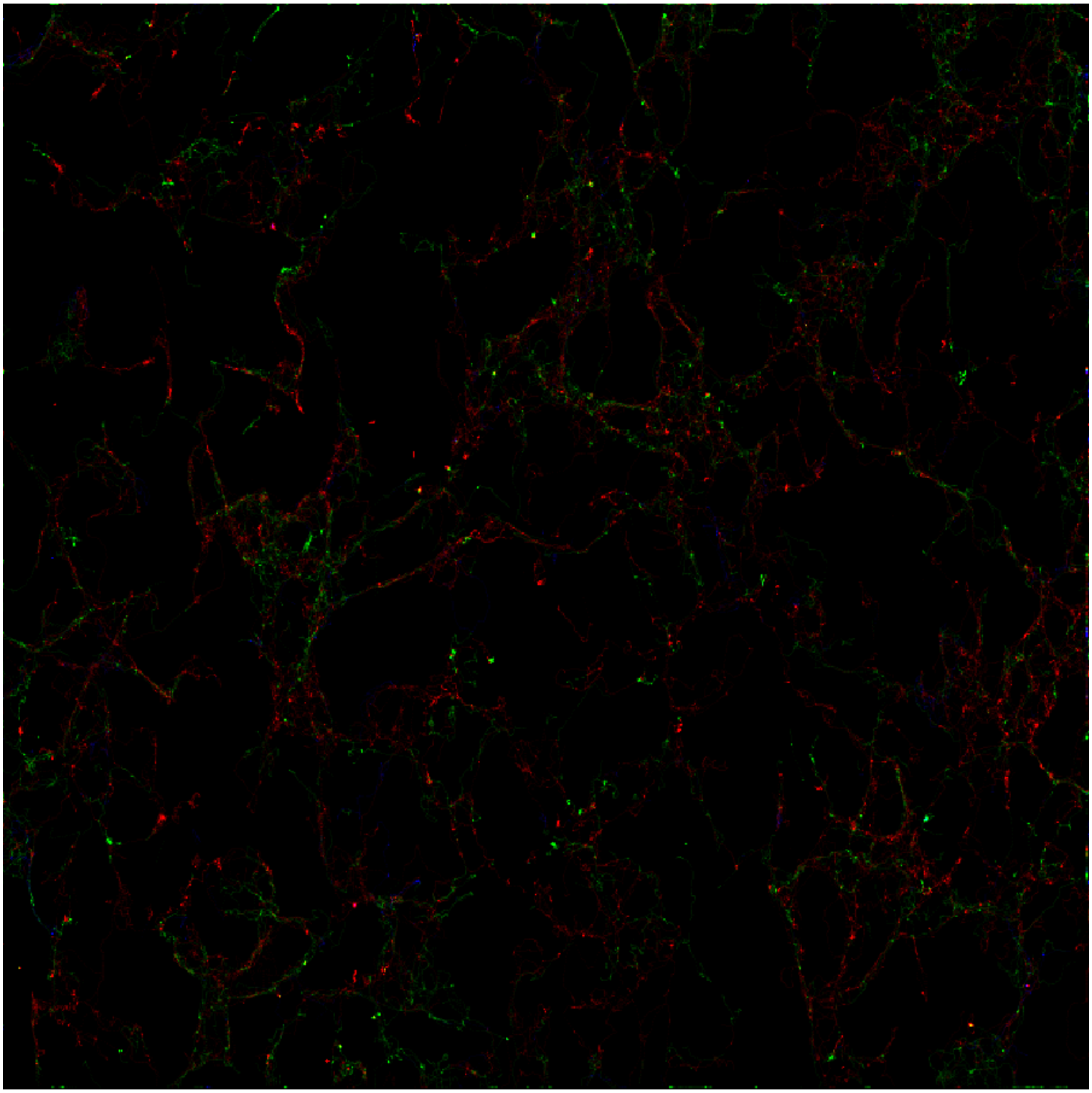


Figure 5.10: Population wide movement heatmap for the first short 84 increment cancer track set is divided via our trained neural net classification into three sub populations (Figure 5.9) and overlaid to generate a colour coded version. *CancerTurning* red, *CancerDirect* green and blue *CancerFrag* tracks are all included.

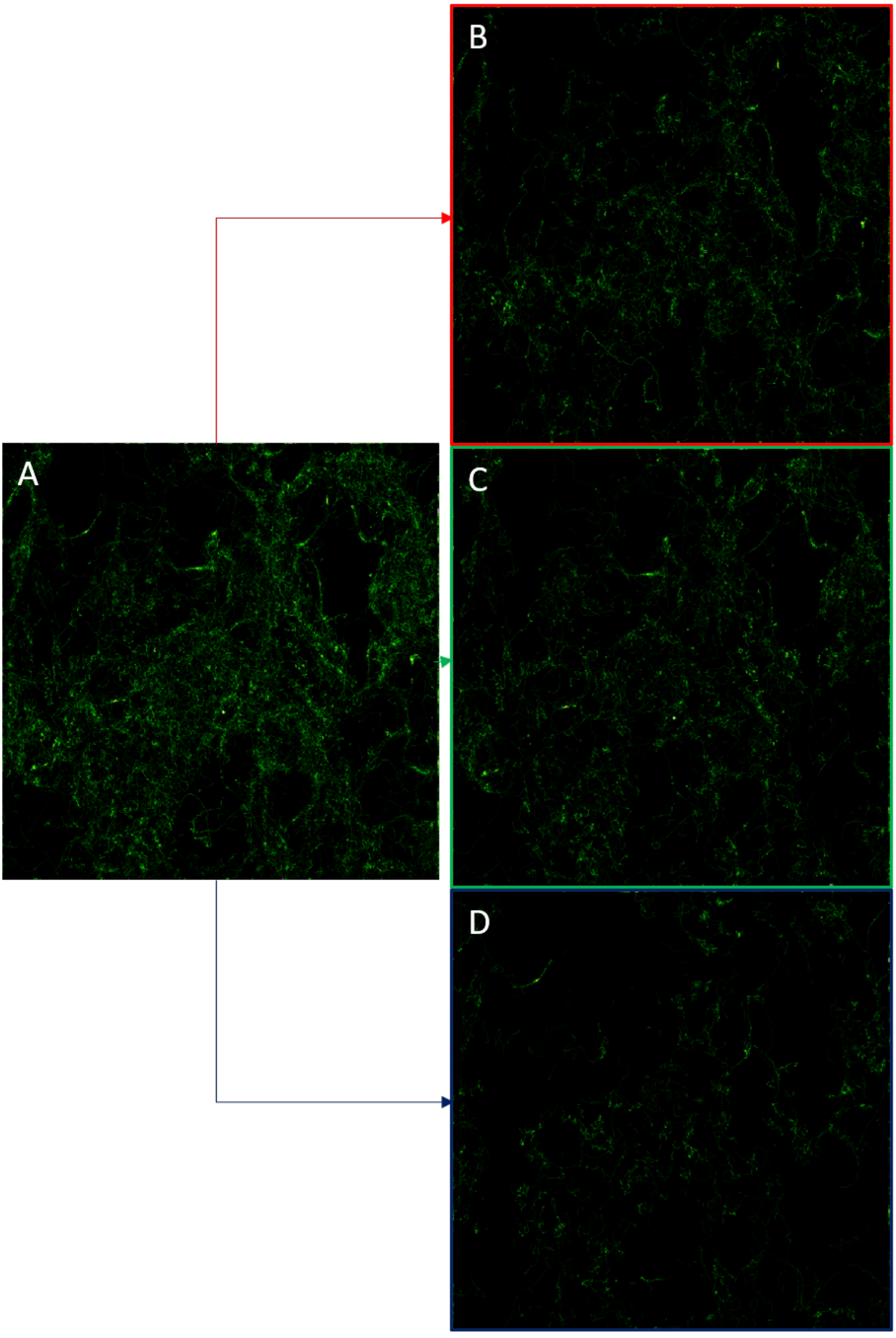



Figure 5.11: The general movement heatmap for the long cancer dataset (A), after separation into *CancerTurning* (B), *CancerDirect* (C) and *fraction* (D) filtered sub populations (Figure 5.12 B:red C:green and D:blue respectively).

The same neural net trained for separation into *CancerTurning*, *CancerDirect* and *CancerFrag* was also applied to the longer 181 increment cancer data set. Previously we observed the longer cancer set consisted of more generalised movement with some strand like patterns (Figure 5.11). Starting with 2,486 original tracks sub populations of 953 CancerTurning, 1,068 CancerDirect and 460 CancerFrag were created.

Movement distance differs across sets, CancerTurning has 195,355 steps, CancerDirect 209,276 and CancerFrag 80,252. While the same overarching visual pattern is repeated across all sets sub-populations with clear differences are still generated in small micro-environmental areas. All appear to have general and strand pattern participation; CancerFrag track ubiquity suggesting tracking loss was also generally present. Some small strands consist of only CancerDirect tracks, more strand like movement within general movement areas are clearer within CancerTurning track representation.

**Overlaying behavioural placement** While clear differentiation of patterns is still difficult even with RGB overlay of the longer cancer dataset some trends can be observed (Figure 5.12). Green/CancerDirect tracks again seem to make up the core of strand movement with red/CancerTurning exploration stretching out and between them, red areas may represent more difficult movement and path-forging. Areas where CancerDirect movement occurs alone tend to be in strong strands, where there is very little general CancerTurning movement many small CancerFrag tracks are visible. Blue/CancerFrag tracks correlate with movement density, more cells make tracking of individuals more difficult more likely to create fragmented tracks. One branch does seem to be entirely made of fractions however contrasting strangely, perhaps a very small cell moving rapidly.

### CancerTurning, CancerDirect and CancerFrag sub-population trends

Once separated into *CancerTurning*, *CancerDirect* and *CancerFrag* sub-populations, the resultant datasets can be passed back through the framework tools from the previous chapters for behaviour analysis. Quantifiable metrics such as population size and turn preferences can again be contrasted with movement heatmaps and phased visualisation.

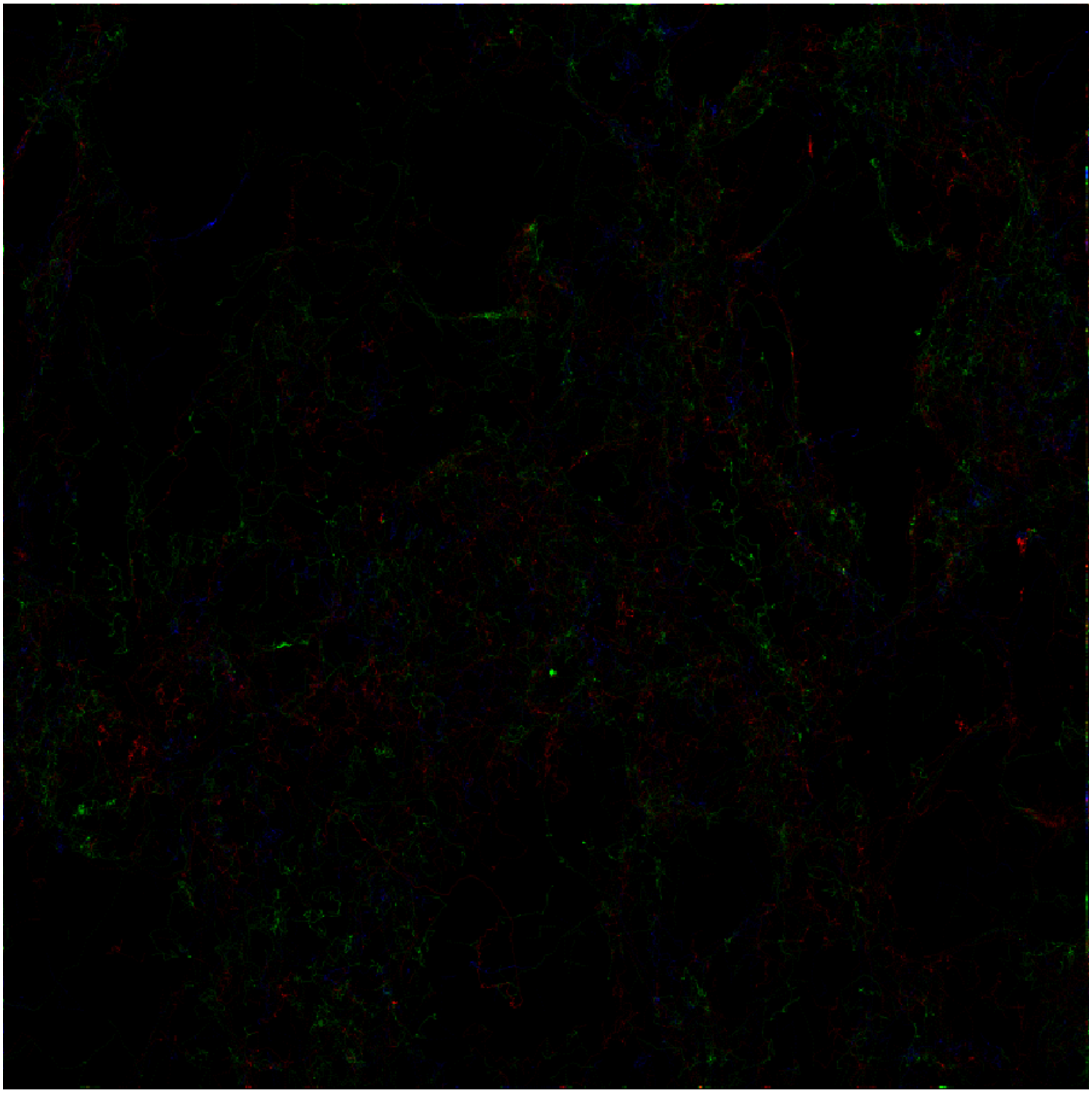


Figure 5.12: Population-wide movement heatmap for the longer 181 increment cancer track set is divided via our trained neural net classification into three sub populations and overlaid to generate a colour coded version. CancerTurning red, CancerDirect green and blue CancerFrag tracks are all included.

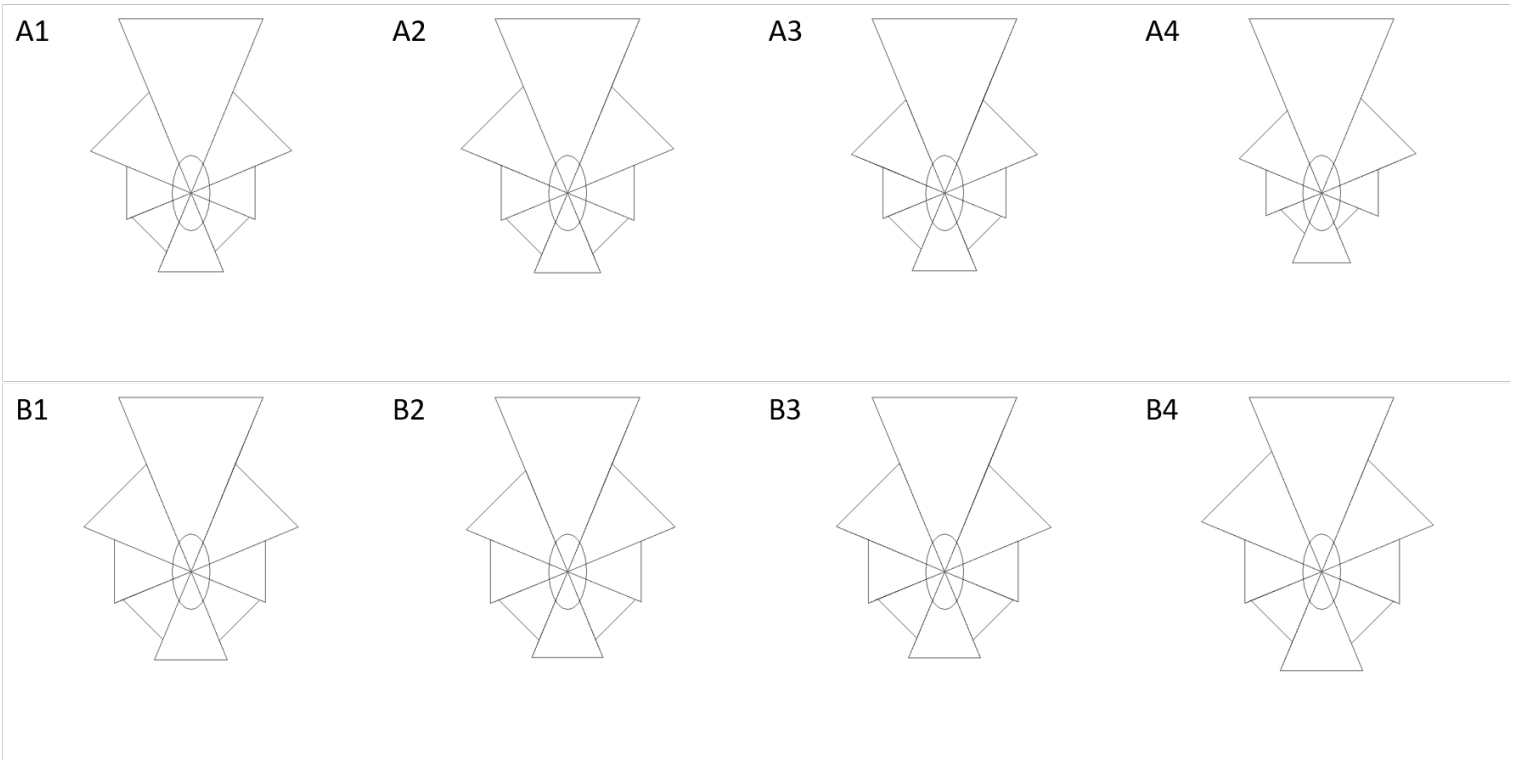



Figure 5.13: Turn diagrams for short (A) and long (B) cancer datasets, pre-filtering (1), CancerTurning (2), CancerDirect (3) and CancerFrag (4) representations. The larger a segment relative to the others the more prevalent selection of that direction was relative to forward (north) across a population and time.

**Turn preference** Before and after filtering, the turn trends displayed by populations of the short and long cancer datasets match previously observed patterns (Figure 5.13). Both short and long sets display strong forward bias indicative of following behaviour, continuing motion in the same direction for extended periods of time. When averaged for short and long runs, the general turn pattern remains uniform. Only the CancerFrag subset of our short cancer set shows a noticeable difference, an even more pronounced forward bias and slight right tilt, attributable to a circular pattern that can be observed in the centre of the corresponding heatmap.

**Very short term behavioural patterns** Phased progression visualisation of the longer cancer set is difficult to interpret (Figure 5.14). However, there is visual indication of a coalescence from light general movement to strand like groupings in both CancerTurning and CancerDirect sets for the long sub-populations. Interestingly, there is also some visual suggestion of an initially quite restricted population spreading out but then coalescing into common strands, more so in the CancerDirect set than CancerTurning. Certainly, visual strand patterns are clearer in later images for both long sets despite little to no increase in overall movement per captured time

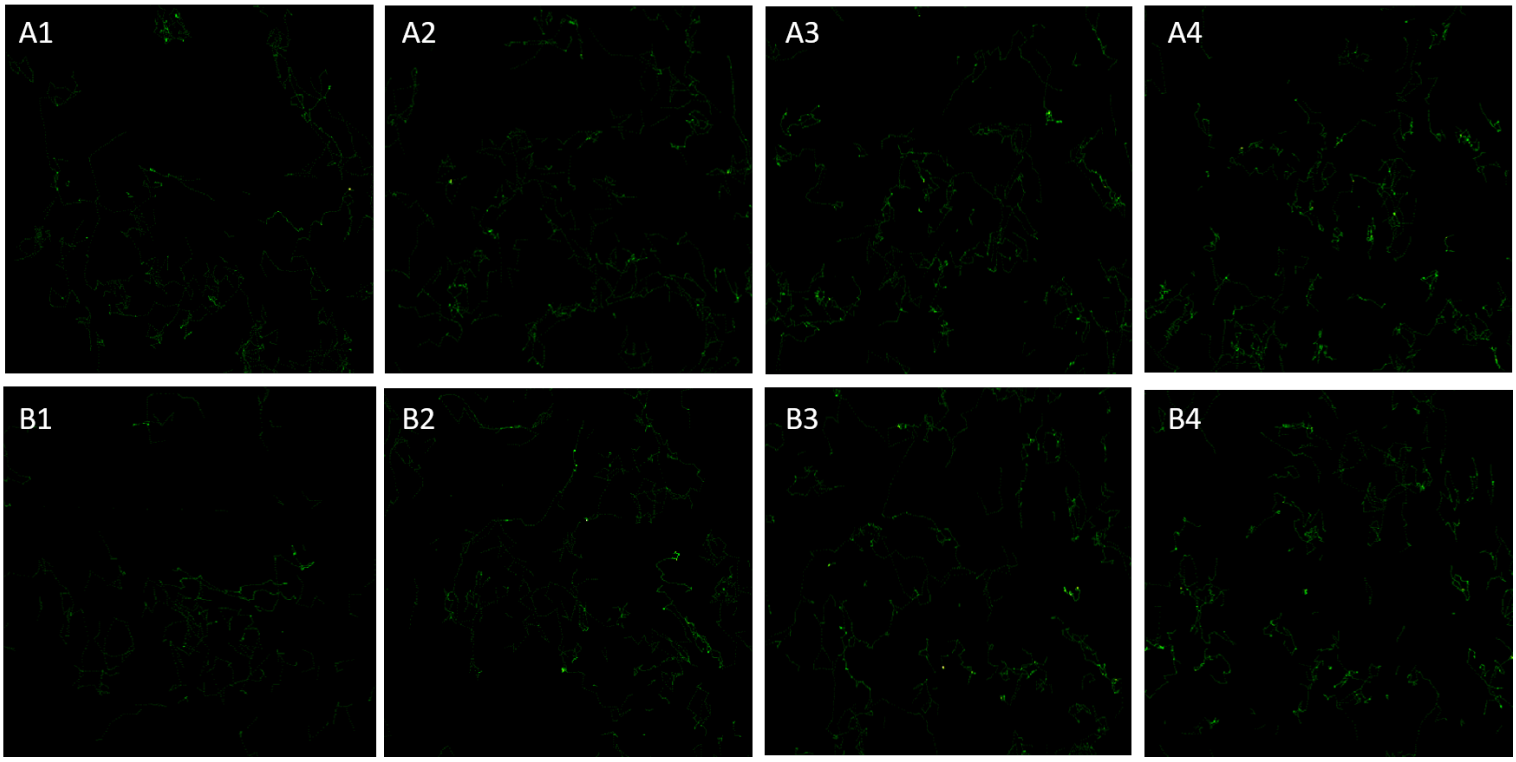



Figure 5.14: Movement heatmaps of long cancer set filtered sub populations zoomed to the centre quarter, CancerTurning (A) and CancerDirect (B) after 45 (1), 90 (2), 135 (3) and 180 (4) time increments.

frame.

#### GPCR and G protein net on Short cancer data set

Separating the short cancer set with the GPCR and G protein trained net yields some interesting visually comparative patterns (Figure 5.15). The full set of 2,433 tracks are sub divided to 502 GPCRBrown, 435 GPCRShiver and 1,492 GPCRFrag sub-populations.

Movement correlating brightness is similar for both GPCRBrown and GPCRFrag sets with suggestions of denser zones for the GPCRFrag set. However, the population preponderance disparity suggests that movement density is due to tracking reliability in the GPCRBrown set and increased population size for GPCRFrag. A few compression based hot-zones are identified and sub-divided from the other movement sets but with very low occurrence.

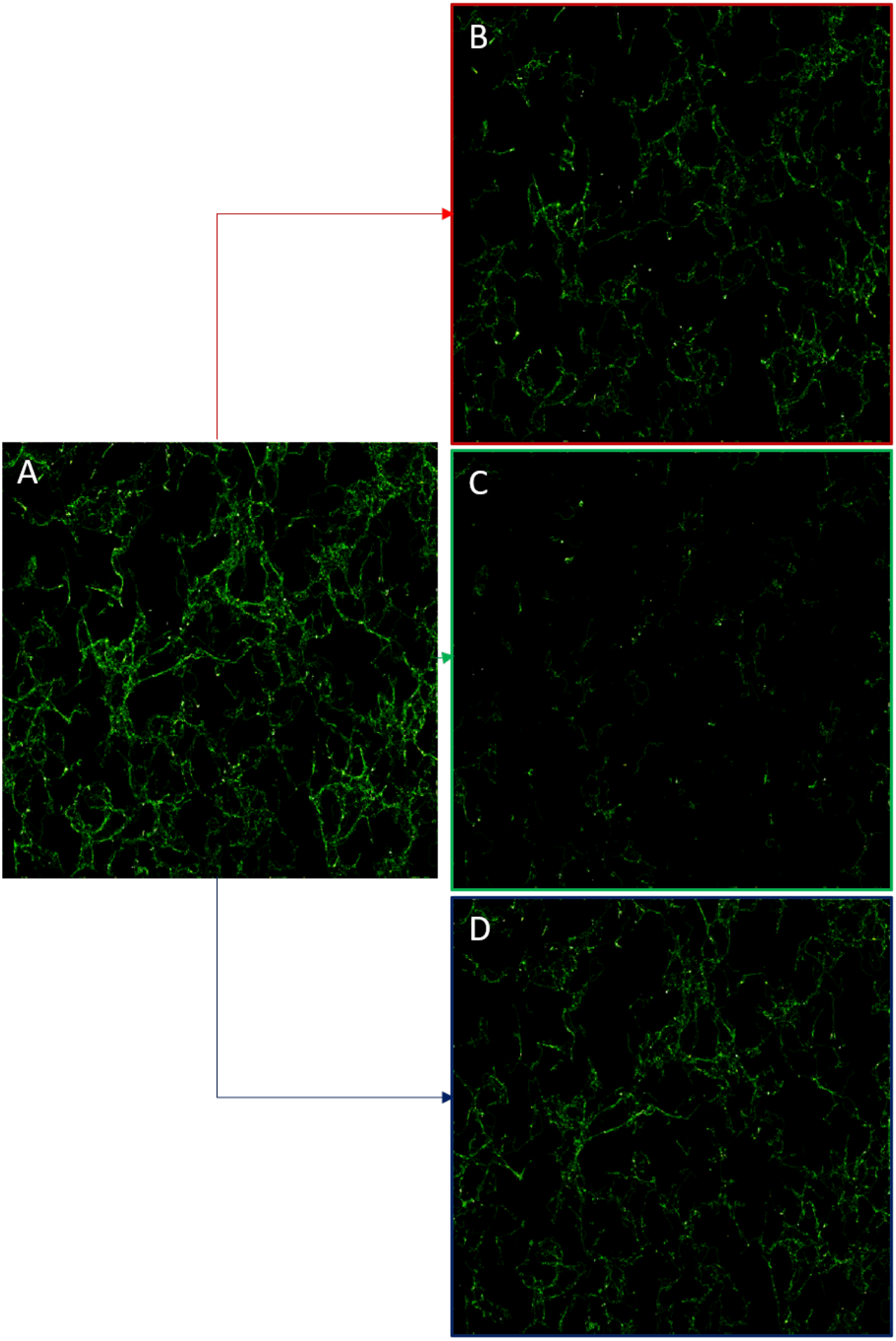



Figure 5.15: Movement heatmaps for the first short cancer dataset pre- (A) and post- filtering to GPCRBrown (B), GPCRShiver (C) and GPCRFrag (D) sets from the GPCR and G protein filter net.

**GPCR and G protein net on long cancer dataset:**

With the longer cancer set, separation again places most movement within the GPCRFrag set at 255,565 movement steps and more visual general movement heat than GPCRShiver movement 112,753 and the GPCRBrown set 116,566 (Figure 5.16). Visualisation of the different subsets was achieved by separation of the starting large 2486-member set into 474 GPCRBrown, 636 GPCRShiver and 1,371 GPCRFrag tracks.

Unlike the short cancer sets, movement density followed most tracks to the GPCRFrag set. However, the remaining sets seem to display a more strand-like behaviour like that of divided short sub-populations. Some of the harder to identify strand perturbations are also only present in the GPCRShiver set representing shivering, suggesting creation via a small subset of tracks with low but constant localised movement.

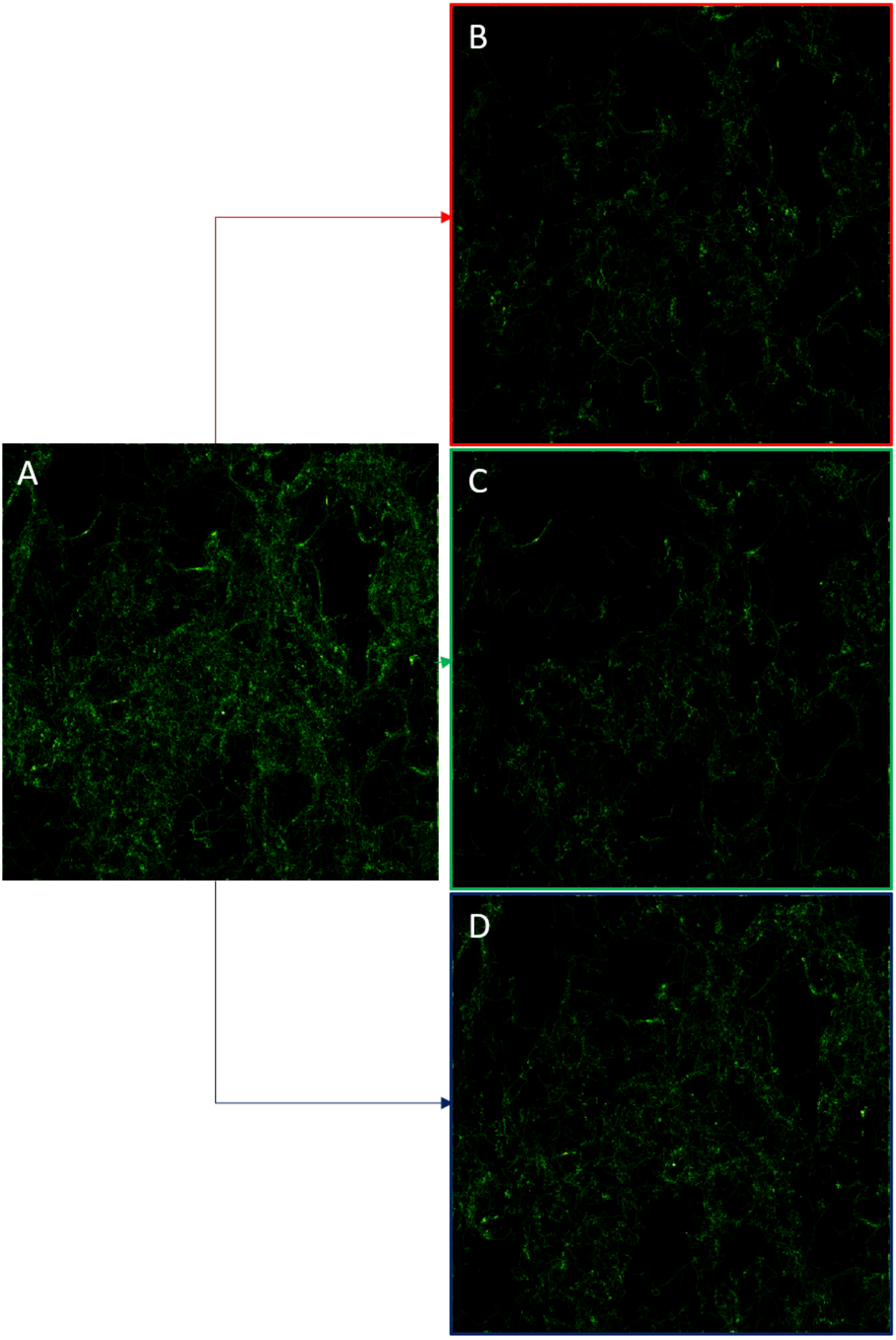



Figure 5.16: The longer 181 increment cancer track set (A) broken into GPCRBrown (B), GPCR-Shiver (C) and GPCRFrag (D) sub sets via a GPCR and G protein trained net then visualised via movement heatmaps.

### 5.3.2 GPCR and G proteins

We are aiming to improve the comparison between real-world *in vitro* observations and model generated heatmaps via the GPCR and G protein specific model similarity pipeline. This should also improve differentiation between hot-zone types and between hot-zones and general background motion.

#### Comparing real-world and representative model movement patterns

Unlike the cancer-specific developed representative models, the comparison between all GPCR and G protein model (Chapter 3 3.2.1) generated results represented a much more difficult training case; not impossible but prohibitively resource and time intensive. To account for the increased comparative difficulty, we reduced the approach to a *two model adversarial* comparison. The CNN is trained to differentiate between models A and B, and has an accuracy classifying those against a held out training set, e.g. 97%. We then present a trained CNN with C (the real-world image) it is 33% confident it is A, 77% confident it is B. The 77% suggests there is more visual similarity between the real-world input and the model trained set B. This is an important assumption: while a two model similarity comparison classification becomes more reliable, the comparison between models becomes more difficult to interpret. To reduce unreliability and obfuscation, we can average across available datasets for GPCR (C1) and G protein (C2) also quantifying the standard deviation of each comparative set. We may not be able to describe exactly how the net is classifying, but, like a POM or ensemble approach, we can support the validity of already observed results with less individually informative analysis and an emphasis upon exploratory hypothesis generation.

**Receptor and model movement heatmap comparison** Using the bi-modal model similarity pipeline, we generate averaged classification values that represent confidence of real-world association with one or other model results (Figure 5.17 and Tables 5.4,5.5). We applied the previous representative models with attractive and deflective areas, just attractive, simple background Brownian motion, and only deflective phenomena (Chapter 3).

The results suggest that for C1 real-world sets, the closest general association is with purely at-

| C1 | AD | AT | BG | DF | SSP |
|---|---|---|---|---|---|
| AD | X | 0.65 | 0.40 | 0.33 | 0.58 |
| AT | 0.35 | X | 0.33 | 0.29 | 0.52 |
| BG | 0.60 | 0.67 | X | 0.42 | 0.6 |
| DF | 0.67 | 0.71 | 0.58 | X | 0.68 |
| SSP | 0.42 | 0.48 | 0.40 | 0.32 | X |

Table 5.4: Trained net comparison confidence averages across all C1 GPCR and G protein datasets for comparative model result pairs top row and left column are the types of representative model available for comparison (Chapter 3). Each value is the confidence average where real-world sets belong to, the top row representative model set against the left. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

| C1 | AD | AT | BG | DF | SSP |
|---|---|---|---|---|---|
| mean | 0.51 | 0.63 | 0.43 | 0.34 | 0.60 |
| max | 0.67 | 0.71 | 0.58 | 0.42 | 0.68 |
| min | 0.35 | 0.48 | 0.33 | 0.29 | 0.52 |
| range | 0.32 | 0.23 | 0.26 | 0.13 | 0.16 |
| median | 0.51 | 0.66 | 0.40 | 0.32 | 0.59 |

Table 5.5: Averaged trained net comparison confidence averages across all C1 GPCR and G protein datasets for comparative model result pairs (Table 5.4 (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

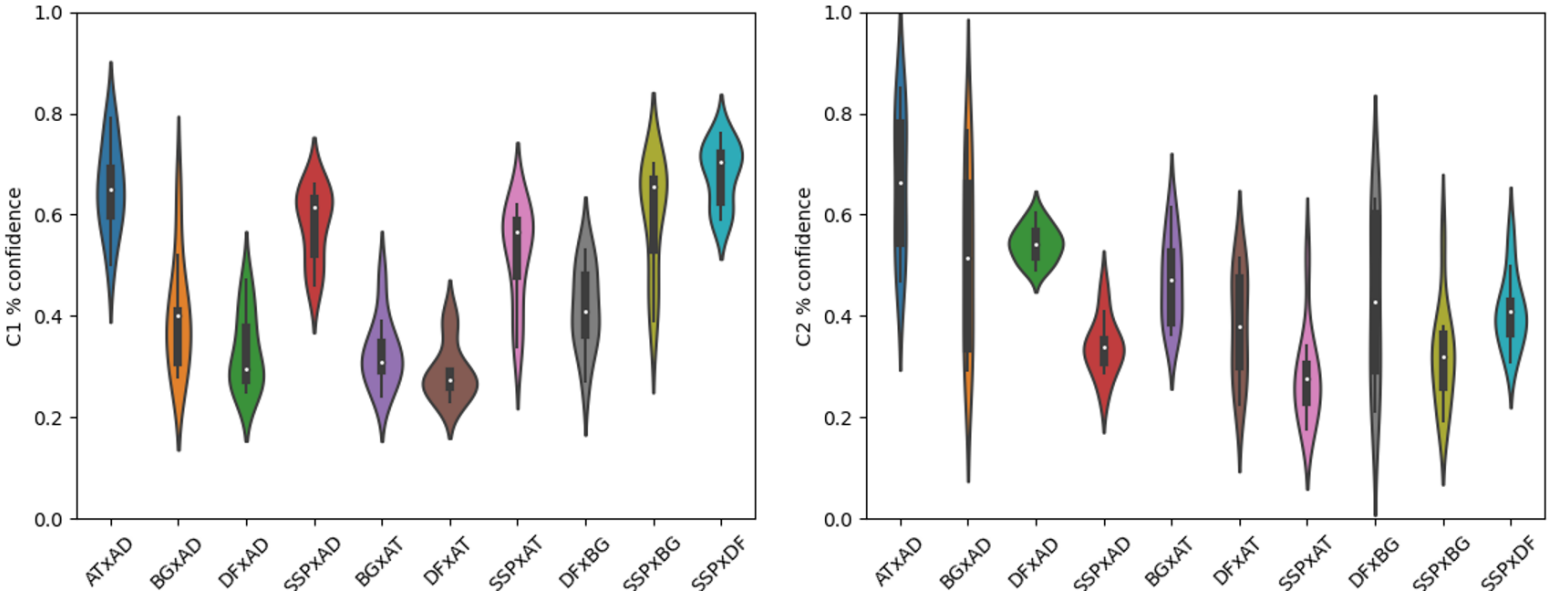



Figure 5.17: Confidences plot for the two model similarity results. All available C1 and C2 heatmaps inputted to adversarial comparison pairs, 'ADxAT' is a net comparing training data from hybrid Attractive and Deflective with attractive only models, the resultant values were then averaged to form a confidence grid (Table 5.4). Results can be horizontally mirrored to flip the relationship, 'ADxAT' becomes 'ATxAD'. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

tractive area representations 0.48-0.71 confidence, stationary sub-populations 0.52-0.68 and then attractive with deflective zones 0.35-0.67. The comparison suggests that real-world C1 entities are more likely directed towards attractive areas in their environment but with some shivering or highly compressed population members present. Between deflective and purely background movement, the model similarity approach suggests that the association of real-world results is closer to simple Brownian motion 0.33-0.58 than with the inclusion of deflective walls 0.29-0.41. Such observations further support our hypothesis that a deflection only system is unlikely.

**GPCR comparison reliability** We lose information on the distribution of C1 and C2 comparative results by encapsulating the range of available model similarity set values in a single average. The spread of classification can be evaluated with the standard deviation (Figure 5.6). A given model type may perform well in general but have wide variation classifying different sets; the pattern or patterns that a CNN uses may not be as distinct or present in the target set. We can scrutinize the score spread that indicates the applicability, or confidence, of model comparison results (Figure 5.17). When a population displays consistent scoring the classification value is

| C1 deviation | AD | AT | BG | DF |
|---|---|---|---|---|
| AT | 0.09 | X | X | X |
| BG | 0.11 | 0.07 | X | X |
| DF | 0.07 | 0.05 | 0.08 | X |
| SSP | 0.07 | 0.09 | 0.11 | 0.06 |

Table 5.6: Standard deviation for each pair of model similarity comparison trained data types when applied to the full set of available GPCR and G protein C1 track movement heatmaps. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

| C2 | AD | Attractive | BG | DF | SSP |
|---|---|---|---|---|---|
| AD | X | 0.66 | 0.51 | 0.54 | 0.34 |
| AT | 0.34 | X | 0.47 | 0.38 | 0.28 |
| BG | 0.49 | 0.53 | X | 0.43 | 0.32 |
| DF | 0.46 | 0.62 | 0.57 | X | 0.41 |
| SSP | 0.66 | 0.72 | 0.68 | 0.59 | X |

Table 5.7: Trained net comparison confidence averages across all C2 GPCR and G protein datasets for comparative pairs top row verses left, each value is the confidence average that sets belong to the top row set of any comparative pair. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

more reliable for the set type comparison. However, when scoring is highly variable, as is more dominant in the C2 case, it is very difficult to determine whether the set itself contains a wide range of pattern adherence or a low applicability to model comparison. The high accuracy on training data should indicate that population patterns vary rather than CNN selection but we cannot be certain.

Attractive zone combined deflective and stationary sub-population models appear to create the highest resultant standard deviation when compared to background movement. However, no single model type displays strong deviation trends in all its comparative pairs; uniformly high deviation indicates possible random selection through lack of clear differentiation. Lower deviation numbers are not entirely clear but the deflective boundary pattern seems to generate least variance indicating it is based upon a strong visual queue.

| C2 | AD | AT | BG | DF | SSP |
|---|---|---|---|---|---|
| mean | 0.48 | 0.63 | 0.56 | 0.49 | 0.34 |
| max | 0.66 | 0.72 | 0.68 | 0.59 | 0.41 |
| min | 0.34 | 0.53 | 0.47 | 0.38 | 0.28 |
| range | 0.32 | 0.19 | 0.21 | 0.21 | 0.13 |
| median | 0.47 | 0.64 | 0.54 | 0.49 | 0.33 |

Table 5.8: Averaged trained net comparison confidence averages across all C2 GPCR and G protein datasets for comparative model result pairs (Table 5.7 (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

| C2 deviation | AD | AT | BG | DF |
|---|---|---|---|---|
| AT | 0.14 | X | X | X |
| BG | 0.18 | 0.09 | X | X |
| DF | 0.03 | 0.11 | 0.17 | X |
| SSP | 0.06 | 0.09 | 0.10 | 0.07 |

Table 5.9: Standard deviation for each pair of model similarity comparison trained data types when applied to the full set of available GPCR and G protein C2 track movement heatmaps. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

**G protein and model movement heatmap comparison** The C2 movement heatmaps were also processed with the model similarity comparative nets to generate classification confidence results (Figure 5.17 and Tables 5.7,5.8).

Attractive zone application models appear to have the closest classification similarity with C2 sets 0.53-0.72 as in C1 0.48-0.71. In particular, the general background model movement has become the second most prevalent classification in contrast to the C1 results 0.47-0.68. Stationary sub-populations are now the least similar set with confidences values of 0.28-0.41 just behind attractive and deflection combined models 0.33-0.67. Therefore, C2 sets consist of more general movement than C1 but with a reasonable number of large attractive zones and fewer shivering sub members. The deflective 0.38-0.59 set is still generally weaker than background alone 0.47-0.68 but less so than C1.

| Validation set | AD | AT | BG | DF |
|---|---|---|---|---|
| AT | 93% | X | X | X |
| BG | 100% | 99% | X | X |
| DF | 97% | 99% | 98% | X |
| SSP | 100% | 100% | 100% | 100% |

Table 5.10: Validation accuracy of trained CNN's when presented with a held out labeled dataset for all comparative model similarity pairs used immediately prior to input of GPCR and G protein data through the same trained net, the same validation sets were used for C1 and C2 runs. (Attractive and Deflective (AD), Attractive (AT), Background (BG), Deflect (DF), Stationary sub-population (SSP))

**GPCR and G protein comparison reliability** Some of the highest deviation values across both C1 and sets can be observed when classifying the C2 group; an average of 0.1 for C2 vs 0.08 for C1(Figure 5.9). Combined attractive and deflective models when compared to just attractive and background results generate deviance much greater than the C1 equivalent. Background and deflect comparison also generate a similarly high standard deviation across real-world datasets. Deflective and attractive comparison shifts from the lowest deviance to the middle of the set with the background and stationary subset pair. Interestingly, despite generally being higher, the lowest standard deviation in C2 0.03 is lower than recorded in C1 0.05 between combined attractive and deflective models when comparing with deflective alone implementations. A greater preponderance of general movement may well make distinct patterns more difficult to identify, in turn affecting results when real-world datasets are presented to the model trained nets.

**Validation accuracy** The previous cancer comparison set functioned with an 89% validation accuracy; when presented with a held out training set, we can record similar validity for each model similarity GPCR and G protein set (Figure 5.10). For the model similarity pairs, the validation set differentiated well, generating almost 100% accurate classification across 100 previously unseen held out model generated and labelled results. Classification of deflective vs other models creates the only consistent inaccuracy 1-3% suggesting a pattern that is difficult for CNN's to recognise; the primary pattern being an absence of movement in an area. Greatest differentiation difficulty seems to occur between combined attractive and deflective models with attractive only examples. Every net was trained with the same number of training and testing examples, 800 and 200 respectively, each required the same number of epochs and steps with identical batch and image size. In contrast, the cancer net required five times as many training steps and a

significant additional compute cost for a potentially less accurate classification result.

#### Trajectory compression classification and population subsets

Utilising our training workflow, framework generated datasets and visualisation libraries, we developed a single-track classification CNN for GPCR and G protein sets.

**Classification and separation of receptor movement** For set TC641 we applied the filtering net to both C1 and C2 runs. The C1 run was comprised of 4776 tracks and broken into subsets of 518 GPCRBrown, 1,186 GPCRShiver and 3,072 GPCRFrag tracks (Figure 5.18). We can also estimate large distinct hot-zones with a transform based upon luminescence and then 3D object count in Fiji [62], from the C1 set of 48 objects GPCRBrown (18), GPCRShiver (33) and GPCRFrag (15) sets are created.

Most of the hot-zone movement has been captured within the GPCRShiver subset. There is still representation in the other two sets but much reduced. Where hot-zones are within the GPCRBrown set, they appear to be larger and more generalised. We previously identified lower forces of attraction or disassembly of the phenomena may be the cause. Additionally, we can observe that where the GPCRBrown set displays some GPCRShiver similarity, it is of entities with pronounced normal movement prior to entering a hot-zone. The GPCRFrag track subset is the largest, suggesting tracking loss is prevalent but also that the group may capture a large portion of edge cases. General movement in the GPCRBrown set also seems more cohesive than the GPCRFrag one, likely long tracks verses lots of small movement jumps in a similar area.

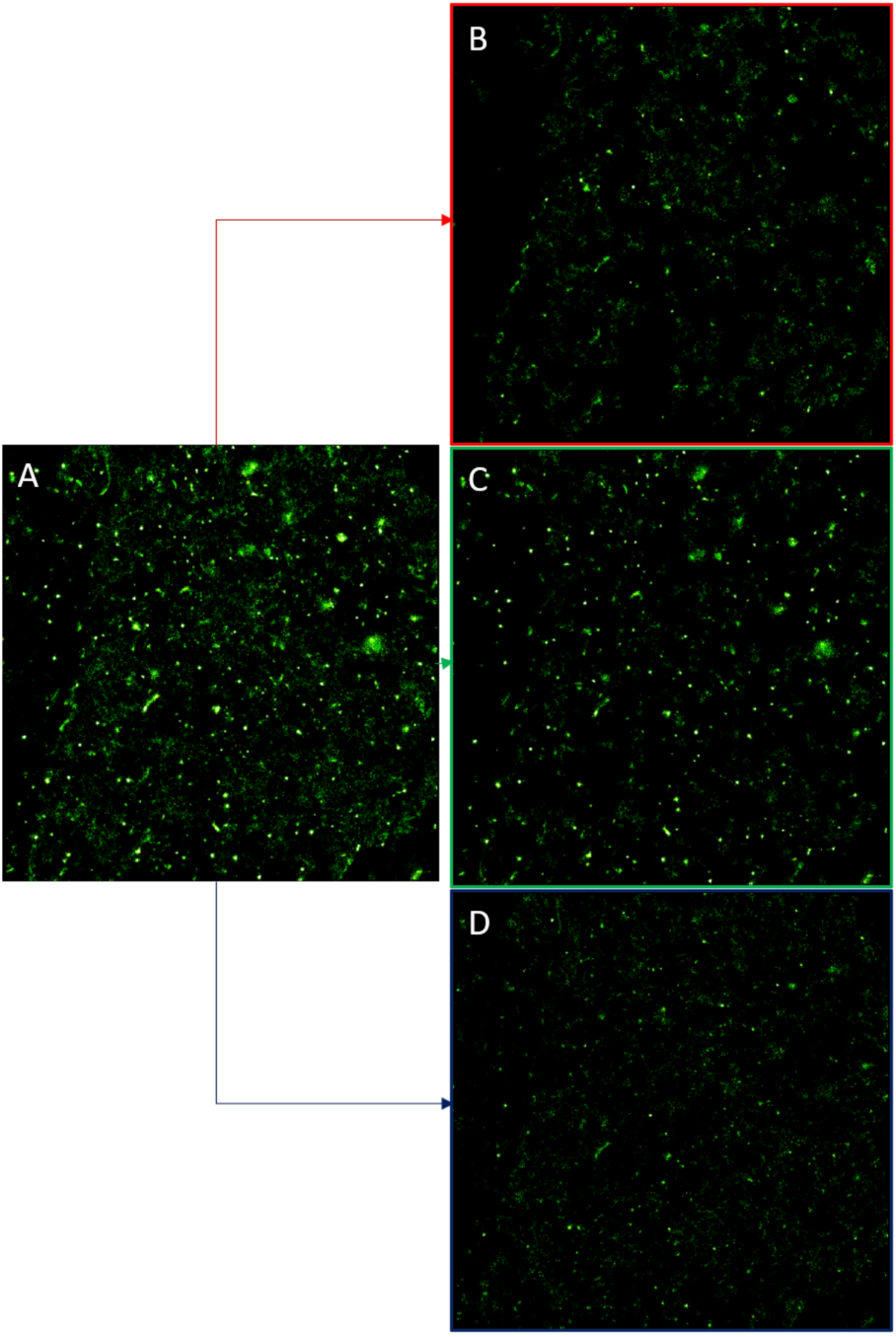



Figure 5.18: General movement heatmaps for TC641 C1 (A) and subsets of GPCRBrown (B), GPCRShiver (C) and GPCRFrag (D) population movement. Also represented as red(B), green(C) and blue(D) tracks (Figure 5.19)

**Overlaying behavioural placement for receptor movement** Reconstructing the original TC641 C1 dataset by overlaying the heatmaps of all three separated sub-populations allows us to further clarify the separation (Figure 5.19). The majority of hot-zones are clearly part of the green GPCRShiver dataset, several larger hot-zones while comprising of parts of each set are also primarily green and therefore consist of GPCRShiver tracks. Again, GPCRFrag movement seems pervasive throughout, short tracks do not seem to be restricted to non-hot-zone movement. In all three colours general movement seems to differentiate quite well, and it is possible they are tracks from isolated entities that was lost intermittently.

**Classification and separation of G protein movement** As with the TC641 C1 version, C2 was processed through our trained filtering neural net. Starting with 6601 tracks sub-populations of 1,665 GPCRBrown, 2,371 GPCRShiver and 2,565 GPCRFrag were generated. We can again also roughly count large distinct hot-zones with a transform based upon luminescence and then 3D object count in Fiji [62], from the C2 set of 138 objects GPCRBrown (41), GPCRShiver (96) and GPCRFrag (10) sets are created.

Again, the GPCRFrag sub-population is the largest, but by a much smaller margin. There also seems to be more general movement in the GPCRShiver set, so movement may trend more towards shivering in general. Hot-zones also seem to be more clearly confined in the GPCRShiver set. There is as well a similar level of general brightness and therefore movement seen across both GPCRBrown and GPCRShiver sets. General movement in the GPCRBrown set is the most cohesive suggesting clear tracking and movement from hot-zones to more permissive environments.

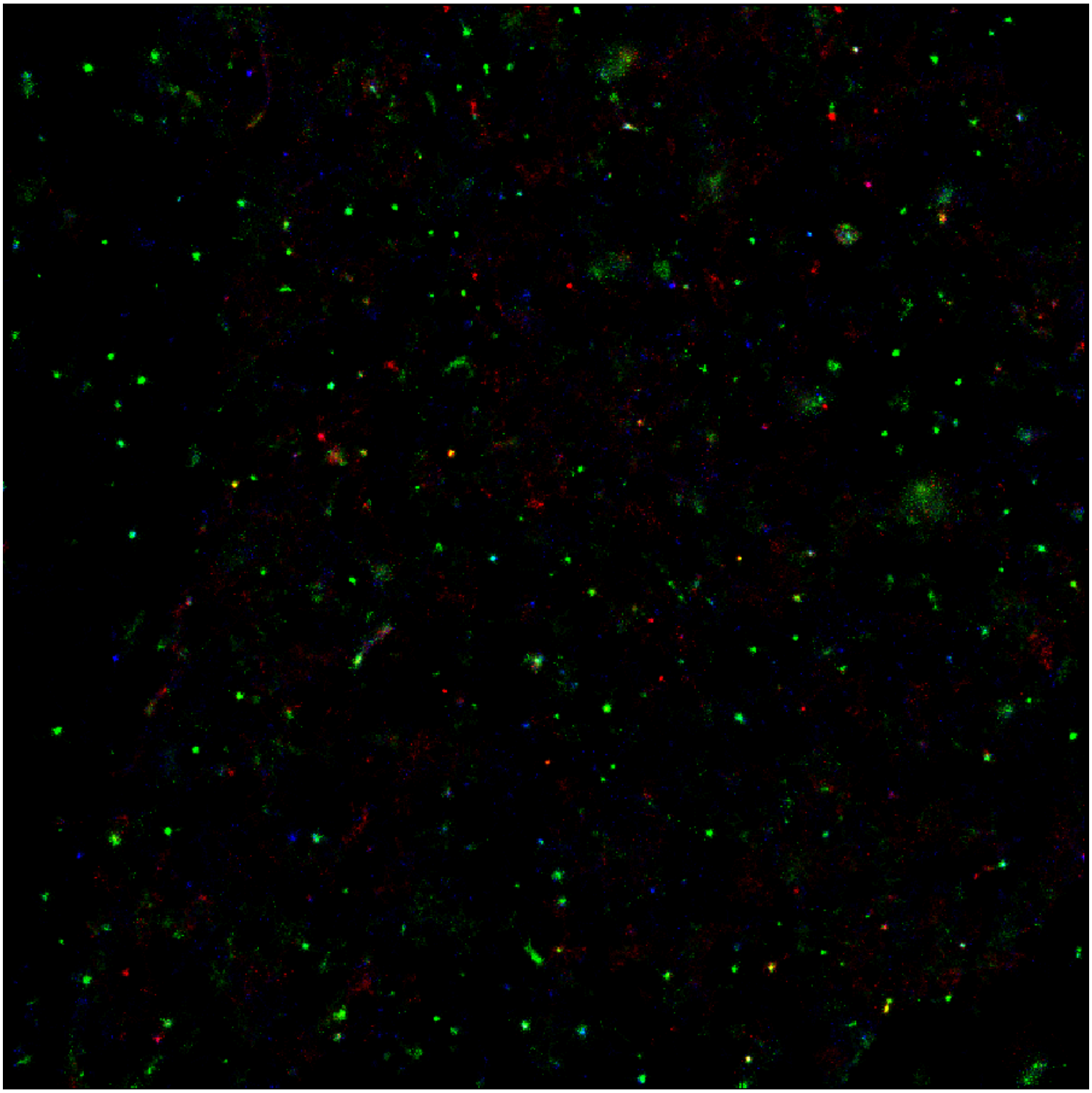


Figure 5.19: After generating subsets from the real-world set TC641 C1 a heatmap representation can be reconstructed by overlaying the three sub sets (Figure 5.18) with red GPCRBrown, green GPCRShiver and blue GPCRFrag tracks present.

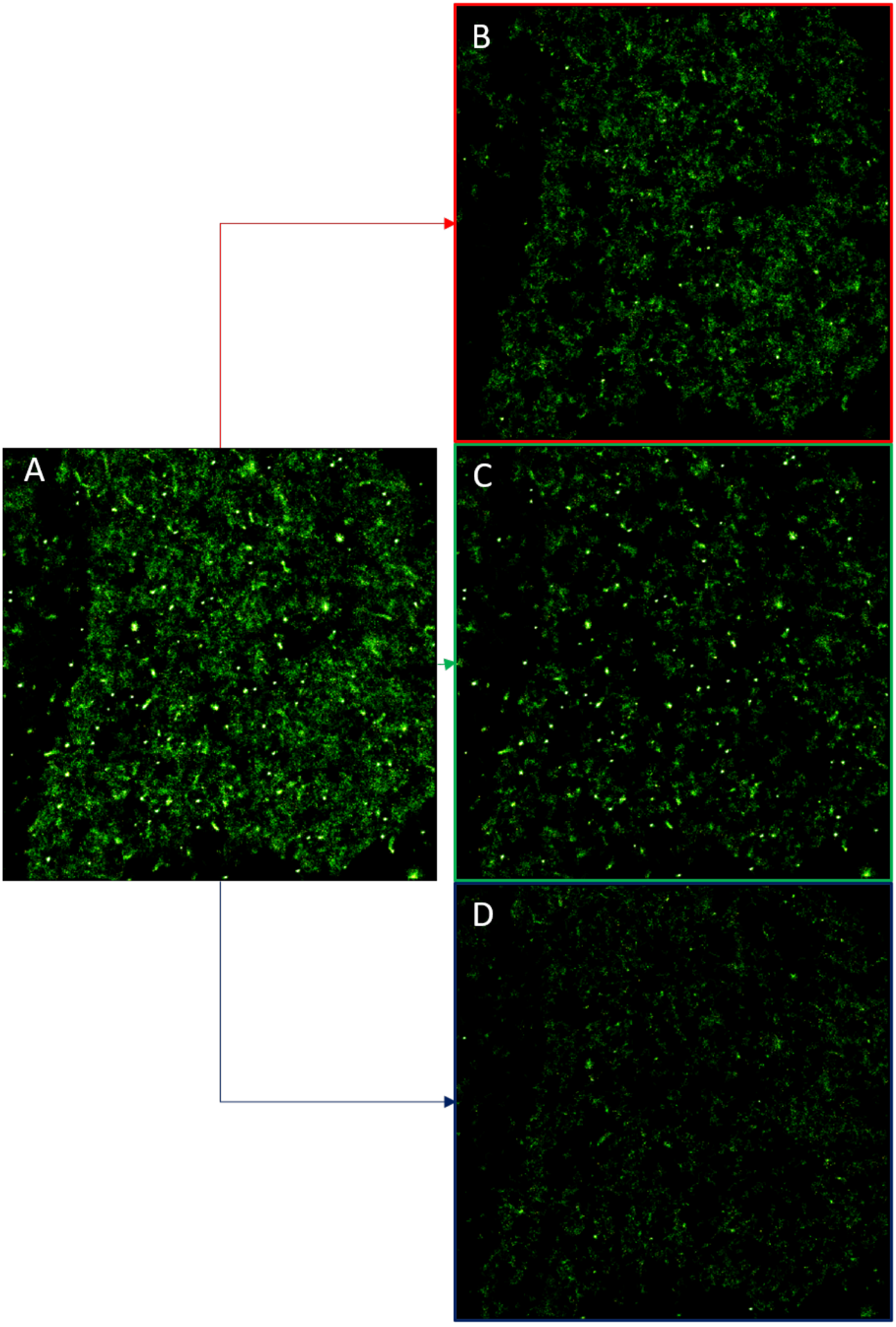



Figure 5.20: General movement heatmaps for GPCR and G protein TC641 C2 (A) and sub sets of GPCRBrown (B), GPCRShiver (C) and GPCRFrag (D) population movement. Also represented as red(B), green(C) and blue(D) tracks (Figure 5.21)

**Overlaying behavioural placement for G protein movement** Overlaying all three filtered sub-populations for the C2 dataset again helps us differentiate behaviour types (Figure 5.21). Green/GPCRShiver and red/GPCRBrown general movement are often clearly differentiated, further suggesting that a change in behaviour may lead to a tracking break and pattern change. GPCRFrag tracks are pervasive and represented in both general and hot-zone movement although less visually pronounced than the previous C1 set. Green/GPCRShiver compression set tracks account for at least part of most hot-zones with many being entirely green or with a light blue/GPCRFrag inclusion suggesting the possibility of tracking loss.

### GPCRBrown, GPCRShiver and GPCRFrag sub-population trends

By taking the sub-population output of the neural net filtering process we can process them through the framework and check whether we manage to isolate the source of turn bias. Quantitatively comparing important metrics such as travel distance and C1 with C2 hot-zone cohesion progression.

**Turn preferences** The turn diagrams generated for C1 and C2 datasets along with their corresponding sub-populations replicate previously observed rear-turn biased movement (Figure 5.22). The northern segment of diagrams still represents their previous direction, each turn is measured relative to that and added to the relevant segment. Therefore, dominance of the southern segment suggests a strong bias towards reverse directional choice at any time point across each population.

As hoped, filtering of the overall data set has noticeably affected the bias presented. Both C1 and C2 GPCRShiver subsets have a much stronger rear bias than GPCRFrag and GPCRBrown, the dominance of hot-zones in that set suggesting we have isolated a strong effect. Where general movement is more present in the C2 GPCRShiver subset the rear bias is also slightly ameliorated. Further, both GPCRBrown and GPCRFrag subsets display a reduced rear bias, nearer random movement expected of general GPCRBrown motion even when compared to the full datasets.

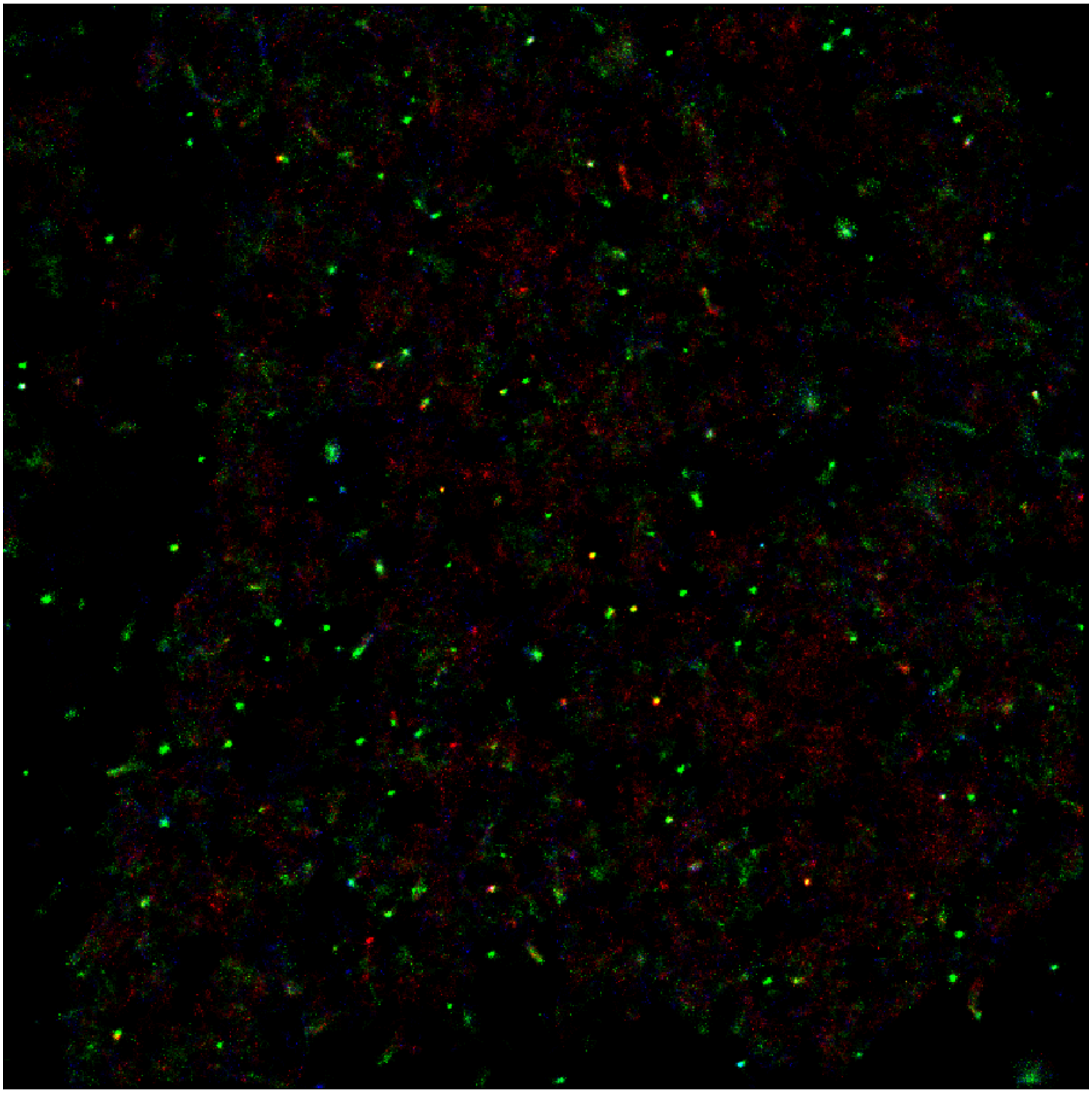


Figure 5.21: After generating subsets from the real-world set TC641 C2 a heatmap representation can be reconstructed by overlaying the three subsets (Figure 5.20) with red GPCRBrown, green GPCRShiver and blue GPCRFrag tracks present.

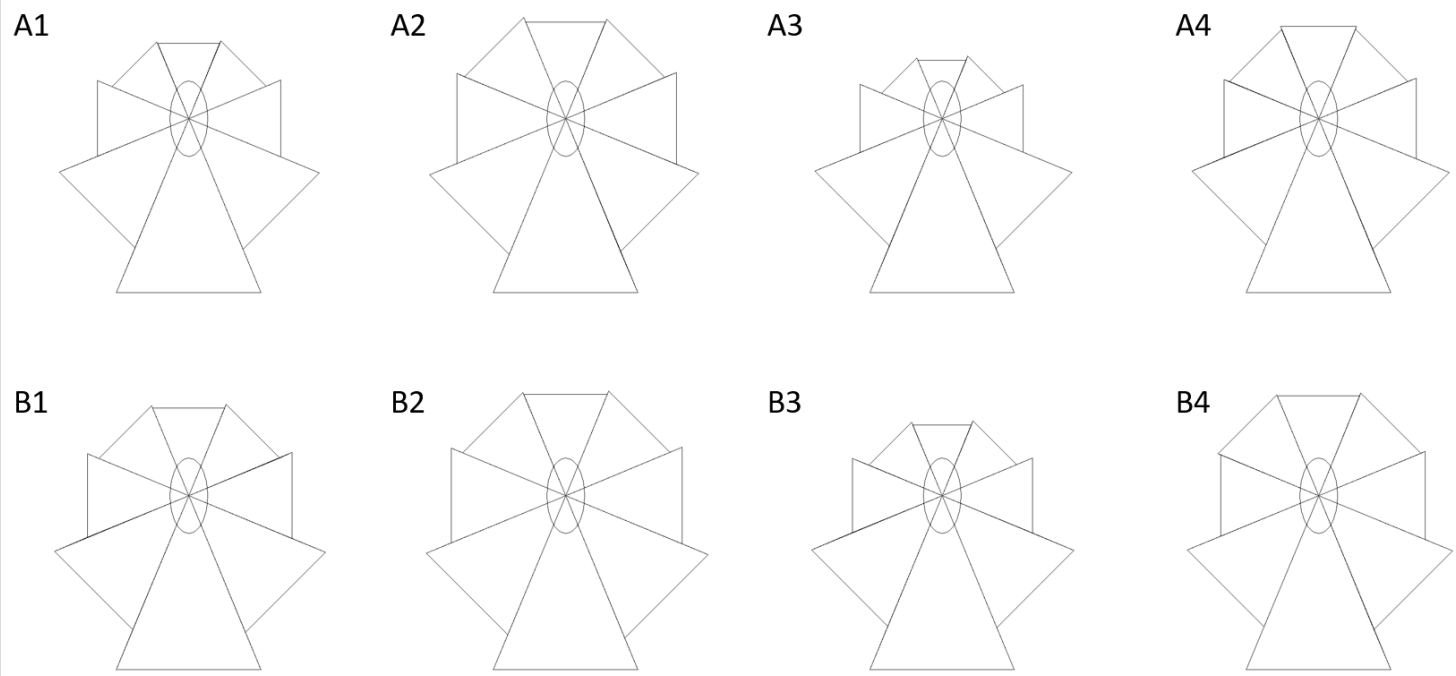



Figure 5.22: C1 (A) and C2 (B) turn diagrams across the entire set (1) and sub-populations: GPCRBrown (2), GPCRShiver (3) and GPCRFrag (4).

**C1 and C2 hot-zone overlap** We have identified that hot-zone morphology changes can be seen in subsets over time but not with clear population-wide trends or patterns. Overlaying C1 and C2 sets may allow us to identify whether coalescence events are common in shared hot-zones (Figure 5.23).

We observe some differentiation of more concentrated small C1 and C2 hot-zones and some overlap, clear co-localization is detected in 24-27 places across time phases (Figure 5.23). Also, a portion of movement previously thought of as random and incidental overlaps to a degree that suggests a larger less cohesive hot-zone is present; filtering clarified new hot-zone candidates. Filtering enables us to identify even larger possible hot-zones, differentiating them from general Brownian movement. At times, after introduction of movement from the other set, the original set coalesces around it, concentrating and eventually becoming more disparate again; possible coincidence or evidence of reactive interaction. While this is potentially a very interesting hypothesis to investigate, it is very difficult to interpret.

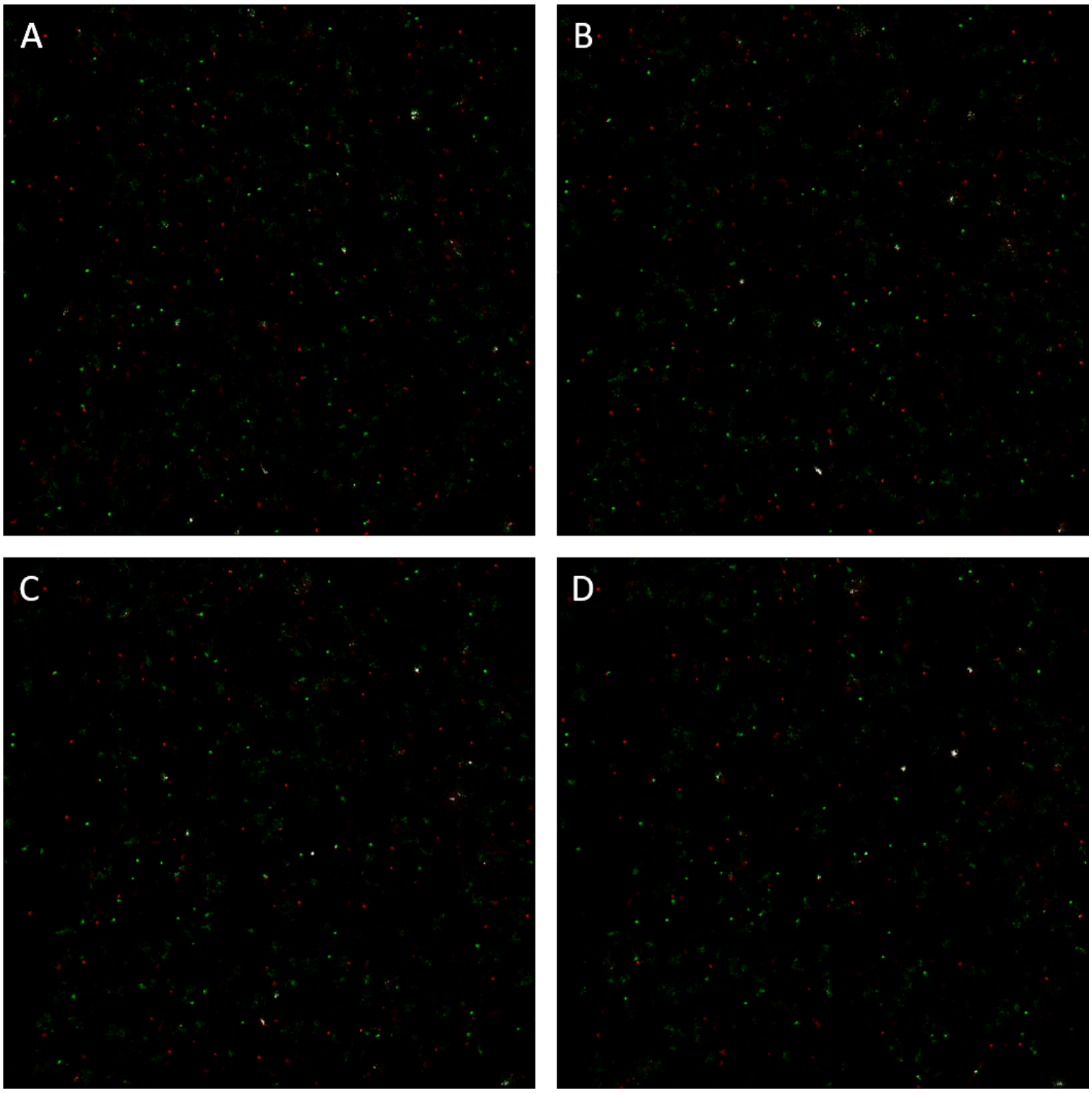



Figure 5.23: Overlaid heatmaps for the C1 (red) and C2 (green) movement generated in the 100 (A), 200 (B), 300 (C) and 400 (D)time phases for GPCRShiver hot-zone subpopulations, clear while is used to colourize areas where both channels closely co-localize

### Cancer net on GPCR and G protein TC641 C1

We were again able to transfer datasets across filtering nets, applying the cancer trained track complexity net to the TC641 C1 data set (Figure 5.24). Three sub-populations were identified from the starting 4,776 tracks, 2,812 CancerTurning classified, 1,953 CancerDirect and 11 fraction. The third was so small a set that tracks were barely visible and therefore have been ignored.

Hot-zones are still present across both visible sets but to varying degrees. The vast majority of motion seems to have fallen into the CancerTurning category; some general movement is present in both with clearer proportions of movement in the CancerDirect set confined to hot-zones. Overlaying CancerTurning and CancerDirect sets supports ubiquity of many hot-zones particularly hot-zones across sub-populations. However, some of the smaller hot-zones seem to be dominantly CancerDirect/green. General movement across the mixed set representation is dominantly red CancerTurning set tracks.

### Cancer net on GPCR and G protein TC641 C2

When applied to the TC641 C2 set, *cancer net* filtering creates a similar CancerTurning set dominance although less pronounced (Figure 5.25). 6,601 starting tracks are separated into 3,773 CancerTurning, 2,409 CancerDirect and 419 fraction sub-populations, although more present this time the movement heatmap for fraction movement was negligible enough to again be ignored.

Clustered general movement is more apparent in the C2 CancerDirect set but with little to no differentiation of hot-zones across sets. Overlaying both sets seems to support this assertion, while a few hot-zones are dominated by movement in either of the sets there are very few present in only one of them. It may be by removal of track heat there is greater general movement clarity in the CancerTurning set over the unfiltered dataset; a form of noise reduction via removal of low confidence tracks.

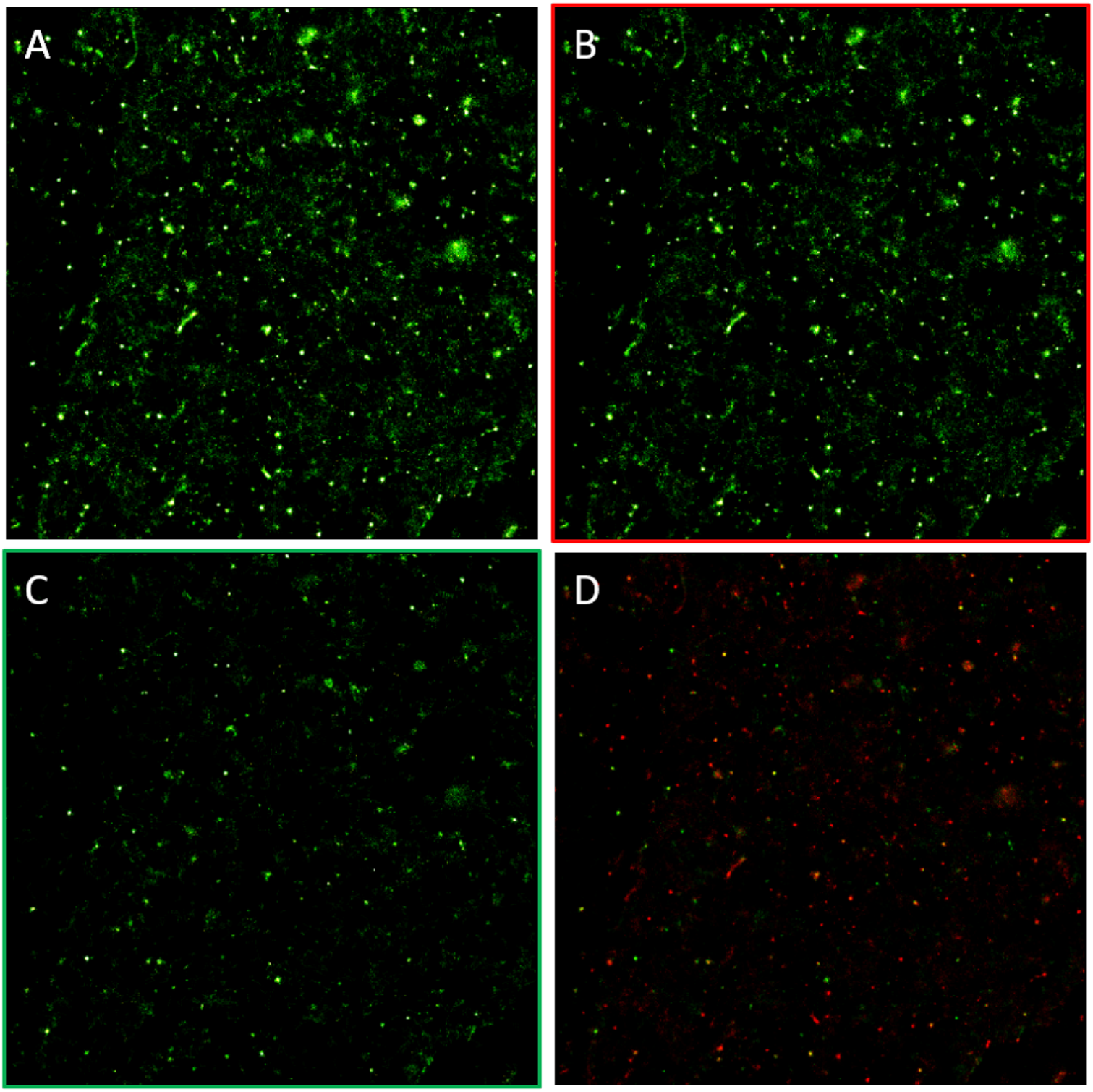



Figure 5.24: GPCR and G protein TC641 C1 data set (A), sub division to red CancerTurning (B) and green CancerDirect (C) sub-populations with a consequent sub-population overlay (D)

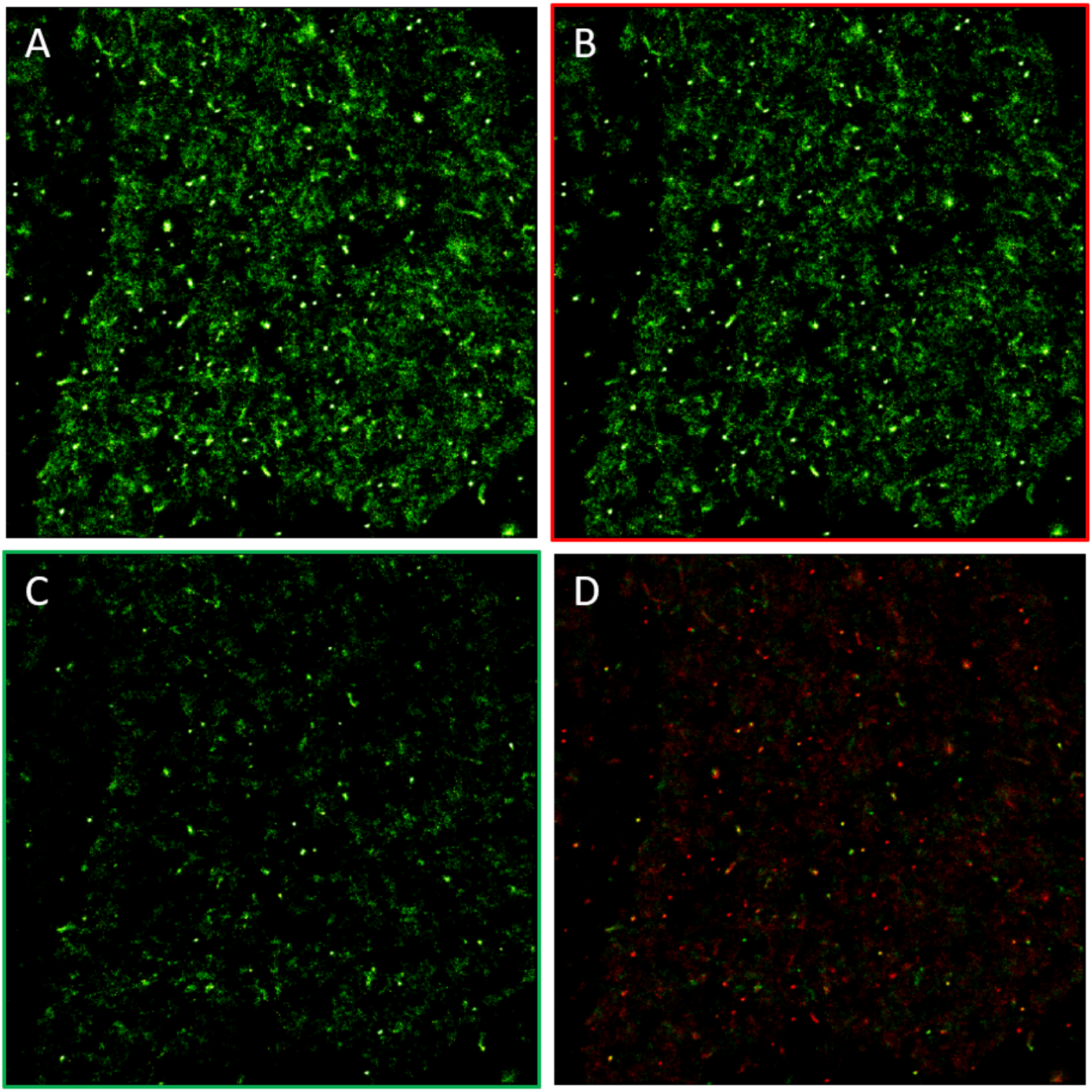



Figure 5.25: GPCR and G protein TC641 C2 data set (A), sub division to red CancerTurning (B) and green CancerDirect (C) sub populations with a consequent sub-population overlay (D)

## 5.4 Discussion

### 5.4.1 Cancer

#### Representative model and real result comparison

Trained net application creates some interesting results in classifying real-world sets against representative model heatmaps. We have developed models more representative than random data generation. CNN comparison primarily suggests the models with combinations of emergent movement patterns are most similar to those in real-world heatmaps; lattice paths and forging are creating an identifiable combined pattern that differentiates from other representations (Figures 5.2, 5.3). Interestingly, path-forging patterns alone are more highly associated than background movement, suggesting that path-forging doesn't detract from model similarity. An emergent combined behaviour of lattice path and forging detracts from the long set similarity but strengthens comparison with short cancer sets. However, path-forging and lattice path following patterns may be classified strongly as a result of stark recognizable visual patterns.

In isolation, simple lattice path results are also visually distinct and do not share similar confidence levels as forging. The combined model is drawing the same lattice comparisons but then taking them a step further. Equally, if such distinct patterns were not present in the real-world set, it would be a strong indicator to move away from such classification. We may be observing a split recognising normal movement, but background and path forging classes are clearly visually distinct so it is unlikely. In general, the real-world to model result classification test strongly supports the presence of both lattice paths of least resistance and exploitative path forging behaviours. The greater association of short to lattice and long to forging very tentatively supports our hypothesis that the cancer cells actively modify and drive change within their environment, generating the movement paths we observe. Such observations match the cancer cells moving through confined areas noted by Paul et al [140], as well as the degrading new paths through surrounding structures and least resistance selection discussed by Deisboeck, Jiao and Tourquato et al [92, 80].

### Breakdown by individual movement track behaviour

By applying the track separation and classification method to our previous short and longer cancer datasets we have further differentiated behavioural patterns within observed strand phenomena. Interestingly, we aimed to differentiate patterns based upon turn number, but a more complex and quantitatively difficult differentiation occurred, arguably more desirable as the CancerTurning class represents more undirected movement and CancerDirect more directed.

Simpler, less deviant tracks seem to make up the core of branches with CancerTurning tracks occurring throughout or as part of possible path forging areas (Figures 5.9,5.10,5.11,5.12). This suggests that cancer cells continue forward more easily in permissive core areas (Figure 5.13), these permissive cores are then examples of low confinement and possible least resistance movement [42, 71, 90, 140]. Strand expansion appears to be composed of more complex tracks of general movement at edges, noisy movement possibly indicative of pushing into these less permissive areas, possibly driven by chemotoaxis and nutrient competition as suggested in the literature [45, 122]. This is supported by the broadening of strands mainly in representations of CancerTurning track sub-populations, it's also possible this complexity is population member collision or splitting interaction interrupting forward motion of cells moving in less permissive areas over time (Figure 5.14).

### Summary

While we can't precisely determine how the trained nets are classifying given images, we can make quantified observations of results and combine them with the hypotheses developed thus far. The real-world classification by model results test supports close association with both path-forging and least resistance paths in combination, suggesting both behaviours of cells are important to environmental interaction. We have previously identified that the longer cancer set may be indicative of an earlier environmental state. This is further supported by short sets being more similar to lattice path models, path forging interaction patterns develop over time and become more pronounced in tightly constructed strand like movement so later data sets are closer to strong lattice than mid generation models.

Even without clear single entity track pattern differentiation, there are single-track results that can be extrapolated to pursue in new models for iterative improvement. We could further develop and compare activity over time; observe if initial coalescence is simply lattice based and moves to general exploitation or vice versa. Further development would also allow comparison of separation based upon more track types or a specifically developed model structure.

### 5.4.2 GPCR and G proteins

#### Representative model and real-world data comparison

By comparing C1 and C2 model similarity results a clear trend appears: attractive areas display the strongest signal for classification and valid representation (Figures 5.4, 5.5,5.7,5.8). Representative of the target hot-zone phenomena [9], attractive areas provide some validation for attractive explanatory models. Brownian motion appears to create strong result deviance across C1 and C2 sets, possibly related to differences in cell surface area and therefore overall dominance and area for general motion. Standard deviation in general seems reasonable, there are some outlying results but not enough to make classification confidence completely stochastic and thereby irrelevant (Figures 5.6,5.9,5.17).

Background classification being stronger for the C2 set might be expected since heatmaps were defined as more movement heavy in previous analysis (3). A higher general deviance for the C2 set however suggests lower general applicability, possibly due to models generated on starting C1 sets then applied across. Such a bias trend suggests reasonable fitting and representation in our initial model definition and comparison with C1 but less representation when comparing with C2. Equally, as net validation success increases, we need to remain aware of over-fitting to training data and model representations.

#### Breakdown by tracks and behaviour type

Previously much of our analysis and discussion centred on hot-zones, application of the single-track classification pipeline therefore focused on attempting to isolate the phenomena. We successfully separated and created a hot-zone dominant model for both C1 and C2 set with neural

net automated classification (Figures 5.18,5.19,5.20,5.21). Additionally, the developed net recognises pre-hot-zone movement and places it in the GPCRBrown subgroup. A potentially very useful serendipitous discovery, it allows us to clearly visualise behaviour before and after hot-zone interaction, important for observing environmental transience discussed in literature [9]. One of the larger hot-zones is of particular interest, previously indicative of an area becoming less confined as time progressed. After filtering, the constituent tracks are still dominated by a compression pattern, the movement is the same but with a larger permissive area to move in. Motility may not always increase with time, boundaries may change but still regulate internal movement dynamics..

Generating turn diagrams for both the C1 and C2 full sets along with their new filtered subsets allows us to evaluate rear-turn bias relative to hot-zones (Figure 5.22). As we are filtering on the phenomena that generates the pattern, rear-turn preference differs across sub populations. Rear bias across all subsets could be the prevalence of hot-zones or difficulty differentiating entirely but the increase with the GPCRShiver set shows a clear pattern and possible causal connection between hot-zones and rear biased movement.

There was a large gap in track complexity between GPCRFrag and GPCRBrown training sets, edge cases may allocate to GPCRFrag for C1 but encountered less cases for C2. It presents an opportunity for future improvement via new more defined classes creating clear subset separation.

### Summary

For the GPCR and G protein sets, our target visual pattern of movement hot-zones is recurrent across a heatmaps with often individual level effects and causes suggested by Calebiro et al [6]. Our model similarity application seems to support differentiation from a purely deflection driven narrative, background motion alone consistently generating closer classification results. Attractive zones alone provide the closest classified similarity, only stationary sub-population models perform better in the C1 set. Attractive zones may better represent the pickets in a fence and picket [56, 58] explanation than the confining cytoskeletal fence further highlighted by Calebiro et al [9]

Separation on an individual basis with our trained neural net has also been more effective. We further isolated and affirmed that hot-zones have a strong effect on the observed cross population rear turn bias. Our net also allows easier identification of pre and post hot-zone interaction tracks. Differentiating between hot-zone types is still however difficult as small and large hot-zone tracks can be classified in GPCRShiver.

### 5.4.3 Transference

Functionally, we were able to apply GPCR net filtering to the cancer sets (Figures 5.15) and vice versa (Figures 5.24,5.25); computational and development cost was low, a couple of interesting interactions were identified.

The main difference between Brownian and random path generation is turn rate and variable jump distance. Therefore, similarity between long term random tracks used to train the CancerTurning class and distance modified GPCRBrown class is not inexplicable. With the hot-zone capture class so clearly defined, the third class when applied to the cancer sets becomes a catchment case consisting of all remaining tracks. Distribution among classes of tracks with transference might be distribution based upon similarity and then elimination into a third set depending upon class definition permissiveness. An inapplicable class with low permissiveness of pattern becomes similar to the fraction filter when applied to GPCR and G protein, too small for visualisation. Future work should generate a final very permissive class type or define a range of acceptable confidence; when filtering results a lack of class confidence could drop into a final 'undefined' class.

The CancerFrag set in the cancer net may be an example where requirements were not met; the net may be unable to differentiate between GPCRFrag, GPCRBrown and GPCRShiver or differentiation between short term tracks created by two different tracking algorithms. We may improve separation and visualisation via combination of cancer and GPCR track filtering. More classes do however lead to higher training complexity.

By cross applying classification nets to different sets we gained some but limited new insight,. However, designed filtering is more effective than simple random application of nets. Both comparative applications function, tailored nets are more effective at sifting and differentiation but with some external applicability.

### 5.4.4 Conclusion

In this chapter we aimed to solve two problems, quantification of representative model to real-world heatmap comparison, and effective population filtering based upon movement profile. We were able to apply CNNs to classify images, heatmaps and tracks based upon training data generated by the framework models.

The cancer set comparison further highlighted a hybrid, least resistance lattice and path forging model design with similarity to the real-world system, analogous to previously noted possible least resistance [42, 71, 90, 140] and stromal degradation behaviours [92, 80]. Track separation was able to create subsets representing complex, simple and fragmented cancer cell movement patterns that were then overlaid to highlight strand makeup. Communal general movement areas seemingly composed of more simple movement than complex strand movement; suggesting the use of strands for travel.

In GPCR to model comparison, deflective representative models representing cytoskeletal confinement [6] performed poorly and again attractive models took precedence in similarity comparison, possibly highlighting the effects of transmembrane protein pickets descussed by Fujiwara et al [56, 58]. We were also able to isolate predominantly hot-zone tracks with confinement based population filtering. Later time phasing suggested that the more general tracks in the GPCR-Shiver set may also belong to larger less distinct hot-zones. The separation enabled us to observe hot-zone coalescence and dispersal along with morphology more clearly.

A larger breadth of investigation into *in vitro* to model comparative CNN fitting would have been valuable and is necessary for further validation of the approach. Similarly, track classifica-

tion would have benefited from more varieties of track class, to differentiate subsets. An internal framework tool for RGB overlay would also allow observation of more than three track subsets simultaneously. When comparing model images to real-world examples for similarity a wider range of training cases could be used. Ensuring CNNs are recognising geographical features not splitting on brightness or other less complex metrics is important. Model design and iteration could be significantly improved through automation, driven by the ability to now quantify model and real-world heatmap comparison.

**Proposed automated framework**

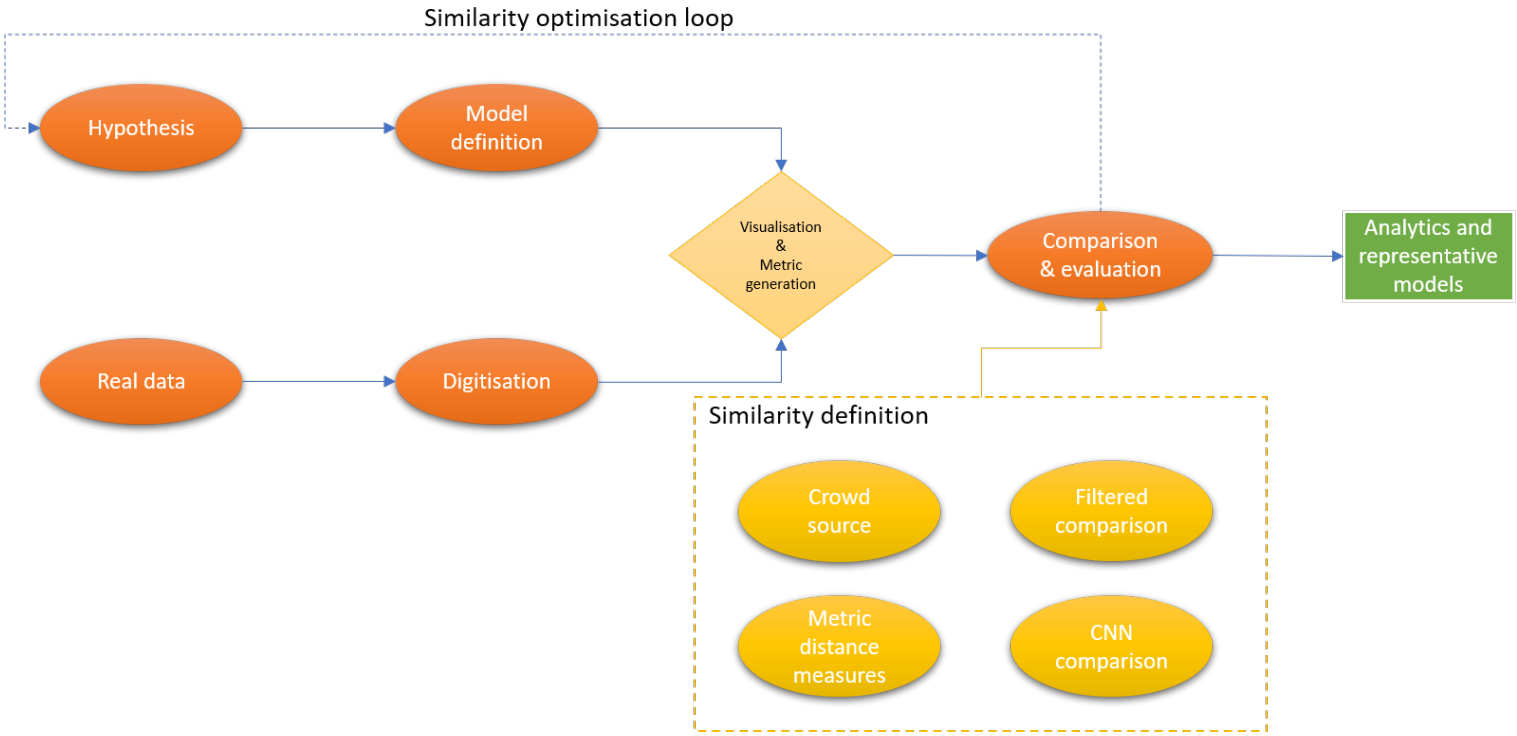



Figure 5.26: Beginning with real tracked data as input, the final proposed automation loop develops representative models through comparison with the real-world results; optimising several possible similarity values between *in vitro* and model generated data to automatically generate new hypothetical insight and re-evaluate to output relevant conclusions and representation iteration.

Possible over- and under-fitting make full automation unlikely. However, the proposed automated framework is a viable semi-supervised approach to initial representative model generation (Figure 5.26). We retain the digitized real-world and representative model sides of the framework but further define the methods for similarity evaluation. It represents a progression from our initial approach to a more defined workflow for multi problem generalised investigative automation

and population movement analysis. In this chapter we have practically evaluated the applicability of a novel, new similarity measure, to be added to those already discussed and implemented in combination with a POM approach.

A great deal of time is spent initially discovering relevant variables for result generation. Similarly, refining ranges to generate over time trends can become expensive. While we can apply some linear regression approaches to the latter problem, initial application of a fully automated loop also substantially cuts back on model development time. An initial implementation of a brute force trend fitting model generation and evaluation loop was already implemented, shortening and improving model definition across chapters.

To automate the comparison segment of model and hypothesis iteration, we need to define similarity; iteration is the increase of this similarity measure. Our trend fitting algorithm evaluates the given quantitative metric with simple deviance from the real set as its similarity measure. Model similarity and single-track classification nets could be used to convert images to metric distance; confidence of comparison between incrementally changed model sets. Single track classification might leverage automatic clustering algorithms to group real-world tracks into classes, a similar model should then be composed of similar sized and distributed subsets.

Unfortunately we were unable to generate a proof of concept for such a step in available time; if computational cost is high it may have been more applicable for late stage model training. An alternative crowd-sourced similarity module was designed and implemented as part of the project, focusing again on visual patterns; citizen-scientists were unfortunately difficult to come by and therefore the contributions were not numerous enough for presentation of any results. It would be a scalable method for pattern classification or holistic comparison. Once similarity is defined, optimisation can become automated using adversarial training and multi objective optimisation to drive emergent model trend generation.

# 6 Discussion

## 6.1 Summary

This work centered on improving our understanding of real-world, complex, dynamic, and spatial biological systems through the analysis of *in vitro* population's movement. With our novel framework, we showed that interactions between population members and their environment can be observed indirectly. By visualising movement we can characterise populations of tracks from real world data as patterns representing possible behaviours (Chapter 1). We examined two types of biological systems: first cancer cells (Chapter 2) and then G protein-coupled receptors (GPCRs) (Chapter 3). Each distinct system evolves over time, leading to emergent patterns of behaviour. The patterns we observed were spatially correlated. We developed tools to isolate and visualise movement choices in relation to position and time phase, further improving interpretation (Chapter 4). Finally, to improve the robustness of population separation as well as quantify the comparison between model and real-world sets, we developed prototype convolutional neural net pipelines (CNN) (Chapter 5).

Our overall objective was to develop and apply a modelling framework to drive understanding for the chosen biological systems. Therefore, our aims were *first*: enable comparison of the selected system with knowledge found in literature, supporting or expanding on existing hypotheses; *second*: to identify new plausible hypotheses. To accommodate the differences between the systems, we also aimed to develop a generalised holistic approach that could be applied to other biological systems.

More specifically, we aimed to:

1. Identify, design and develop our modelling framework for biological systems, first focusing on cancer at a cell level with individual free movement (Chapter 2)

2. Being able to examine the spatial relationships between GPCR and G protein to understand how their environment affects patterns of movement and behaviour (Chapter 3)

3. Improve our framework to enable the characterisation of directional trends within micro-environmental patterns, their construction and morphology over time and apply to both systems(Chapter 4)

4. Finally, use artificial neural networks to develop novel workflow; classifying identifiable visual patterns for model/system comparison and filtering populations into sub sets on track morphology (Chapter 5)

We began with video of dynamic biological systems as the main data source [48, 6]. Tracks were extracted from these videos as a set of entity positions, representing population movement in a given environment over time. A framework was developed to take sets of population tracks and digitise them in a unifying representation, then we could design representative models through iterative comparison with the digitized real-world data. We reduced the risk of introduced artificial bias by applying a pattern oriented modelling (POM) approach and by using common visualisation and metric gathering tools throughout.

## 6.2 Cancer cell movement - original design

Cancer is one of the leading causes of death worldwide with wide reaching effects that are on the rise [1]. Improved understanding of the underlying metastatic mechanisms [120, 121, 45, 122] could lead to improved treatment, early diagnosis, prevention and monitoring [123]. Therefore, we developed our initial framework design for a cancer cell biological system with a focus on movement (Chapter 2). We were given access to *in vitro* non-small lung cancer cell videos, 5 of which were of sufficient length for tracking and analysis [48].

To analyse this system, we first developed tools for visualisation of cells and gathering of movement and population data over time. Movement information was then compiled into heatmap meta representations to display the patterns that each population generated over time. Three key components were identified in the dataset:

- Directed forward focused movement

- Adherence to strand-like paths over time
- Possible sub-populations or predominant behavioural states

We next moved to the definition and generation of representative models within the framework, in order to unify tools for data gathering and representation. Representative interactions were implemented to create baseline movement models, comparing with forward movement, lattice following and path forging for striated strand like paths and entity definition with variance to represent multiple sub-populations. We found that forward biased movement alone did not create the strand like patterns present in real-world sets. Attractive lattice following created clear strands of movement but removed general undirected patterns, path forging created some strands but with more general movement than observed. A final lattice following and path forging model was compelling and suggested that real-world behaviours likely consisted of similar phenomena. The application of our generalised framework for real-world systems and representative model design provided interesting insights into the cancer dataset.

We were able to generate representative models with comparable results, also identifying limitations: limitation in size and timescale of the dataset, difficulty to correlate the large number of metrics for spatial movement patterns. Since numerical trends were not necessarily representative of the phenomena, we focused upon visual comparison and exploratory analysis. The availability of a very different system and corresponding dataset then allowed us to develop our approach further.

## 6.3 Modelling G protein coupled receptor and G protein population movement and interaction

To further develop the framework and extend its capabilities, we next studied G protein coupled receptors (GPCR) and G protein movement (Chapter 3). We were able to improve our range of applications and improve tools for application to the original cancer sets. This second system: GPCR and G protein, a large and diverse group of cell surface receptors, was represented in 10 paired data sets, 10 GPCR and 10 G protein, one of each for every recorded system within a different *in vitro* cell. They facilitate communication of extracellular behavioural changes, receptors interact with proteins in the plasma membrane and react when co-localizing to send signals into a cell [6, 9, 7, 8]. Studying GPCR and G protein behaviour over time allows us to further understand how a cell identifies, controls and apprehends its environment.

Again, we first generated movement and metric data for the new data; analyzing the generated patterns. Movement within the available data sets was characterised by:

- Predominantly Brownian motion
- Entity catchment zones referred to as 'hot-zones'
- Protein receptor co localisation or differentiation

Developing new model definition tools we added possibly representative phenomena for the GPCR set. Initial implementation used Brownian motion and population change to create a baseline background model. We then implemented deflective barriers and attractive zones to compare and replicate representative patterns. Deflective barriers did not create patterns similar to those in real-world sets, attractive zones were more but not completely representative. To further explain the variable morphology of observed hot-zones in the dataset, we developed static or shivering sub-populations. We found that a relatively small shivering sub population could: generate patterns representative of some small hot-zones, create a rear turn bias for the overall population and that not all small zones consisted of a single entity. Again, we were able to identify emergent visual patterns and produce representative models, highlighting possible hypothesis development or questions through comparison.

It was difficult to correlate numerical and visual patterns, quantifying comparative analysis. A pronounced and unexpected rear turn bias was identified, however, localising the bias was difficult with existing tools. Therefore, further tools for spatial representation of directional preference were developed. Populations were also separated by behaviour and time into data subsets for more targeted observation.

## 6.4 An expanded micro-environmental view: methods for further pattern identification

The first two applications highlighted the need for greater information relating to spatial complexity and directional selection of population members. Therefore, new tools such as directional movement heatmaps were implemented to overlay entity choice with location information (Chapter 4). Time phased data was also collected to isolate the effects of movement upon patterns

within specific time frames. A population filtering tool was also investigated using the numerical data our framework generated. All new tools could be applied to the cancer and GPCR and G protein systems via the frameworks backward compatibility.

We identified that cancer sets continued with directed forward movement within strands and through areas of less cohesive general movement. Time phases suggested that strands were initially developed and then characterised by consistent, consequent use for travel, possibly representing exploratory and then exploitative behaviour. For the GPCR and G protein sets, we observed that rear-bias was predominantly within hot-zones, explaining the overall population bias toward rear turns(Chapter 3). Time phases also suggested that a restrictive barrier phenomena could be present and creating hot-zones; generating an effect, but not consistently, to create observed patterns within differing time frames.

Separating sets into sub-populations based upon current quantified metrics, such as movement distance or turn preference, had very limited effect; attempting to isolate hot-zones or strands. The primary remaining limitations were quantified comparison of visual patterns and filtering of populations, to isolate the behaviours that generate them. Therefore, we proposed to use an artificial intelligence (AI) approach in the form of *Convolutional Neural Nets* (CNN), trained for classification as a solution.

## 6.5 Artificial neural nets for movement pattern classification

We developed two distinct artificial neural net pipelines using CNNs: one for comparison of real and model derived heatmaps and a second separating and filtering populations (Chapter 5). We were able to train CNNs to differentiate between each model type. We then present the trained model with our real-world *in vitro* data for quantification of model similarity. The GPCR and G protein set was more complex and required pairwise model training. We generated paired nets that could differentiate two types of model reliably, each was then presented with movement heatmaps from the GPCR and G protein data. Finally, results were compiled into a confidence grid indicating each models similarity to the *in vitro* results.

The second process for track-based separation generated sets of labeled model tracks for use as training data. CNNs were trained to differentiate based upon track complexity for the cancer set and compression for GPCR and G protein data. Once trained the nets were used to sort

all the real-world data tracks into subsets. Each subset was then handed to the framework for visualisation and metric gathering, allowing us to observe and compare with the originating systems. The approach was particularly effective at isolating hot-zones within its own separate set. Finally, we attempted to cross-apply the track separation models GPCR trained to cancer set and vice versa for transference filtering. The resultant sets showed that each net was filtering on different metrics and that some new if limited insight could be gained through transference.

With these initial pipelines we were able to generate similarity metrics between *in vitro* biological systems and representative models entirely based upon visual patterns. Furthermore, the filtered sets displayed clear and very useful visual pattern distinctions, thus allowing further focused analysis on cancer strand as well as GPCR and G protein hot-zone composition and morphology over time. Further, more detailed analysis of the approach could be accomplished with more nuanced CNN designs along with a wider spread of test cases and applicable biological systems in future work.

## 6.6 Biological perspective

We aimed to support or develop existing hypotheses while also identifying new possible interactions relating to our targeted biological systems. As such our approach was exploratory, we mixed quantitative and qualitative investigation to generate possible new insight. Therefore, there are several identifiable possible biologically representative patterns indicative of causal narratives throughout our work.

### 6.6.1 Cancer cell movement

When observing the cancer data sets, movement generated distinct visual patterns that could be broken down into three main behaviours; directed movement, striated strand like cohesiveness and sub-populations active over time. Directed movement may be representative of gradient following and chemotoaxis phenomena discussed in literature [122, 45, 123], strand cohesion representative of least resistance following described by Jiao et al [80] or confined motility by Paul et al [140] and sub populations relating to possible population heterogeneity widely discussed in the literature [80, 116, 117, 118]. We also had videos taken from different time points in environmental interaction with differences in strand pattern progression.

### Directed movement

Directed movement refers to the proclivity for a population member to travel in non random directions. When analysing the cancer cell data set, we observed a very strong forward bias (Chapter 2). We assumed this forward bias was environmentally driven, possibly representative of phenomena in literature such as chemotaxis [122, 45, 123]. Movement heatmaps with localised direction preference indicated that forward biased behaviour continued outside the scope of established strands (Chapter 4) supporting the narrative of localised micro-environment interaction. So, we designed a baseline directed movement representative model with parameters suggested by the real-world observations and a pronounced forward bias (Chapters 2, 4). Heatmap movement distribution suggested that for models with only directed movement, environmental exploration was increased and primarily dependant upon several indirect factors; length of time, travel speed and population size. Purely forward turn biased representative models scored poorly in our comparative CNN, although better than purely random motion (Chapter 5); likely due to the absence of key environmental patterns caused by interactions with phenomena such as nutrient or resistance gradients.

For purely directed movement models environmental exploration could vary and common patterns were identified based upon initial placement, clear strand and environmental development was not observed unlike in the real-world sets. Any single population member under the effects of directed movement can create small narrow paths with some similarity to the strands seen in *in vitro* data, but without the consistency of adherence over time or density of strands created by multiple population members.

### Lattice path and strand patterns

A lattice is an overlaying mesh pattern; when attractive, a lattice overlay may lead to a population congregating to the mesh and creating strands of increased movement over time along their length. Visual strands of movement were common in the cancer data set and therefore likely attractive, representing cancer cell interaction with phenomena such as paths of least resistance [80] or movement restriction [140]. Strands of movement could be caused by a lattice of depressions in their environment offering easier paths of least resistance (Chapter 2). Such strands showed strong forward preferential movement throughout their length (Chapter 4), again, possi-

bly indicative chemotoaxis and gradient following discussed in literature [122, 45, 123]. Phased observation of the longer cancer set showed signs of strands developing over time, from general movement to a more cohesive pattern, suggesting catchment, but also path expansion and development (Chapter 4) possibly similar to stromal degradation in permissive environments [80].

In representative models a high and consistent attachment rate was needed to generate distinct lattice like patterns over time (Chapter 2). Lattice only representations were good at creating strand patterns without reproducible varied morphology. Therefore, we also developed a forging-path design; the more cells had travelled an area, the more attractive it was for future movement, representing gradual degradation of micro-environment developing paths of least resistance similar to those modeled by Jiao et al [80]. Forging models were able to produce strands with more representative morphology and retained less directed general movement areas. However, forged strands lacked clarity and ability to retain patterns when meeting an area of more general movement unlike in real-world sets. Our comparative CNN showed a strong similarity between cancer sets and hybrid following and forging models, when alone simple forging was preferred over only attractive models (Chapter 5). By breaking the cancer tracks down upon track morphology, we were able to identify two further ways of separating strand behaviour; strand following and general exploration via strand divergence (Chapter 5). It is possible that cells broaden strands by diverging to respond to nutrient availability, chemotaxis taking priority over environmental permissibility. We observe a core of strong directed motion along the spine of strands, at the edges, forefront or sides of strands, less directed movement occurred.

#### Sub-populations/behavioural states

By separating a real-world set upon a metric such as turn preference, we could create subsets representing that specific split. We assume they are a separate sub-population, referred to as subsets. Therefore, a subset in this case is a group of cells showing similar behaviours over our observed run. We have identified strong following and expansionist behaviour types possibly reacting to very localised nutrient or confinement changes [80, 140]. We can observe them with sets split by turn preference (Chapter 4)or track complexity (Chapter 5), both generate different patterns, but without complete differentiation; no visual pattern is entirely isolated. Forward motile sets move more and further, less motile sets tend to be static or forging at the edges of strands. Cells might cycle through states over time, adapt or be differentiated at a genetic level

for behavioural preference. One narrative that often occurs in the literature is that of invasive versus proliferative sub-populations [124, 92, 80]. Similarly, such observation is often tied to the ability for cells to degrade surrounding stromal structure, similar to the forging behaviour we identified.

The cancer cells are likely evolving and moving within an environment with lattice paths of least resistance and reasonably even nutrient distribution. Strands were likely created by highly motile cells and thickened by use over time and density shifts via path forging. Unfortunately, this doesn't suggest a solution to the common problem of nutrient starvation driving metastatic exploration versus nutrient availability driving proliferative expansion. However, it may offer further insight into controlling cancerous development and cell behaviour over time.

### 6.6.2 GPCR and G protein movement

For the GPCR and G protein datasets, we focused upon baseline Brownian motion [56, 58], distinct hot-zones[6] and differences between protein and GPCR behaviour as described by Calebiro et al [9, 152]. Both motion and hot-zones were present in G protein and GPCR datasets. Representative models were used to replicate and differentiate background Brownian motion from the hot-zones. Further, the CNN single track filter application we developed, was able to separate general and restricted subsets to improve direct observation over time.

#### Brownian motion

Brownian motion refers to an entity that moves at every available time step, in a random direction and a normally distributed distance; it is expected of GPCR and G protein movement [56, 58]. The framework identifies Brownian motion as an entirely random turn preference; the GPCR and G protein data when visualised produced a pronounced rear turn preference (Chapter 3), indicating some possibly non purely Brownian motion and a potentially interesting interaction.

We generated a representative Brownian motion only model showing that over time a pop-

ulation of baseline GPCR and G protein model entities would spread out with highly random trajectories (Chapter 3). Similar general movement patterns were observed when population size and travel distance were applied from *in vitro* results with Brownian motion. General turn patterns were also spatially similar in directional heatmaps (Chapter 4). However, general spread, areas of movement starvation and movement dense hot-zones were not visible in the baseline purely Brownian model indicating the presence of other interactions in real-world sets as discussed in literature [9]. Also, we generated models with differing initial distribution (Chapter 3). While general spread was reasonably representative the amount of localised movement starvation and hot-zones were not replicated without the inclusion of other environmental interactions.

### Hot-zones

A movement hot-zone is an area of dense entity placement across time; one or many entities moving within a small area will gradually increase the brightness in a movement heatmap. The brighter, or in other words 'hotter', the zone, the more movement has occurred over time. Hot-zone patterns are a key feature of the *fence and picket* model of the plasma membrane, actin-based skeleton *fences* and transmembrane protein *pickets*, fences confining motion [56, 58, 9]. The phenomena is important to GPCR and G protein observation, a hot-zone is believed to be an area of catchment and possible co-localisation of G proteins and GPCRs for communication of information [54, 52, 144, 6].

We found it very difficult to reproduce similar patterns with confinement models representing cytoskeletal restrictive areas alone, in isolation this method of confinement was relatively ineffective. We developed two sub-hypotheses: a small sub-population of highly localised virtually immobile shivering entities representing one of several sub-population types discussed in literature [54] and attractive zones representing transmembrane protein picket catchment [56, 58]. In the case of shivering entities, small hot-zones could be generated and the previously observed reverse bias also showed across population trends (Chapter 3). Shivering provided a possible explanation for small hot-zones and showed that even a small subset could bias turn preference visualisation. Shivering alone did not explain differences in hot-zone size, morphology and changes over time.

Shape and size of hot-zones were not directly tied to rear bias. Directional heatmaps showed that while larger hot-zones were often more permissive and not entirely rear turn dominated, many were not (Chapter 4). By varying parameters for attraction strength, reach and hold, an attractive zone representative model was capable of producing: central permissive movement, varied morphology and surrounding area movement starvation. All key patterns of the GPCR and G protein datasets, CNN's also scored them highly on similarity (Chapter 5).

We were able to use CNN's to separate a majority of hot-zone movement from Brownian via trajectory filtering into subsets for direct observation (Chapter 5). We could then identify some cases of small hot-zones progressing to larger, and then again smaller areas over time within the GPCR and G protein data set 4 possibly representative of transient catchment [152] or conditional change [9]. Movement starved areas also often changed over time; catchment and release of population members is likely a common process. Observed GPCR and G protein populations interact with their environment in a complex manner indicative of multiple competing phenomena. We can examine the differences between GPCR and G protein behaviour; perhaps different catchment types have differing effects depending upon target.

### GPCRs and G protein interaction

Our hypothesis, was that proteins and receptors can co-localise within hot-zones, in line with the literature [6, 9]. We also analysed GPCR and G protein movement behaviour separately. Both GPCR and G protein movement features general Brownian motion, but showed a marked rear bias across the population (Chapter 3). Greater movement spread over time was observed in the G protein sets, but with similar average distance traveled, indicative of possibly greater hot-zone capture for GPCRs. CNNs found increased similarity between representative models with attractive and shivering sub-population representation than simple baseline movement; indicating likely causal phenomena beyond random chance (Chapter 5).

By overlapping the movement heatmaps for both protein and GPCRs we observed that they overlap particularly within hot-zones (Chapters 3,4), co-localisation being important for signal communication [6, 9]. Also, there is some general movement overlap but at a lower rate indicating an amount of expected random co-localisation, hot-zones may aid but do not solely dictate

interaction. We can also observe that overlap occurs in small and large hot-zones. Occasionally one type of entity leaves and a hot-zone can be seen to disintegrate, possibly disassociation of dimer localisation. Interestingly, there is also some indication of the opposite phenomena, i.e. areas of mutual disassociation.

There are areas where GPCR and G protein movement patterns mutually differentiate, i.e. movement starved zones in one population map are filled by the other when overlaid (Chapter 4). Additionally, in the starvation area of some hot-zones there are smaller sub hot-zones of only one of the two populations near by. Both phenomena might be explained by environmental density based differentiation, i.e as the density of an area changes to trap one type, it may restrict the other. A possible competitive measure, effective association may need a balance of available members from each set, but not overabundance.

## 6.7 Conclusion

We successfully developed a framework using trajectory information from the observable biological systems, providing analyses through observation and comparison with representative models. We developed an applied approach with generalised implementation that allowed input of multiple varied data sources and successful subsequent analysis. The modular design also allowed addition of numerous model definition and metric generation tools, without combined computational overhead increases.

Insights into cancer cell behaviour were successfully obtained with a focus upon individual movement over time. We followed by investing the spatial relationships that dictate GPCR and G protein movement pattern emergence; focusing upon hot-zones. Application to a new biological system further improved the framework as a tool set, we specifically developed tools for visualising directional selection within very small micro-environments and pattern development over time. Finally, CNNs were developed and applied in a proof of concept pipeline investigation to generate similarity metrics for model to *in vitro* result comparison and for filtering sets of tracks upon visual profiles. Both cancer cell and GPCR systems were investigated, with interesting results leading to development of novel hypotheses.

Recurrent limitations came from the number of available datasets and the difficulty generating representative numerate analysis for spatially complex patterns. From the successful development and application of our framework, we have generated valuable insight and can suggest several directions for possible future work or development.

In the case of cancer movement, literature and our own analysis suggests cells prefer paths of least resistance when moving, but not the priority order of selection or expansion behaviour; nutrient availability versus ease of movement. As a practical application, such information could support treatment attempting to predict and trap cancerous cells [159]; knowing when cells sense and select direction and upon what criteria could be used to manipulate their behaviour on a population-wide scale. We also identified that future work could improve models through the inclusion of greater population behavioural and morphological heterogeneity [141, 142], often indicative of motility methods [140].

For GPCR and G protein systems, we also observed patterns suggestive of some phenomena such as density variation for GPCR and G protein differentiation, as well as possible co-localisation. Work to define the effect of medium density on GPCR and G protein travel may drive further insights into environmental interaction. Models could also be improved by the inclusion of representations for lipid-protein complexes and nanodomains, affecting population member movement and confinement [9]. Similarly, environmental transience may account for many observations and even the lack of confinement patterns in out restrictive boundary models [9].

Application toward datasets from other biological systems would continue to drive framework development and improve analysis of new and old. More direct improvement of the overall approach is also possible, primarily involving increased automation (Chapter 5). Further formalising model and system comparison pipelines, along with integration into the framework, would allow us to leverage the generated visual patterns. Improvements in those areas would increase accuracy, with both un- and semi-supervised approaches becoming more applicable with evidence of representative outcomes. We do not propose a fully unsupervised approach. However,

being able to automatically eliminate large numbers of prospective hyperparameter combinations, would be advantageous, allowing focus upon broadening exploratory area and analysis with manual iteration.

In conclusion, we have demonstrated the power of our framework to tackle a variety of biological problems. Further extension of the framework capability to both implement more complex systems as well as better quantify their impact would yield significant benefits. Finally, the application of our framework to new biological systems would also result in new tools and improvements that can be applied across other systems, a priori and a posteriori.

# Bibliography


[1] Freddie Bray, Ahmedin Jemal, Lindsey A. Torre, David Forman, and Paolo Vineis. Long-term realism and cost-effectiveness: Primary prevention in combatting cancer and associated inequalities worldwide. *Journal of the National Cancer Institute*, 107(12):273, sep 2015.

[2] R. J. Murphy, P. R. Buenzli, R. E. Baker, and M. J. Simpson. A one-dimensional individual-based mechanical model of cell movement in heterogeneous tissues and its coarse-grained approximation. 475(2227):20180838, 2019-07.

[3] Peter Friedl, Joseph Locker, Erik Sahai, and Jeffrey E. Segall. Classifying collective cancer cell invasion. 14(8):777–783, 2012-08.

[4] Katarina Wolf, Yi I. Wu, Yueying Liu, Jörg Geiger, Eric Tam, Christopher Overall, M. Sharon Stack, and Peter Friedl. Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. 9(8):893–904, 2007-07.

[5] Jan Brábek, Claudia T Mierke, Daniel Rösel, Pavel Veselý, and Ben Fabry. The role of the tissue microenvironment in the regulation of cancer cell motility and invasion. 8(1), 2010-09.

[6] Davide Calebiro and Zsombor Koszegi. The subcellular dynamics of GPCR signaling. *Molecular and Cellular Endocrinology*, 483:24–30, mar 2019.

[7] Kristen L Pierce, Richard T Premont, and Robert J Lefkowitz. Seven-transmembrane receptors. *Nature reviews. Molecular cell biology*, 3:639–650, September 2002.

[8] Robert J. Lefkowitz. Historical review: A brief history and personal retrospective of seven-transmembrane receptors. *Trends in Pharmacological Sciences*, 25(8):413–422, aug 2004.

[9] Davide Calebiro, Zsombor Koszegi, Yann Lanoiselée, Tamara Miljus, and Shannon O'Brien. G protein-coupled receptor-g protein interactions: a single-molecule perspective. 101(3):857–906, 2021-07.

[10] Akihiro Kusumi and Yasushi Sako. Cell surface organization by the membrane skeleton. 8(4):566–574, 1996-08.

[11] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310(5750):pp. 987–991, 2005.

[12] Rui Chen and Wenhua Zeng. Multi-objective optimization in dynamic environment: A review. In *2011 6th International Conference on Computer Science & Education (ICCSE)*. IEEE, aug 2011.

[13] Oliver Giel and Per Kristian Lehre. On the effect of populations in evolutionary multi-objective optimisation. *Evolutionary Computation*, 18(3):335–356, sep 2010.

[14] Ali Masoudi-Nejad, Gholamreza Bidkhori, Saman Hosseini Ashtiani, Ali Najafi, Joseph H. Bozorgmehr, and Edwin Wang. Cancer systems biology and modeling: Microscopic scale and multiscale approaches. *Seminars in Cancer Biology*, 30:69, Feb 2015.

[15] Le Zhang, Yao Xue, Beini Jiang, Costas Strouthos, Zhenfeng Duan, Yukun Wu, Jing Su, and Xiaobo Zhou. Multiscale agent-based modelling of ovarian cancer progression under the stimulation of the stat 3 pathway. *IJDMB*, 9(3):235, 2014.

[16] Robert G. Abbott, Stephanie Forrest, and Kenneth J. Pienta. Simulating the hallmarks of cancer. *Artificial Life*, 12(4):634, Oct 2006.

[17] Lois M. L. Delcambre, Stephen W. Liddle, Oscar Pastor, and Veda C. Storey. A reference framework for conceptual modeling. In *Conceptual Modeling*, pages 27–42. Springer International Publishing, 2018.

[18] Roger J Brooks and Wang Wang. Conceptual modelling and the project process in real simulation projects: a survey of simulation modellers. *Journal of the Operational Research Society*, 66(10):1669–1685, oct 2015.

[19] Stewart Robinson. Conceptual modelling for simulation: Progress and grand challenges. *Journal of Simulation*, 14(1):1–20, may 2019.

[20] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, nov 2002.

[21] Chwee Beng Lee. The interactions between problem solving and conceptual change: System dynamic modelling as a platform for learning. *Computers & Education*, 55(3):1145–1158, nov 2010.

[22] Anastasia Gogi, Antuela A. Tako, and Stewart Robinson. An experimental investigation into the role of simulation models in generating insights. *European Journal of Operational Research*, 249(3):931–944, mar 2016.

[23] Gibin G. Powathil, Maciej Swat, and Mark A.J. Chaplain. Systems oncology: Towards patient-specific treatment regimes informed by multiscale mathematical modelling. *Seminars in Cancer Biology*, Mar 2014.

[24] Benjamin Titz, Kevin R Kozak, and Robert Jeraj. Computational modelling of anti-angiogenic therapies based on multiparametric molecular imaging data. *Phys. Med. Biol.*, 57(19):6101, Sep 2012.

[25] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Stefan Flasche, Billy J Quilty, Nicholas Davies, Yang Liu, Samuel Clifford, Petra Klepac, Mark Jit, Charlie Diamond, Hamish Gibbs, and Kevin van Zandvoort. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496, apr 2020.

[26] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, mar 2002.

[27] Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, may 2012.

[28] Thomas S. Deisboeck, Zhihui Wang, Paul Macklin, and Vittorio Cristini. Multiscale cancer modeling. *Annual Review of Biomedical Engineering*, 13(1):127–155, Aug 2009.

[29] G.E.P. Box. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. Elsevier, 1979.

[30] Le Zhang, Zhihui Wang, Jonathan A. Sagotsky, and Thomas S. Deisboeck. Multiscale agent-based cancer modeling. *J. Math. Biol.*, 58(4-5):545–559, Sep 2008.

[31] Mohammad Hossein Zangooei and Jafar Habibi. Hybrid multiscale modeling and prediction of cancer cell behavior. *PloS one*, 12:e0183810, 2017.

[32] Lucas B. Edelman, James A. Eddy, and Nathan D. Price. In silico models of cancer. *WIREs Syst Biol Med*, 2(4):438–459, Nov 2009.

[33] Craig W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21(4):25–34, aug 1987.

[34] Ludmil B Alexandrov. Understanding the origins of human cancer. *Science (New York, N.Y.)*, 350:1175, December 2015.

[35] Samra Turajlic and Charles Swanton. Metastasis as an evolutionary process. *Science (New York, N.Y.)*, 352:169–175, April 2016.

[36] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, 366(10):883–892, 2012.

[37] Rebecca A. Burrell, Nicholas Mcgranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467), 2013.

[38] Michael R Stratton. Exploring the genomes of cancer cells: progress and promise. *Science (New York, N.Y.)*, 331(6024):1553–1558, 2011.

[39] Joan Massagué. G1 cell-cycle control and cancer. *Nature*, 432(7015):298–306, Nov 2004.

[40] Robert Axelrod, David E. Axelrod, and Kenneth J. Pienta. Evolution of cooperation among tumor cells. *Proc Natl Acad Sci U S A*, 103(36):13474–13479, Sep 2006.

[41] Hermann B. Frieboes, John S. Lowengrub, S. Wise, X. Zheng, Paul Macklin, Elaine L. Bearer, and Vittorio Cristini. Computer simulation of glioma growth and morphology. *NeuroImage*, 37:S59–S70, Jan 2007.

[42] Alexander R. A. Anderson, Katarzyna A. Rejniak, Philip Gerlee, and Vito Quaranta. Microenvironment driven invasion: a multiscale multimodel investigation. *J. Math. Biol.*, 58(4-5):579–624, Oct 2008.

[43] Adrian L. Harris. Hypoxia – a key regulatory factor in tumour growth. *Nature Reviews Cancer*, 2(1):38–47, Jan 2002.

[44] Jun Chen, Kathleen Sprouffske, Qihong Huang, and Carlo C. Maley. Solving the puzzle of metastasis: The evolution of cell migration in neoplasms. *PLoS ONE*, 6(4):e17933, Apr 2011.

[45] Evanthia T. Roussos, John S. Condeelis, and Antonia Patsialou. Chemotaxis in cancer. *Nature Reviews Cancer*, 11(8):573–587, jul 2011.

[46] D. O. Croci and M. Salatino. Tumor immune escape mechanisms that operate during metastasis. *Curr Pharm Biotechnol*, 12(11):1923–1936, Nov 2011.

[47] Neele Drobnitzky. *Novel therapeutic strategies for targeting EGFR mutated non-small cell lung cancer*. PhD thesis, 2017.

[48] Jacopo Credi. Collective behaviour and stigmergy in populations of cancer cells, 2015.

[49] Alexander S. Hauser, Misty M. Attwood, Mathias Rask-Andersen, Helgi B. Schiöth, and David E. Gloriam. Trends in GPCR drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery*, 16(12):829–842, oct 2017.

[50] Dustin E Bosch and David P Siderovski. G protein signaling in the parasite entamoeba histolytica. *Experimental & Molecular Medicine*, 45(3):e15–e15, mar 2013.

[51] Davide Calebiro and Titiwat Sungkaworn. Single-molecule imaging of GPCR interactions. *Trends in Pharmacological Sciences*, 39(2):109–122, feb 2018.

[52] Peter Hein, Monika Frank, Carsten Hoffmann, Martin J Lohse, and Moritz Bünemann. Dynamics of receptor/g protein coupling in living cells. *The EMBO Journal*, 24(23):4106–4114, nov 2005.

[53] Céline Galés, Joost J J Van Durm, Stéphane Schaak, Stéphanie Pontier, Yann Percherancier, Martin Audet, Hervé Paris, and Michel Bouvier. Probing the activation-promoted structural rearrangements in preassembled receptor–g protein complexes. *Nature Structural & Molecular Biology*, 13(9):778–786, aug 2006.

[54] Titiwat Sungkaworn, Marie-Lise Jobin, Krzysztof Burnecki, Aleksander Weron, Martin J. Lohse, and Davide Calebiro. Single-molecule imaging reveals receptor–g protein interactions at cell surface hot spots. *Nature*, 550(7677):543–547, oct 2017.

[55] Kenichi Suzuki, Ken Ritchie, Eriko Kajikawa, Takahiro Fujiwara, and Akihiro Kusumi. Rapid hop diffusion of a g-protein-coupled receptor in the plasma membrane as revealed by single-molecule techniques. *Biophysical Journal*, 88(5):3659–3680, may 2005.

[56] Takahiro Fujiwara, Ken Ritchie, Hideji Murakoshi, Ken Jacobson, and Akihiro Kusumi. Phospholipids undergo hop diffusion in compartmentalized cell membrane. *Journal of Cell Biology*, 157(6):1071–1082, jun 2002.

[57] Nao Hiramoto-Yamaki, Kenji A. K. Tanaka, Kenichi G. N. Suzuki, Koichiro M. Hirosawa, Manami S. H. Miyahara, Ziya Kalay, Koichiro Tanaka, Rinshi S. Kasai, Akihiro Kusumi, and Takahiro K. Fujiwara. Ultrafast diffusion of a fluorescent cholesterol analog in compartmentalized plasma membranes. *Traffic*, 15(6):583–612, mar 2014.

[58] Akihiro Kusumi, Kenichi G.N. Suzuki, Rinshi S. Kasai, Ken Ritchie, and Takahiro K. Fujiwara. Hierarchical mesoscale domain organization of the plasma membrane. *Trends in Biochemical Sciences*, 36(11):604–615, nov 2011.

[59] Michael J. Saxton. Chemically limited reactions on a percolation cluster. *The Journal of Chemical Physics*, 116(1):203, 2002.

[60] Ton De Jong and Wouter R. Van Joolingen. Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2):179–201, jun 1998.

[61] Thomas S Deisboeck, Le Zhang, Jeongah Yoon, and Jose Costa. In silico cancer modeling: is it ready for prime time? *Nat Clin Prac Oncol*, 6(1):34–42, Oct 2009.

[62] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, jun 2012.

[63] Peter Goldsborough. A tour of tensorflow. *arXiv*, 2016.

[64] V. Grimm and S. F. Railsback. Pattern-oriented modelling: a 'multi-scope' for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):298–310, dec 2011.

[65] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

[66] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, dec 1986.

[67] Gabriele Lohmann, Kerstin Erfurth, Karsten Müller, and Robert Turner. Critical comments on dynamic causal modelling. *NeuroImage*, 59(3):2322–2329, feb 2012.

[68] M. San Miguel, J. H. Johnson, J. Kertesz, K. Kaski, A. Díaz-Guilera, R. S. MacKay, V. Loreto, P. Érdi, and D. Helbing. Challenges in complex systems science. *The European Physical Journal Special Topics*, 214(1):245–271, nov 2012.

[69] Gang Luo. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1), may 2016.

[70] Eleftheria Tzamali, Georgios Grekas, Konstantinos Marias, and Vangelis Sakkalis. Exploring the competition between proliferative and invasive cancer phenotypes in a continuous spatial model. *PLoS ONE*, 9(8):e103191, Aug 2014.

[71] Thomas S. Deisboeck and Iain D. Couzin. Collective behavior in cancer cell populations. *BioEssays*, 31(2):190–197, Feb 2009.

[72] Thomas S Deisboeck. Personalizing medicine: a systems biology perspective. *Molecular Systems Biology*, 5, Mar 2009.

[73] Andriy Burkov. *The Hundred-Page Machine Learning Book*. 2019.

[74] Yuzhen Niu, Lingling Ke, and Wenzhong Guo. Evaluation of visual saliency analysis algorithms in noisy images. *Machine Vision and Applications*, 27(6):915–927, jun 2016.

[75] Dave Tahmoush. Image similarity to improve the classification of breast cancer images. *Algorithms*, 2(4):1503–1525, dec 2009.

[76] Metty Mustikasari, Sarifuddin Madenda, Eri Prasetyo, Djati Kerami, and Suryadi Harmanto. Content based image retrieval using local color histogram. *International Journal of Engineering Research*, 3(8):507–511, aug 2014.

[77] Nicholas A. Cilfone, Denise E. Kirschner, and Jennifer J. Linderman. Strategies for efficient numerical implementation of hybrid multi-scale agent-based models to describe biological systems. *Cellular and Molecular Bioengineering*, 8(1):119–136, nov 2014.

[78] A. Braun, S.R. Musse, L.P.L. de Oliveira, and B.E.J. Bodmann. Modeling individual behaviors in crowd simulation. In *Proceedings 11th IEEE International Workshop on Program Comprehension*. IEEE Comput. Soc.

[79] Mathematical Games. The fantastic combinations of john conway's new solitaire game "life" by martin gardner. *Scientific American*, 223:120–123, 1970.

[80] Yang Jiao and Salvatore Torquato. Emergent behaviors from a cellular automaton model for invasive tumor growth in heterogeneous microenvironments. *PLoS Comput Biol*, 7(12):e1002314, Dec 2011.

[81] A. R. Kansal, S. Torquato, G.R. Harsh, E. A. Chiocca, and T. S. Deisboeck. Simulated brain tumor growth dynamics using a three-dimensional cellular automaton. *Journal of Theoretical Biology*, 203(4):367–382, Apr 2000.

[82] Ángel Monteagudo and José Santos. Treatment analysis in a cancer stem cell context using a tumor growth model based on cellular automata. *PLoS ONE*, 10(7):e0132306, Jul 2015.

[83] Katarzyna A. Rejniak and Alexander R. A. Anderson. Hybrid models of tumor growth. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(1):115–125, Jul 2010.

[84] D.C. Walker, N.T. Georgopoulos, and J. Southgate. Anti-social cells: Predicting the influence of e-cadherin loss on the growth of epithelial cell populations. *Journal of Theoretical Biology*, 262(3):425–440, Feb 2010.

[85] Johannes Pollmächer and Marc Thilo Figge. Agent-based model of human alveoli predicts chemotactic signaling by epithelial cells during early aspergillus fumigatus infection. *PLoS ONE*, 9(10):e111630, Oct 2014.

[86] Nabila Kazmi, M.A. Hossain, and Roger M. Phillips. A hybrid cellular automaton model of solid tumor growth and bioreductive drug transport. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(6):1595–1606.

[87] Michael Rubenstein, Christian Ahler, Nick Hoff, Adrian Cabrera, and Radhika Nagpal. Kilobot: A low cost robot with scalable operations designed for collective behaviors. *Robotics and Autonomous Systems*, 62(7):966–975, jul 2014.

[88] M. Rubenstein, A. Cornejo, and R. Nagpal. Programmable self-assembly in a thousand-robot swarm. *Science*, 345(6198):795–799, aug 2014.

[89] Erik Meijering, Oleh Dzyubachyk, Ihor Smal, and Wiggert A. van Cappellen. Tracking in cell and developmental biology. 20(8):894–902, 2009-10.

[90] H. P. Zhang, A. Beer, E.-L. Florin, and H. L. Swinney. Collective motion and density fluctuations in bacterial colonies. *Proceedings of the National Academy of Sciences*, 107(31):13626–13630, jul 2010.

[91] V. Narayan, S. Ramaswamy, and N. Menon. Long-lived giant number fluctuations in a swarming granular nematic. *Science*, 317(5834):105–108, jul 2007.

[92] T. S. Deisboeck, M. E. Berens, A. R. Kansal, S. Torquato, A. O. Stemmer-Rachamimov, and E. A. Chiocca. Pattern of self-organization in tumour systems: complex growth dynamics in a novel brain tumour spheroid model. *Cell Prolif*, 34(2):115–134, Apr 2001.

[93] David R. Sherwood and Julie Plastino. Invading, leading and navigating cells inCaenorhabditis elegans: Insights into cell movementin vivo. *Genetics*, 208(1):53–78, jan 2018.

[94] Keith C. Clarke. Cellular automata and agent-based models. *Handbook of Regional Science*, pages 1217–1233, Jul 2013.

[95] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520*, 2018.

[96] David Johnson, Steve McKeever, Georgios Stamatakos, Dimitra Dionysiou, Norbert Graf, Vangelis Sakkalis, Konstantinos Marias, Zhihui Wang, Thomas Deisboeck, David Johnson, and et al. Dealing with diversity in computational cancer modeling. *Cancer Informatics*, page 115, May 2013.

[97] Sandra K. Hanneman. Design, analysis, and interpretation of method-comparison studies. *AACN Advanced Critical Care*, 19(2):223–234, 4 2008.

[98] J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, apr 1999.

[99] Junhwan Jeon, Vito Quaranta, and Peter T. Cummings. An off-lattice hybrid discrete-continuum model of tumor growth and invasion. *Biophysical Journal*, 98(1):37–47, Jan 2010.

[100] M.A. Al-Mamun, L.J. Brown, M.A. Hossain, C. Fall, L. Wagstaff, and R. Bass. A hybrid computational model for the effects of maspin on cancer cell dynamics. *Journal of Theoretical Biology*, 337:150–160, Nov 2013.

[101] MunJu Kim, Damon Reed, and Katarzyna A. Rejniak. The formation of tight tumor clusters affects the efficacy of cell cycle inhibitors: A hybrid model study. *Journal of Theoretical Biology*, 352:31–50, jul 2014.

[102] Gibin Powathil and Mark A. J. Chaplain. A hybrid multiscale approach in cancer modelling and treatment prediction. *Mathematical Oncology 2013*, pages 237–263, 2014.

[103] R.A Bernards and R.A. Weinberg. Metastasis genes: A progression puzzle. *Nature*, 418(6900):823–823, Aug 2002.

[104] Alireza Fakhrizadeh Esfahani, Philippe Dreesen, Koen Tiels, Jean-Philippe Noël, and Johan Schoukens. Parameter reduction in nonlinear state-space identification of hysteresis. *Mechanical Systems and Signal Processing*, 104:884–895, may 2018.

[105] J. Buhl, D J T. Sumpter, I. D. Couzin, J. J. Hale, E. Despland, E. R. Miller, and S. J. Simpson. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406, Jun 2006.

[106] David R. Bickel. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Canadian Journal of Statistics*, 39(4):610–631, jul 2011.

[107] Orin J. Robinson, Viviana Ruiz-Gutierrez, and Daniel Fink. Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, 24(4):460–472, dec 2017.

[108] Rebecca M. Turner, David J. Spiegelhalter, Gordon C. S. Smith, and Simon G. Thompson. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47, jan 2009.

[109] Steven P. Reise, Richard Scheines, Keith F. Widaman, and Mark G. Haviland. Multidimensionality and structural coefficient bias in structural equation modeling. *Educational and Psychological Measurement*, 73(1):5–26, jul 2012.

[110] Neil D.B. Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, nov 2015.

[111] Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, mar 1963.

[112] Michael Mak, Fabian Spill, Roger D Kamm, and Muhammad H Zaman. Single-cell migration in complex microenvironments::mechanics and signaling dynamics. 2016.

[113] Natalia L. Komarova. Spatial interactions and cooperation can change the speed of evolution of complex phenotypes. *Proc Natl Acad Sci U S A*, 111 Suppl 3:10789–10795, Jul 2014.

[114] Xin Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, September 1999.

[115] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin. Artificial neural networks: a tutorial. *Computer*, 29(3):31–44, March 1996.

[116] James M. Greene, Doron Levy, King L. Fung, Paloma Silva de Souza, Michael M. Gottesman, and Orit Lavi. Modeling intrinsic heterogeneity and growth of cancer cells. *Journal of Theoretical Biology*, Nov 2014.

[117] Anwoy Kumar Mohanty, Aniruddha Datta, and Vijayanagaram Venkatraj. A model for cancer tissue heterogeneity. *IEEE Transactions on Biomedical Engineering*, 61(3):966–974, mar 2014.

[118] Louise J Barber, Matthew N Davies, and Marco Gerlinger. Dissecting cancer evolution at the macro-heterogeneity and micro-heterogeneity scale. *Current Opinion in Genetics & Development*, 30:1–6, Feb 2015.

[119] Cicely K. Macnamara. Biomechanical modelling of cancer: Agent-based force-based models of solid tumours within the context of the tumour microenvironment. 1(2), 2021-05.

[120] Kent W Hunter, Nigel PS Crawford, and Jude Alsarraj. Mechanisms of metastasis. *Breast Cancer Research*, 10(S1), feb 2008.

[121] Thomas R. Geiger and Daniel S. Peeper. Metastasis mechanisms. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1796(2):293–308, dec 2009.

[122] W. K. Chang, C. Carmona-Fontaine, and J. B. Xavier. Tumour-stromal interactions generate emergent persistence in collective cancer cell migration. *Interface Focus*, 3(4):20130017–20130017, Jun 2013.

[123] Trenis D. Palmer, William J. Ashby, John D. Lewis, and Andries Zijlstra. Targeting tumor cell motility to prevent metastasis. *Advanced Drug Delivery Reviews*, 63(8):568–581, jul 2011.

[124] Eric R. Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, jun 1990.

[125] Manuela Ferracin, Massimo Pedriali, Angelo Veronese, Barbara Zagatti, Roberta Gafà , Eros Magri, Maria Lunardi, Gardenia Munerato, Giulia Querzoli, Iva Maestri, and et al. Microrna profiling for the identification of cancers with unknown primary tissue-of-origin. *J. Pathol.*, 225(1):43–53, Jun 2011.

[126] F.K. Ahmad, S. Deris, and N.H. Othman. The inference of breast cancer metastasis through gene regulatory networks. *Journal of Biomedical Informatics*, 45(2):350–362, Apr 2012.

[127] Kung-Jeng Wang, Bunjira Makond, and Kung-Min Wang. Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: A case study of Taiwan. *Computers in Biology and Medicine*, 47:147–160, Apr 2014.

[128] Salman Habib, Carmen Molina-Parıs, and Thomas S. Deisboeck. Complex dynamics of tumors: modeling an emerging brain tumor system with coupled reaction–diffusion equations. *Physica A: Statistical Mechanics and its Applications*, 327(3-4):501–524, Sep 2003.

[129] Parmeshwar Khurd, Claus Bahlmann, Peter Maday, Ali Kamen, Summer Gibbs-Strauss, Elizabeth M. Genega, and John V. Frangioni. Computer-aided gleason grading of prostate cancer histopathological images using texton forests. In *Proceedings of the 2010 IEEE International Conference on Biomedical Imaging: From Nano to Macro*, ISBI'10, pages 636–639, Piscataway, NJ, USA, 2010. IEEE Press.

[130] Joaquin Chapa, Ryan J. Bourgo, Geoffrey L. Greene, Swati Kulkarni, and Gary An. Examining the pathogenesis of breast cancer using a novel agent-based model of mammary ductal epithelium dynamics. *PLoS ONE*, 8(5):e64091, May 2013.

[131] K.-A. Norton and A. S. Popel. An agent-based model of cancer stem cell initiated avascular tumour growth and metastasis: the effect of seeding frequency and location. *Journal of The Royal Society Interface*, 11(100):20140640–20140640, Sep 2014.

[132] Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, Jan 2012.

[133] François-Clément Bidard, Jean-Yves Pierga, Anne Vincent-Salomon, and Marie-France Poupon. A "class action" against the microenvironment: do cancer cells cooperate in metastasis? *Cancer Metastasis Rev*, 27(1):5–10, Mar 2008.

[134] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, may 2018.

[135] Aleksandar S. Dimovski, Claus Brabrand, and Andrzej Wasowski. Variability abstractions for lifted analyses. *Science of Computer Programming*, 159:1–27, jul 2018.

[136] Alessandra R. Brazzale and Anthony C. Davison. Accurate parametric inference for small samples. *Statistical Science*, 23(4):465–484, nov 2008.

[137] https://java.com/en/. Java webpage.

[138] Simon J. Thompson, Stephanie E.M. Thompson, and Jean-Baptiste Cazier. Castles (compute and storage for the life sciences): a collection of compute and storage resources for supporting research at the university of birmingham. 2019.

[139] J. A. Lawrence and P. S. Steeg. Mechanisms of tumor invasion and metastasis. *World J Urol*, 14(3):124–130, 1996.

[140] Colin D. Paul, Panagiotis Mistriotis, and Konstantinos Konstantopoulos. Cancer cell motility: lessons from migration in confined spaces. 17(2):131–140, 2016-12.

[141] Zhihui Wang, Joseph D. Butner, Romica Kerketta, Vittorio Cristini, and Thomas S. Deisboeck. Simulating cancer growth with multiscale agent-based modeling. *Seminars in Cancer Biology*, May 2014.

[142] Christopher Z. Eddy, Helena Raposo, Aayushi Manchanda, Ryan Wong, Fuxin Li, and Bo Sun. Morphodynamics facilitate cancer cells to navigate 3d extracellular matrix. 11(1), 2021-10.

[143] R S Ostrom, S R Post, and P A Insel. Stoichiometry and compartmentation in g protein-coupled receptor signaling: implications for therapeutic interventions involving g(s). *The Journal of pharmacology and experimental therapeutics*, 294:407–412, August 2000.

[144] Stéphanie M. Pontier, Yann Percherancier, Ségolène Galandrin, Andreas Breit, Céline Galés, and Michel Bouvier. Cholesterol-dependent separation of the 2-adrenergic receptor from its partners determines signaling efficacy. *Journal of Biological Chemistry*, 283(36):24659–24672, jun 2008.

[145] S. J. Singer and G. L. Nicolson. The fluid mosaic model of the structure of cell membranes. *Science*, 175(4023):720–731, feb 1972.

[146] Davide Calebiro, Finn Rieken, Julia Wagner, Titiwat Sungkaworn, Ulrike Zabel, Alfio Borzi, Emanuele Cocucci, Alexander ZÃrn, and Martin J Lohse. Single-molecule analysis of fluorescently labeled g-protein-coupled receptors reveals complexes with distinct dynamics and organization. *Proceedings of the National Academy of Sciences of the United States of America*, 110:743–748, January 2013.

[147] Graeme Milligan. G protein-coupled receptor dimerization: Function and ligand pharmacology. *Molecular Pharmacology*, 66(1):1–7, may 2004.

[148] Martin J. Lohse, Susanne Nuber, and Carsten Hoffmann. Fluorescence/bioluminescence resonance energy transfer techniques to study g-protein-coupled receptor activation and signaling. *Pharmacological Reviews*, 64(2):299–336, mar 2012.

[149] Konstantinos Lefkimmiatis and Manuela Zaccolo. cAMP signaling in subcellular compartments. *Pharmacology & Therapeutics*, 143(3):295–304, sep 2014.

[150] D Axelrod. Cell-substrate contacts illuminated by total internal reflection fluorescence. *The Journal of Cell Biology*, 89(1):141–145, apr 1981.

[151] Daniel Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, 2(11):764–774, nov 2001.

[152] Yann Lanoiselée, Jak Grimes, Zsombor Koszegi, and Davide Calebiro. Detecting transient trapping from a single trajectory: A structural approach. 23(8):1044, 2021-08.

[153] Hippolyte Verdier, Maxime Duval, François Laurent, Alhassan Cassé, Christian L. Vestergaard, and Jean-Baptiste Masson. Learning physical properties of anomalous random walks using graph neural networks. 54(23):234001, 2021-05.

[154] Patrycja Kowalek, Hanna Loch-Olszewska, and Janusz Szwabiński. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. 100(3):032410, 2019.

[155] Martin Schrimpf. Should i use tensorflow. *arXiv*, 2016.

[156] Facundo Bre, Juan M. Gimenez, and Víctor D. Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158:1429–1441, jan 2018.

[157] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014-09-04.

[158] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[159] Mathie Najberg, Muhammad Haji Mansor, Frank Boury, Carmen Alvarez-Lorenzo, and Emmanuel Garcion. Reversing the tumor target: Establishment of a tumor trap. *Frontiers in Pharmacology*, 10, aug 2019.